USP

Campus de São Carlos

USO DE REDES NEURAIS ARTIFICIAIS PARA DESCOBERTA DE CONHECIMENTO SOBRE A ESCOLHA DO MODO DE VIAGEM

FÁBIO GLAUCO WERMERSCH

ORIENTADOR: PROF. DR. EIJI KAWAMOTO

UNIVERSIDADE DE SÃO PAULO

ESCOLA DE ENGENHARIA DE SÃO CARLOS

Serviço de Pós-Graduação EESC/USP
EXEMPLAR REVISADO
Data de entrada no Serviço: 12/07/02
Ass.: 02

Uso de redes neurais artificiais para descoberta de conhecimento sobre a escolha do modo de viagem

Fábio Glauco Wermersch

DEDALUS - Acervo - EESC
31100037048

Dissertação apresentada à Escola de Engenharia de São Carlos da Universidade de São Paulo como parte dos requisitos para obtenção do título de Mestre em Engenharia Civil com ênfase em Transportes.

Orientador Prof. Dr. Eiji Kawamoto.

São Carlos
Abril
2002

Class. TESE-EESC
Cutt. 6303
Tombo T0166/02
Sysno 1263958

Ficha catalográfica preparada pela Seção de Tratamento
da Informação do Serviço de Biblioteca – EESC/USP

W489u

Wermersch, Fábio Glauco
Uso de redes neurais artificiais para descoberta de conhecimento sobre a escolha do modo de viagem / Fábio Glauco Wermersch. -- São Carlos, 2002.

Dissertação (Mestrado) -- Escola de Engenharia de São Carlos-Universidade de São Paulo, 2002.
Área: Transportes.
Orientador: Prof. Dr. Eiji Kawamoto.

1. Análise de demanda por transporte. 2. Modelos de escolha. 3. Inteligência artificial (IA). 4. Redes neurais artificiais (RNA). 5. Árvores de decisão. 6. Mineração de dados (*Data Mining*). 7. Descoberta de conhecimento em banco de dados. I. Titulo.

FOLHA DE JULGAMENTO

Candidato: Engenheiro **FÁBIO GLAUCO WERMERSCH**

Dissertação defendida e julgada em 09-05-2002 perante a Comissão Julgadora:

APROVADO

Prof. Dr. **EIJI KAWAMOTO (Orientador)**
(Escola de Engenharia de São Carlos/USP)

APROVADO

Prof. Dr. **ALUIZIO FAUSTO RIBEIRO ARAÚJO**
(Escola de Engenharia de São Carlos/USP)

APROVADO

Prof. Dr. **ORLANDO STRAMBI**
(Escola Politécnica/USP)

Prof. Dr. **EDSON MARTINS DE AGUIAR**
Coordenador do Programa de Pós-Graduação
em Engenharia de Transportes

Profa. Assoc. Maria do Carmo Calijuri
Presidente da Comissão de Pós-Graduação da EESC

À minha avó Angelina (*in memorian*) por deixar tão patente a mim o valor do conhecimento.

AGRADECIMENTOS

Aos meus pais pelo apoio, incentivo e paciência.

Ao Prof. Eiji Kawamoto por ter aceitado guiar-me ao longo deste trabalho, mantendo-se sempre presente nos momentos mais difíceis do desenvolvimento deste.

A Edmar Martineli pela inestimável colaboração à face computacional deste trabalho, sem a qual este não seria possível.

Ao meu primo André pela revisão cuidadosa e enriquecedora do texto, preenchida com notas bem humoradas tão importantes nesta jornada.

Ao Masaki Kawabata Neto e a Edgar Bacarji pelo auxílio irrestrito, até fora de horário, na solução dos entraves técnicos informáticos surgidos.

Ao Prof. Benedito Galvão Benze pela colaboração na área estatística.

Aos amigos do Departamento de Transportes da EESC/USP, Ricardinho, Pastor, Carlos (Toco), Charlie, Norma, Márcia, Cíntia Egami e Cíntia Lotti, Juliana, Adriana, Gauchinha, Patrícia, Beth, Heloísa, Lílian, Magali, Sueli e em especial à Ana Paula Larocca (Aninha) por tanta atenção e incentivo em todas as horas.

Aos amigos Prof. Antônio Nelson R. Silva, Prof. Glauco T. P. Fabbri, Mauro Kenji e Pieter Jan Slaets pelo apoio.

À Coordenadoria de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudos concedida.

SUMÁRIO

1 INTRODUÇÃO 1
1.1 Objetivo da pesquisa 2
1.2 Método 2
1.3 Justificativa 2
1.4 Estrutura do texto 4
2 MODELOS DE DEMANDA POR TRANSPORTE 5
2.1 Modelos de escolha em transportes 7
2.1.1 Modelos compensatórios 8
2.1.1.1 Modelo de utilidade estrita 8
2.1.1.2 Modelo de utilidade aleatória 9
2.1.2 Modelos não-compensatórios 14
2.1.3 Modelo semi-compensatório 18
2.2 Considerações finais 21
3 MINERAÇÃO E DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS 24
3.1 Redes neurais artificiais 26
3.1.1 Usando redes neurais para *data mining* 30
3.2 Extração de conhecimento de RNA 32
3.2.1 Algoritmo Trepan 34
3.2.1.1 Perguntas de associação e o oráculo 36
3.2.1.2 Expansão da árvore 37
3.2.1.3 Testes de divisão 37
3.2.1.4 Critério de parada 37
3.3 Redes neurais no domínio da escolha modal em transportes 39
4 DESCRIÇÃO DO MÉTODO 42

4.1 Base digital de dados geo-referenciados 42
4.2 Fase 1: coleta e análise de dados 44
4.2.1 Variáveis de saída 44
4.2.2 Atributos de entrada 46
4.2.3 Filtragem dos dados 46
4.2.4 Amostragem 46
4.3 Fase 2: pré-processamento dos dados para entrada na simulação de rede neural artificial 48
4.3.1 Normalização e codificação 48
4.3.2 Formação das partições: treinamento, validação e teste 49
4.4 Fase 3: simulação de redes neurais artificiais 49
4.4.1 Treinamento de RNA 50
4.4.2 Seleção da arquitetura 51
4.5 Fase 4: pós-processamento da RNA para entrada no algoritmo Trepan 53
4.5.1 Conversão da RNA 53
4.5.2 Conversão dos dados 53
4.6 Fase 5: extração de conhecimento 54
4.6.1 Seleção da árvore de decisão 54
4.6.2 Desnormalização 54
4.7 Fase 6: análise de resultados 55
4.7.1 Análise e interpretação 55
4.7.2 Consolidação do conhecimento extraído 56
4.8 Considerações finais 56
5 COLETA E ANÁLISE DE DADOS 57
5.1 Caracterização da cidade em estudo 57
5.2 Pesquisa origem-destino 58
5.3 Preparação e análise dos dados 60
5.3.1 Utilização de SIG 60
5.3.2 Definição das alternativas de escolha modal 61
5.3.3 Definição dos atributos ou variáveis de entrada 62
5.3.3.1 Atributos sócio-econômicos 63
5.3.3.2 Atributos de viagem 64
5.3.3.2.1 O atributo tempo de viagem 64
5.3.4 Filtragem dos dados 67

5.3.5 Amostragem 68
5.3.5.1 Tamanho da amostra 69
5.3.5.2 Segmentação da base de dados e seleção de registros 71
5.3.5.3 Coleta de distâncias de viagem por modo ônibus 75
6 MINERAÇÃO: OBTENÇÃO E ANÁLISE DE RESULTADOS 79
6.1 Diretriz de busca de paradigmas 79
6.2 A busca 80
6.2.1 Resultados preliminares 81
6.2.2 Resultados conclusivos 92
6.3 Considerações finais 101
7 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS 102
7.1 Apresentação dos resultados 102
7.2 Discussão dos resultados 119
8 CONCLUSÕES E RECOMENDAÇÕES 130
8.1 Conclusões 130
8.2 Recomendações 132
ANEXO A: figuras de árvores de decisão 134
ANEXO B: figuras de árvores de decisão e suas planilhas de condições. 141
ANEXO C: método de distinção entre os modos a pé e bicicleta 148
ANEXO D: atributos de entrada utilizados nos experimentos 156
REFERÊNCIAS BIBLIOGRÁFICAS 162
APÊNDICE

LISTA DE FIGURAS

2.1 Estrutura do processo de decisão de modelos compensatórios..... 13

2.2 Estrutura do processo de decisão do modelo não-compensatório de Escolha Lexicográfica..... 16

2.3 Estrutura do processo de decisão do modelo semi-compensatório..... 20

3.1 Uma visão geral do processo de KDD..... 25

3.2 Uma rede neural MLP com 4 unidades de entrada, 3 unidades intermediárias e uma unidade de saída (4-3-1)..... 27

3.3 Exemplo de árvore de decisão de nível de profundidade igual a três.... 35

4.1 Estrutura do método proposto..... 43

5.4 Exemplo de código para um registro de viagem..... 72

5.5 Gráfico ilustrativo da seleção de registros codificados..... 73

5.6 Localização das origens das viagens de toda a base de dados..... 74

5.7 Localização das origens das viagens da amostra..... 74

5.8 Rede de transporte público formada por sobreposição das 85 rotas (linhas) de ônibus existentes, em destaque sobre a rede do sistema viário..... 76

5.9 Exemplo de viagem realizada utilizando uma linha de ônibus..... 77

5.10 Exemplo de viagem realizada utilizando duas linhas de ônibus..... 77

6.1 Situações típicas de conceitos descobertos para uma classe alvo..... 80

6.2 Gráfico de resultados do experimento 1 para 4 modos de viagem..... 82

6.3 Gráfico dos resultados do experimento 2 para 4 modos de viagem..... 83

6.4 Gráfico dos resultados do experimento 3 para 4 modos de viagem..... 84

6.5 Gráfico de resultados do experimento 4 para 4 modos de viagem..... 85

6.6 Gráfico dos resultados do experimento 5 para 4 modos de viagem..... 86

6.7 Continuação do gráfico da Figura 6.6..... 86

6.8 Continuação do gráfico da Figura 6.6..... 86

6.9 Continuação do gráfico da Figura 6.6..... 87

6.10 Continuação do gráfico da Figura 6.6 87
6.11 Gráfico de importância relativa das entradas da rede 31-11-4 89
6.12 Gráfico dos resultados do experimento 6 para 4 modos de viagem 90
6.13 Gráfico dos resultados do experimento 7 para 4 modos de viagem 90
6.14 Gráfico dos resultados do experimento 8 para 4 modos de viagem 91
6.15 Gráfico dos resultados do experimento 1 para 3 modos de viagem 92
6.16 Detalhe da raiz da árvore extraída da rede 32-10-10 93
6.17 Detalhe da raiz da árvore extraída da rede 32-15-15 94
6.18 Gráfico dos resultados do experimento 2 para 3 modos de viagem 94
6.19 Continuação do gráfico da Figura 6.18 95
6.20 Gráfico dos resultados do experimento 3 para 3 modos de viagem 96
6.21 Continuação do gráfico da Figura 6.20 96
6.22 Gráfico dos resultados do experimento 4 para 3 modos de viagem 97
6.23 Continuação do gráfico da Figura 6.22 98
6.24 Gráfico dos resultados do experimento 5 para 3 modos de viagem 99
6.25 Continuação do gráfico da Figura 6.24 99
6.26 Continuação do gráfico da Figura 6.24 100
7.1 Árvore de decisão extraída da rede 11-11-3 com profundidade 7 103
7.2 Árvore de decisão extraída da rede 5-9-3 com profundidade 7 107
7.3 Árvore de decisão extraída da rede 5-12-9-3 com profundidade 7 111
7.4 Árvore de decisão extraída da rede 5-3-3-3 com profundidade 7 114
7.5 Diagrama das regras da árvore da rede 5-9-3 para o modo a pé, com o número de instâncias compreendidas por local 120
7.6 Diagrama das regras coincidentes 15, 16, 17 e 18 da árvore da rede 5-9-3 para o modo ônibus, com o número de instâncias compreendidas por local 121
A.1 Árvore de profundidade 30 extraída da rede 32-10-10-3 135
A.2 Árvore de profundidade 31 extraída da rede 32-15-15-3 136
A.3 Árvore de profundidade 48 extraída da rede 31-10-10-3 137
A.4 Árvore de profundidade 29 extraída da rede 31-30-15-3 138
A.5 Árvore extraída da rede 33-30-20-3 com profundidade 35 139
A.6 Árvore extraída da rede 33-20-15-3 com profundidade 21 140
B.1 Árvore extraída da rede 11-30-3 com profundidade 46 142
B.2 Árvore extraída da rede 11-10-5-3 com profundidade 21 144
B.3 Árvore extraída da rede 11-15-15-3 com profundidade 22 146

C.1 Viagens a pé X Comprimento da viagem..... 151
C.2 Probabilidade de escolha modo a pé X Comprimento da viagem..... 151
C.3 Probabilidade de escolha do modo a pé em função da distância..... 152

LISTA DE TABELAS

5.7 Tamanhos necessários da amostra..... 70

5.8 Erros de estimativa para intervalo de confiança de 95% e tamanho de amostra de 1.451 elementos..... 78

C.1 Tabela distância X velocidade para o motivo de viagem "trabalho"..... 153

C.2 Tabela distância X velocidade para o motivo de viagem "escola"..... 153

C.3 Tabela distância X velocidade para o motivo de viagem "compras"..... 153

C.4 Tabela distância X velocidade para o motivo de viagem "saúde"..... 154

C.5 Tabela distância X velocidade para o motivo de viagem "lazer"..... 154

C.6 Tabela distância X velocidade para o motivo de viagem "outros"..... 154

D.1 Atributos de entrada do primeiro experimento da etapa preliminar..... 157

D.2 Atributos de entrada do segundo experimento da etapa preliminar..... 157

D.3 Atributos de entrada do terceiro experimento da etapa preliminar..... 157

D.4 Atributos de entrada do quarto experimento da etapa preliminar..... 158

D.5 Atributos de entrada do quinto experimento da etapa preliminar..... 158

D.6 Atributos de entrada do sexto experimento da etapa preliminar..... 158

D.7 Atributos de entrada do sétimo experimento da etapa preliminar..... 159

D.8 Atributos de entrada do oitavo experimento da etapa preliminar..... 159

D.9 Atributos de entrada do primeiro experimento da etapa conclusiva..... 159

D.10 Atributos de entrada do segundo experimento da etapa conclusiva..... 160

D.11 Atributos de entrada do terceiro experimento da etapa conclusiva..... 160

D.12 Atributos de entrada do quarto experimento da etapa conclusiva..... 160

D.13 Atributos de entrada do quinto experimento da etapa conclusiva..... 161

LISTA DE PLANILHAS

7.1 Condições para os modos de viagem da árvore extraída da rede 11-11-3 com profundidade 7 104

7.2 Condições simplificadas para os modos de viagem da árvore extraída da rede 11-11-3 com profundidade 7 105

7.3 Regras para os modos de viagem da árvore extraída da rede 11-11-3 com profundidade 7 106

7.4 Condições para os modos de viagem da árvore extraída da rede 5-9-3 com profundidade 7 108

7.5 Condições simplificadas para os modos de viagem da árvore extraída da rede 5-9-3 com profundidade 7 109

7.6 Regras para os modos de viagem da árvore extraída da rede 5-9-3 com profundidade 7 110

7.7 Condições para os modos de viagem da árvore extraída da rede 5-12-9-3 com profundidade 7 112

7.8 Condições simplificadas para os modos de viagem da árvore extraída da rede 5-12-9-3 com profundidade 7 113

7.9 Condições para os modos de viagem da árvore extraída da rede 5-3-3-3 com profundidade 7 115

7.10 Condições simplificadas para os modos de viagem da árvore extraída da rede 5-3-3-3 com profundidade 7 116

7.11 Regras para os modos de viagem da árvore extraída da rede 5-3-3-3 com profundidade 7 117

7.12 Número de instâncias corretamente classificadas e correspondente percentual de acerto de classificação das regras para os modos de viagem da árvore extraída da rede 5-9-3 com profundidade 7 aplicadas individualmente, fora da estrutura hierarquica da árvore, a um conjunto de 567 instâncias 123

7.13 Número de instâncias adicionais corretamente classificadas e correspondente percentual de acerto de classificação das regras para os modos de viagem da árvore extraída da rede 5-9-3 com profundidade 7, aplicadas de acordo com a hierarquia da árvore apenas em relação ao modo de viagem, a um conjunto de 567 instâncias..... 124

B.1 Condições para os modos de viagem da árvore extraída da rede 11-30-3 com profundidade 46 apresentadas conforme ordenamento por modo..... 143

B.2 Condições para os modos de viagem da árvore extraída da rede 11-10-5-3 com profundidade 15 apresentadas conforme ordenamento por modo..... 145

B.2 Condições para os modos de viagem da árvore extraída da rede 11-15-15-3 com profundidade 22 apresentadas conforme ordenamento por modo..... 147

LISTA DE ABREVIATURAS E SIGLAS

ANFAVEA	Associação Nacional dos Fabricantes de Veículos Automotores
EMDURB	Empresa Municipal de Desenvolvimento Urbano e Rural de Bauru
FHWA	Federal Highway Administration
FINEP	Financiadora de Estudos e Projetos
GEIPOT	Empresa Brasileira de Planejamento de Transportes
KDD	Knowledge Discovery in Database
RECOPE	Rede Cooperativa de Pesquisa
RNA	Redes Neurais Artificiais
SEADE	Fundação Sistema Estadual de Análise de Dados
SIG	Sistema de Informações Geográficas
Trepan	Trees Parroting Network

RESUMO

WERMERSCH, F. G. (2002) *Uso de redes neurais artificiais para descoberta de conhecimento sobre a escolha do modo de viagem*. São Carlos, 2002. 175p. Dissertação (Mestrado) – Escola de Engenharia de São Carlos, Universidade de São Paulo.

Esta pesquisa objetivou uma melhor compreensão do processo de escolha do modo de viagem. Empregou-se a abordagem indutiva dirigida a dados livre de suposições *a priori* da mineração em banco de dados (*Data Mining*), utilizando redes neurais artificiais (RNA) como ferramenta mineradora, à procura de conhecimento, ou informação útil, a respeito do processo de escolha e capaz de indicar qual das estruturas de decisão subjacentes aos modelos de escolha modal considerados mais se aproximaria do observado. Partindo-se da idéia de que nesse processo exista um padrão o qual pode ser captado por uma RNA, ajustou-se um modelo de RNA aos dados e extraiu-se então o conhecimento contido no modelo de RNA ajustado através de um algoritmo de extração de árvore de decisão de RNA chamado Trepan (Trees *parroting network*), que foi analisado e interpretado à luz dos objetivos desta pesquisa. Os dados que foram utilizados nesse processo de descoberta de conhecimento são provenientes de uma pesquisa de entrevista domiciliar realizada na cidade de Bauru – SP, para fins de estimativa da matriz de deslocamentos origem-destino dessa cidade. Obteve-se quatro árvores de decisão com estruturas simples e com acurácia preditiva de 75% aproximadamente para os três modos de viagem estudados. Embora o conhecimento extraído dos modelos neurais ajustados não tenham proporcionado a indicação de qual das estruturas de decisão subjacentes aos modelos de escolha modal mais se aproximaria da obtida com o modelo neural, foi constatada nas árvores resultantes do processo de descoberta de conhecimento uma relação de compensação entre o atributo sexo e os atributos relacionados à capacidade econômica do domicílio na decisão de escolha do modo carro para a realização de uma viagem. Os resultados também sugerem a não necessidade de mais de um atributo de entrada referente ao deslocamento realizado em uma viagem para a modelagem por RNA do processo de escolha do modo de viagem no contexto estudado.

Palavras-chave: Análise de Demanda por Transporte,
Modelos de Escolha,
Inteligência Artificial (IA),
Redes Neurais Artificiais (RNA),
Árvores de Decisão,
Mineração de Dados (*Data Mining*) e
Descoberta de Conhecimento em Banco de Dados.

ABSTRACT

WERMERSCH, F. G. (2002) *Using artificial neural network for the discovery of mode travel choice knowledge.* São Carlos, 2002. 175p. Dissertation (M.Sc.) – São Carlos School of Engineering, Universidade de São Paulo.

This research aimed at a better understanding of the mode travel choice process. The inductive data driven free from a priori assumptions of the *data mining* approach was employed, using artificial neural networks (ANN) as a mining tool, looking for knowledge or useful information, concerning the choice process and capable of indicating which of the underlying decision structures to the considered modal choice models would come closer to the observed one. Taking into consideration that there is a pattern in this process that can be captured by ANN, an ANN model was fitted (trained) to the data, and the knowledge contained in the trained ANN model was extracted by employing an ANN decision tree extraction algorithm called Trepan (Trees parroting network), which was analysed and interpreted in the light of the object of this research. The data which was employed in this knowledge discovery process come from a household survey carried out in Bauru – SP in order to estimate the O-D matrix in this city. Four decision trees with simple structures and predicting accuracy of approximately 75% for the three travel modes studied were obtained. Even though the knowledge extracted from the trained ANN model has not yielded the indication of which of the underlying decision structures to the modal choice models was closer to the neural model, a compensating relation between the sex attribute and the household economic-related attribute in the decision of choosing the car mode in order to travel was evidenced in the trees resulting from the process of knowledge discovery. The results also suggest the lack of necessity of more than one input travel attribute concerning the displacement performed in a trip for the ANN modelling of the mode travel choice process in the studied context.

Key words: Transport Demand Analysis,
Choice Models,
Artificial Intelligence (AI),
Artificial Neural Network (ANN),
Decision Trees ,
Data Mining and
Knowledge Discovery in Database (KDD).

1 INTRODUÇÃO

Por várias razões, atualmente tornou-se evidente a necessidade de medidas que reduzam o número de veículos individuais em circulação. Para que se possa conhecer os efeitos destas medidas antecipadamente, são utilizados de forma ampla modelos comportamentais, isto é, modelos matemáticos baseados no comportamento humano da escolha do modo de transporte. Entre eles estão os modelos compensatórios, os modelos não-compensatórios e os modelos semi-compensatórios, que serão discutidos adiante. Cada modelo existente fundamenta-se em diferentes hipóteses de representação da estrutura de decisão.

Apesar dos constantes aperfeiçoamentos destes modelos, principalmente os compensatórios, não se conseguiu ainda atingir o resultado esperado (MCFADDEN, 2000). As causas podem ser várias: a) o processo de escolha do modo de viagem não está bem compreendido; b) as hipóteses a respeito das estruturas de decisão não condizem com a realidade; c) a especificação dos modelos não reflete corretamente a maneira como as diferentes informações são processadas na mente dos usuários; e d) as informações consideradas por usuários estão parcialmente distorcidas. É possível ainda que algumas das causas acima estejam agindo em conjunto ao invés de isoladamente.

Neste contexto, este estudo enfoca as estruturas de decisão subjacentes aos modelos de escolha modal. Para tanto, aplica-se um processo computacional de descoberta de conhecimento em banco de dados, o qual utiliza redes neurais artificiais (RNA) - que têm como sua principal característica o reconhecimento de padrões através de métodos de aprendizado – como ferramenta mineradora de informação associada a um algoritmo de extração de conhecimento de RNA, na tentativa de detectar o processo mental seguido por usuários de transporte na escolha de um modo de viagem.

1.1 OBJETIVO DA PESQUISA

O objetivo da pesquisa é compreender melhor o processo de escolha do modo de viagem.

Tem-se como meta a obtenção de considerações a respeito do mecanismo subjacente ao comportamento do usuário de transportes em relação à escolha modal, obtida através do conhecimento contido no modelo de RNA ajustado para o domínio estudado.

1.2 MÉTODO

O estudo será realizado a partir de dados coletados na cidade de Bauru, localizada a 340 km a oeste da cidade de São Paulo.

Procura-se uma estrutura de decisão que se ajuste o melhor possível às escolhas observadas em Bauru. Para tanto, será desenvolvida uma arquitetura de RNA que melhor classifique os padrões de escolha modal observados e então, será extraído o conhecimento contido na RNA desenvolvida (treinada), na forma de árvore de decisão que melhor o aproxime. A partir desse conhecimento procura-se formular considerações sobre o comportamento de escolha modal observado no espaço de instâncias utilizado.

1.3 JUSTIFICATIVA

A aplicação de redes neurais artificiais em diversos domínios tem sido intensificada nos últimos anos devido aos bons resultados de constantes pesquisas. As RNA, também conhecidas como "modelos conexionistas" da Inteligência Artificial (IA), foram propostas para resolver problemas em que os seres humanos alcançam boa performance e são de difícil tratamento computacional tradicional, como por exemplo o reconhecimento de rostos (MARTINELI, 1999). Com bom desempenho na solução de problemas de reconhecimento de padrões (BISHOP, 1995), e como aproximador universal de funções multi-variáveis (CYBENKO 1988 & 1989), as RNA não requerem conhecimento prévio sobre os modelos matemáticos utilizados no domínio das aplicações (BITTENCOURT, 1996). Essa característica de não necessidade de hipóteses estabelecidas *a priori* das RNA a configura como uma

das ferramentas da abordagem indutiva, dirigida a dados e livre de hipóteses da mineração de dados (*data mining*), para a captura de relações funcionais subjacentes aos dados, mesmo quando essas são desconhecidas ou apresentem dificuldade de descrição (ZHANG et al., 1998).

Porém, todo o poder oferecido pelas RNA na modelagem de domínios esbarra em um problema: sua incapacidade para explicar de forma compreensível suas decisões. Procura-se hoje por sistemas que tenham capacidade de aprendizado e generalização e que possam, também, apresentar um alto grau de compreensibilidade (BRAGA et al., 1998). O processo de descoberta de conhecimento em banco de dados, que inclui a mineração de dados, tem como principais objetivos em termos práticos a previsão de valores futuros ou desconhecidos para variáveis de interesse e a descrição de padrões interpretáveis ao ser humano que estejam escondidos nos dados (FAYYAD, 1996). Assim, esse processo envolve a extração, interpretação e representação do conhecimento de um dado domínio. Algoritmos de extração de conhecimento de RNA permitem a descrição de sua estrutura em forma lingüística, gerando um conjunto de conceitos a partir da RNA treinada. O algoritmo de extração de conhecimento de RNA chamado Trepan (CRAVEN 1996), a ser aplicado neste trabalho, gera uma árvore de decisão que aproxima a função representada por uma RNA treinada, explicitando a estrutura de decisão contida nesta.

Como dito anteriormente, apesar dos aperfeiçoamentos dos modelos de escolha modal, não se conseguiu ainda atingir o resultado esperado. Ao fazer uma revisão do desenvolvimento dos modelos de escolha em transporte nos últimos 30 anos, MCFADDEN (2000) conclui que o desenvolvimento de um unificado modelo comportamental ainda não foi completado e que os modelos atuais apoiam-se pesadamente em métodos de calibração de engenharia e não representam um refinamento cumulativo de estudos comportamentais. Assim, espera-se que as considerações sobre o comportamento do usuário de transportes em relação à escolha do modo de viagem, obtidas através do conhecimento extraído da RNA treinada, possam auxiliar o estabelecimento de hipóteses realistas na elaboração de modelos de divisão modal, indicar quais as variáveis e como deverão ser consideradas no modelo e ainda qual, dentre as estruturas do processo de escolha modal até hoje propostas, a que mais se aproxima da realidade.

1.4 ESTRUTURA DO TEXTO

Este trabalho está organizado em sete capítulos, constituindo a presente introdução o capítulo primeiro, na qual apresentam-se os objetivos da pesquisa, um resumo do método usado e a justificativa para o seu desenvolvimento.

O capítulo 2 apresenta uma visão geral dos tipos de modelos matemáticos de escolha utilizados para estimativa da divisão modal em transportes.

Por sua vez, o capítulo 3 define o conceito do processo de descoberta de conhecimento em banco de dados, apresenta o modelo de redes neurais artificiais empregado na mineração dos dados disponíveis a este trabalho, e o paradigma extrator de conhecimento de rede neural utilizado chamado Trepan.

O método usado no desenvolvimento do trabalho proposto encontra-se descrito, em termos das fase e procedimentos, no capítulo 4. No capítulo 5 é feita a descrição pormenorizada dos procedimentos realizados na fase de aquisição do conjunto de dados estudado.

O capítulo 6 contém a apresentação e a análise da etapa de mineração de dados deste trabalho.

A apresentação e discussão dos resultados obtidos fazem parte do capítulo 7. As conclusões e recomendações deste trabalho estão apresentadas no capítulo 8. Adiante neste texto estão dispostos os anexos A, B, C e D. O anexo A contém figuras de árvores de decisão e o anexo B contém figuras de árvores de decisão conjugadas com correspondentes planilhas das condições contidas nas árvores. O anexo C descreve o método utilizado para a separação dos modos de viagem bicicleta e caminhada e o anexo D apresenta tabelas com os atributos de entrada utilizados em cada experimento. Há ainda as referências bibliográficas e um apêndice.

2 MODELOS DE DEMANDA POR TRANSPORTE

Até o início da década de 70 a demanda por transporte de passageiros era estimada através da projeção de fluxo de passageiros para o horizonte de projeto, conforme tendência observada em dados históricos, ou através de modelos matemáticos que relacionavam as variáveis consideradas explicativas com o tráfego. Estes modelos matemáticos, por empregarem as variáveis agregadas da zona de tráfego, ficaram conhecidos como modelos agregados. A partir do início dos anos 70 iniciou-se o desenvolvimento sistemático dos modelos para previsão da demanda chamados "comportamentais desagregados", quando os administradores públicos europeus e norte-americanos passaram a manifestar preocupações com a escalada do congestionamento do sistema viário de grande cidades, e com a elevação do preço do petróleo. Estes administradores públicos estavam, então, convencidos de que se fazia necessária a adoção de medidas visando aliviar tanto o problema do congestionamento quanto o do consumo de derivados de petróleo. Mas era desejável ter-se antecipadamente a previsão dos efeitos de tais medidas. Entretanto, os modelos disponíveis na época, os modelos agregados, cujas variáveis são índices médios da zona de tráfego, não eram suficientemente sensíveis para refletir o impacto das medidas que fossem tomadas sobre os usuários. Vários tipos de problemas associados aos modelos agregados foram documentados, como as chamadas falácias ecológicas (SPEAR, 1977) e ajustes devidos à agregação, guiando pesquisas na tentativa de contorno de tais problemas (OPENSHAW, 1977). Segundo WEBER (1980), os pesquisadores, na época, passaram a procurar uma teoria para os modelos de demanda por transporte que enfocasse uma unidade decisória mais significante que as arbitrárias unidades espaciais, as zonas de tráfego. "Zonas não viajam, pessoas viajam" era o slogan do projeto *Travel Demand Forecasting* da universidade de Berkeley - E.U.A, sob direção de Daniel L. McFadden em 1972. Foi neste contexto que surgiu a

necessidade do desenvolvimento de modelos comportamentais, isto é, modelos baseados no comportamento do homem, com os quais fosse possível avaliar, no curto prazo, a reação dos usuários frente a medidas tomadas, medidas estas políticas ou técnicas (para maiores detalhes ver RICHARDS & BEN-AKIVA, 1975). O termo "desagregado" surgiu portanto para diferenciar os novos modelos daqueles que empregam as variáveis agregadas em nível zonal.

Os novos modelos, desagregados então, trouxeram diferentes necessidades em relação aos dados para modelagem. Pesquisas detalhadas de coleta de dados e instrumentos como questionários foram desenvolvidos para o registro das viagens, realizadas por pessoas, e para a obtenção de suas preferências e atitudes a respeito de modos de viagem, rotas, e destinos.

Os primeiros modelos desagregados foram aplicados à escolha do modo de viagem, para viagens realizadas por motivo trabalho. Viagens por motivo trabalho com suas rotas fixas, suas observáveis regularidades, e sua característica de congestionamento na hora pico, atraíram muitos estudos de demanda por transportes em grandes áreas metropolitanas. Ainda com relação ao paradigma comportamental desagregado de escolha, segundo PIPKIN (1995), durante os anos 70 havia um consenso sobre os melhores modelos de comportamento de viagem: os modelos probabilísticos de escolha, em especial os da família logit multinomial (MNL), que serão enfocados mais adiante neste texto. Até o final da década seguinte tais modelos probabilísticos de escolha foram continuamente refinados e estendidos (ver BEN-AKIVA & LERMAN, 1985 e FISCHER & NIJKAMP, 1987). Utilizando estes modelos e na busca pela melhoria da performance predictiva, alguns estudos foram feitos separando as viagens por diferentes motivos, além do trabalho, como por exemplo por motivo compras, lazer e social (GORDON et al,1988); outros estudos lançaram mão da segmentação de mercado e buscaram modelar o comportamento de viagem de grupos especiais, incluindo mulheres, minorias, deficientes físicos (por exemplo, OPACIC & POTTER, 1986), idosos (SMITH, 1992), e o grupo dos que não possuem carro (por exemplo, BROMLEY & THOMAS, 1993).

Mas já na década de 70 a insatisfação com a primeira geração dos modelos de escolha desagregado, fez emergir novas propostas, como o modelo de escolha baseada na eliminação por aspectos (TVERSKY, 1972), e a idéia de que um

indivíduo teria suas atividades sujeitas a restrições no espaço e no tempo, que mais tarde gerou os modelos baseados em atividades (HAGERSTRAND, 1970).

Na seção seguinte serão apresentados os principais tipos de modelos de escolha, bem como as estruturas subjacentes a cada um deles.

2.1 MODELOS DE ESCOLHA EM TRANSPORTES

As principais idéias sobre as quais se assentam as teorias sobre o comportamento de escolha do usuário de transporte foram incorporadas, com algumas adaptações e simplificações, das teorias econômicas e psicológicas.

Da Economia tomou-se de empréstimo a noção de utilidade. De acordo com KAWAMOTO (1987), o termo utilidade foi concebido pelos economistas para refletir o nexo existente entre as coisas e as necessidades e os desejos do homem, tendo sido um pouco distorcido no âmbito do estudo de transportes, passando a depender da renda do consumidor e do preço entre outros, sob a denominação de utilidade indireta.

As teorias econômicas desenvolvidas para o comportamento do consumidor postulam que o mesmo percebe a utilidade dos bens de consumo. Por exemplo, segundo Samuelson (SIMONSEN, 1971), o consumidor possui o senso de comparação para discriminar se um conjunto é mais útil do que outro. Tais teorias postulam também que o consumidor maximiza sua satisfação, proporcionada pelos bens adquiridos e reserva monetária (ex.: poupança), conforme permite sua renda. Ainda que fosse possível construir teorias do comportamento do usuário de transporte baseadas nas teorias econômicas tradicionais, DALVI (1978) ressalta que existem duas importantes diferenças entre as características dos bens de consumo e do transporte. Primeiramente, em transportes as alternativas de escolha são qualitativas e discretas, isto é, os indivíduos têm de escolher uma opção a partir de um conjunto definido de alternativas, e daí advém o fato de os modelos de escolha em transportes usualmente serem chamados de *modelos de escolha discreta*. E em segundo lugar, as alternativas com que os usuários de transportes se defrontam não necessariamente estão bem definidas como um conjunto de bens. O autor acredita, porém, que os usuários muitas vezes exercem a escolha

sobre algo mais abstrato, que poderia ser o conjunto de atributos de cada alternativa. A este respeito, LANCASTER apud ORTÚZAR & WILLUNSEN (1994) sustenta que alternativas de escolha *per se* não produzem utilidade, mas esta seria derivada dos atributos de cada alternativa e dos atributos do indivíduo que exerce a escolha.

Da Psicologia emprestou-se a idéia de que o comportamento de escolha é um fenômeno probabilístico e a expressão matemática, a função logística, para descrever a variabilidade na percepção humana (KAWAMOTO, 1987).

Uma revisão sobre a influência das áreas de Economia e Psicologia na construção dos atuais modelos de previsão do comportamento dos usuários de transporte pode ser encontrada em KAWAMOTO (1987), BEN-AKIVA (1999) e MCFADDEN (2000).

2.1.1 MODELOS COMPENSATÓRIOS

2.1.1.1 Modelo de Utilidade Estrita

De acordo com a psicologia, as decisões humanas seriam fenômenos probabilísticos por natureza, e seriam apoiadas na avaliação de utilidades das alternativas de escolha. A probabilidade de escolha de uma dada alternativa seria, então, uma relação entre as utilidades das alternativas:

$$p_j = \frac{v_i}{\sum_{j=1}^{n} v_j} \tag{2.1}$$

onde p_j é a probabilidade da j-ésima opção a ser escolhida; e v_j são as utilidades associadas com as alternativas à escolha. Entretanto, esta probabilidade de escolha refere-se a um grupo e não a um indivíduo.

Originalmente formulada por LUCE (1959), a idéia por trás deste modelo sustenta que qualquer alternativa j tem utilidade v_{ji} determinada por seus atributos,

que são representados pelo vetor X_j, e modificada de alguma forma pelos atributos do indivíduo i, estes representados pelo vetor S_i, como mostra a Equação 2.2.

$$v_{ji} = v(X_j, S_i) \tag{2.2}$$

De acordo com esta teoria, cada indivíduo percebe a utilidade deterministicamente em função das variáveis, ou atributos, incluídos no modelo e escolhe a alternativa que se apresenta como aquela de maior utilidade. Porém, devido à heterogeneidade do grupo, ocorrem escolhas diversas que podem ser tratadas estatisticamente. Assim, esta teoria apoia-se na hipótese de que quanto maior for a utilidade de uma alternativa tanto maior será a probabilidade desta ser escolhida.

Esse modelo que parte do pressuposto de que a estimativa da utilidade é determinística, é denominado "Modelo de Utilidade Estrita" ou "Modelo de Utilidade Constante", e foi aplicado aos estudos de transporte pela primeira vez por Warner em 1962 (STOPHER & MEYBURG, 1976).

2.1.1.2 Modelo de Utilidade Aleatória

Uma alternativa de modelo de escolha comportamental desagregado foi proposta por MCFADDEN (1975), o modelo de utilidade aleatória. A teoria por trás deste modelo preserva o princípio da maximização da utilidade para escolha de uma alternativa e postula que indivíduos pertencentes a um segmento homogêneo de mercado – grupo de pessoas com características sócio-econômicas observadas e oportunidades de transportes similares – atribuem, apesar das similaridades entre si, diferentes valores de utilidade a uma dada alternativa de escolha. A variação na quantificação da utilidade de uma alternativa poderia ter como causa fatores subjetivos individuais, difíceis de serem incorporados ao modelo, dada sua própria natureza.

Para McFadden, a percepção da utilidade de uma dada alternativa por um indivíduo é descrita em função da média dos efeitos das variáveis observadas para um segmento que influenciam a escolha, e do provável desvio individual em relação a essa média. Assim, a utilidade atribuída por um indivíduo a uma dada alternativa,

também chamada de utilidade indireta, consistiria de um componente determinístico ou sistemático (v) e um aleatório ou residual (e), podendo ser representada de acordo com a Equação (2.3). Vale lembrar que a utilidade direta de um bem refere-se à sua qualidade de afastar um estado de necessidade ou de desejo de uma pessoa, e portanto é desvinculada do seu preço ou do poder aquisitivo do indivíduo. Já na utilidade indireta não há essa desvinculação.

$$U(X_j, X'_j, S, S') = v(X_j, S) + e(X'_j, S') \tag{2.3}$$

ou simplesmente,

$$U_j = v_j + e_j \tag{2.4}$$

onde v é o valor médio de utilidade atribuído à alternativa j pelo segmento homogêneo de mercado com atributos da alternativa X e características sócio-econômicas S incluídas no modelo; e e é um desvio da média que tem uma distribuição de probabilidade induzida pela distribuição das variáveis não incluídas (X' e S') no modelo.

Pelo princípio de maximização de utilidade, a probabilidade de se escolher a alternativa j é dada pela probabilidade percebida de j superar as utilidades percebidas de todas as demais alternativas, como mostra a expressão a seguir:

$$p_j = \text{prob}\{U_j > U_k\} \qquad \text{para } \forall\, k \neq j \tag{2.5}$$

Diferentes formas do modelo surgem quando diferentes distribuições de probabilidade são assumidas para o termo e. Uma forma que se popularizou durante a década de 70 como a mais simples para implementação do modelo é a chamada Multinomial Logit (MNL). O modelo Multinomial Logit é também referido como o "modelo padrão" (*standard model*) ou ainda "modelo convencional" na literatura, dada sua extensiva aplicação na modelagem de demanda em transportes, desde sua criação até os dias atuais (HOROWITZ, 1985; PIPKIN, 1995; MCFADDEN, 1997; BEN-AKIVA et al, 1999). Uma completa apresentação da família de modelos MNL, com várias alternativas, extensões e procedimentos

estatísticos de calibração, pode ser encontrada em BEN-AKIVA & LERMAN (1985). De uma maneira geral o MNL apresenta-se da seguinte forma:

$$p_{ji} = \frac{\exp(U_{ji})}{\sum_k \exp(U_{ki})} \qquad \text{para } \forall\, k \neq j \qquad (2.6)$$

$$U_{ji} = e_{ji} + v_{ji} \qquad (2.7)$$

$$v_{ji} = \sum_n \theta_{nj} X_{nji} \qquad (2.8)$$

onde v é função dos n atributos X incluídos no modelo e varia para cada alternativa j (j variando de 1 até k alternativas à escolha) e para cada indivíduo i (ORTÚZAR & WILLUMSEN, 1994).

Apesar de ambas as teorias de utilidade, a Estrita e a Aleatória, resultarem em modelos de formas semelhantes, a representação matemática da utilidade se difere entre elas. O Modelo de Utilidade Estrita, de acordo com WARNER (1962) apud STOPHER & MEYBURG (1976), assume que a utilidade é uma função multiplicativa cujos fatores correspondem a funções exponenciais dos atributos. Por sua vez, o Modelo de Utilidade Aleatória, de acordo com MCFADDEN (1975), assume que a utilidade é uma função linear dos atributos. A principal diferença entre os enfoques de Warner e McFadden reside no fato de que o primeiro assumiu o caráter determinístico da utilidade e a heterogeneidade do grupo de observação, enquanto que o segundo atribuiu determinada aleatoriedade à utilidade percebida por um grupo homogêneo de indivíduos (KAWAMOTO, 1987). Atualmente a interpretação da utilidade não mais passa pela idéia de grupo homogêneo, porém é tomada de forma individual. Dessa forma, cada indivíduo percebe a utilidade de uma alternativa a qual pode ser representada pela Equação 2.4.

Embora existam diferenças na representação matemática entre os modelos de utilidade Estrita e Aleatória, esses são classificados como *modelos compensatórios*, pois em ambos está implícito que o processo de escolha é compensatório, ou seja, que há uma compensação entre atributos, ou ainda que a

mudança de um ou mais atributos, na função utilidade, pode ser compensada pela mudança, em sentido contrário, de alguns dos demais atributos (RECKER & GOLOB, 1979).

Para um usuário do sistema de transporte, o processo de decisão compensatório equivale a dizer, por exemplo, que um aumento no preço da passagem de ônibus deve ser acompanhado de uma elevação na qualidade do serviço para que a satisfação do usuário continue a mesma. Ainda para elucidar o caso no âmbito da escolha por modo de viagem, considera-se a Equação 2.7 como um exemplo simplificado de índice de utilidade da alternativa de modo j, percebido pelo usuário i, dependente dos atributos da alternativa (RAPIDEZ e PREÇO) e dependente dos atributos sócio-econômicos, como por exemplo a renda familiar e o número de dependentes na família (RENDA e N°_DEPENDENTES):

$$U_{ji} = e + \alpha \text{RAPIDEZ} - \beta \text{PREÇO} + \gamma \text{RENDA/N°_DEPENDENTES} \qquad (2.7)$$

onde e é o componente aleatório devido aos atributos não considerados e à percepção do usuário; e α, β, γ e δ são coeficientes constantes determinados. A noção de rapidez, representada aqui por um atributo de viagem, se refere ao tempo gasto para a transposição de uma distância independentemente da velocidade desenvolvida pela alternativa de modo escolhida durante o trajeto. Para a avaliação da utilidade de uma alternativa de modo de viagem através da equação acima, pode-se supor que é indiferente usar um modo caro e rápido ou um modo barato e lento, ou até supor que um alto número de dependentes na receita doméstica poderia ser "compensado" pela rapidez daquele modo de viagem (!). Seja como for, em uma função aditiva ou multiplicativa está implícito que uma desvantagem decorrente de um atributo pode ser compensada com uma vantagem decorrente de outro, o que conforma um caráter compensatório aos modelos que utilizam tais funções para estimativa da utilidade das alternativas de escolha.

A figura abaixo apresenta um esboço da estrutura de decisão subjacente aos modelos compensatórios.

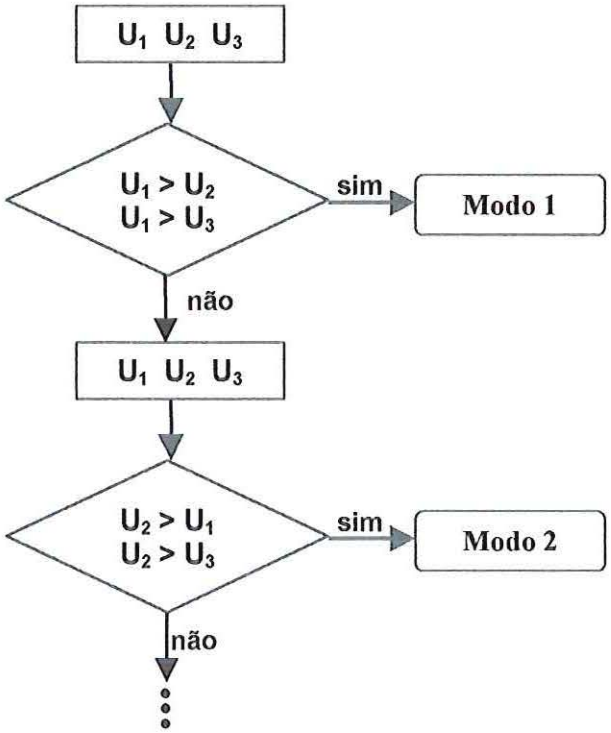


Figura 2.1: Estrutura do processo de decisão de modelos compensatórios

A estrutura de decisão acima mostra que o indivíduo escolhe a alternativa de máxima utilidade percebida. Para tanto, dados os índices de utilidade referentes aos modos, compara-se o índice ou a utilidade de um modo em relação às dos demais; se a utilidade de um modo for maior que as demais, então este modo é escolhido, senão passa-se à comparação do modo seguinte, repetindo-se o processo até que um modo seja escolhido.

Apesar de os modelos de utilidade, Estrita ou Aleatória, principalmente o último, representarem o estado da prática na modelagem de demanda por viagem, alguns pesquisadores (ver TVERSKY, 1972; FOERSTER, 1979; GOLOB & RICHARDSON, 1981; e BARNES & SHEPPARD, 1992) têm sugerido que tais modelos não são apropriados para análises do comportamento de viagem por não considerarem as restrições ao comportamento de viagem, os processos seqüenciais de decisão e a decisão de caráter não-compensatório.

Certamente os modelos compensatórios não refletem a realidade relacionada ao comportamento de viagem daqueles que estão submetidos a uma condição extrema, seja em termos de falta de tempo ou de falta de dinheiro. Infelizmente, um país em desenvolvimento geralmente apresenta um enorme contingente de pessoas mal remuneradas que não se dispõem a pagar pelo conforto adicional, simplesmente pela impossibilidade de fazê-lo (KAWAMOTO, 1987). Assim, não é por acaso que GAUDRY (1981) encontrou em suas observações uma porcentagem residual de pessoas que não abandonam um modo de viagem mesmo piorando muito as condições dele. Esse fato pode estar indicando que o processo subjacente à escolha é não-compensatório.

A despeito do fato de que a maior parte das pesquisas sobre escolha desenvolvida em transportes tenha sido conduzida através de modelos compensatórios, pesquisas em outras áreas tais como Psicologia, Marketing e Administração encontraram evidências empíricas de que, em determinadas situações, o modelo que melhor se ajusta ao comportamento de escolha observado não segue o processo de decisão compensatório, e sim o não-compensatório, o qual será apresentado no item seguinte.

Uma retrospectiva dos últimos 30 anos a respeito do desenvolvimento do modelo de utilidade aleatória, bem como do modelo MNL, desde seu surgimento pode ser encontrada em MCFADDEN (2000).

2.1.2 MODELOS NÃO-COMPENSATÓRIOS

O processo de decisão não-compensatório foi inicialmente apresentado por TVERSKY (1972), e baseia-se na idéia da "eliminação por aspectos". Nessa teoria, cada alternativa é vista como um conjunto de aspectos, ou atributos, e em cada estágio do processo um aspecto é selecionado e todas as alternativas que não o incluem são eliminadas. O processo continua até restar, em tese, apenas uma alternativa, que será a escolhida.

A relevância dos modelos não-compensatórios, baseados na idéia da comparação atributo por atributo, justifica-se pela possibilidade que eles apresentam para a explicação de casos onde o comportamento de escolha se dá

de forma não compensatória. Pode-se observar no dia a dia que por mais confortável e rápido que seja um veículo, se o seu custo de utilização é alto, os indivíduos que recebem baixa remuneração não poderão utilizá-lo. Assim, os modelos não-compensatórios, comparativamente aos modelos compensatórios, descrevem o comportamento de escolha das pessoas com maior realismo.

Embora os princípios básicos de racionalismo e hedonismo implícitos nos modelos de utilidade (compensatórios), e os atributos considerados na análise de alternativas sejam os mesmos, a maior capacidade de explicação dos modelos não-compensatórios relacionada a alguns comportamentos de escolha observados (ver FOERSTER, 1979) deve-se à simples suposição de que a comparação é estabelecida entre atributos e não entre alternativas, ou objetos, como um todo.

Alguns modelos não-compensatórios foram relacionados por FOERSTER (1979). Para efeito de exemplificação alguns destes modelos encontram-se descritos abaixo.

Modelos de Escolha Lexicográfica representam os casos em que os atributos das alternativas são ordenados hierarquicamente. O atributo preponderante será o primeiro na escala de hierarquia, sendo a escolha determinada por este, independentemente dos demais atributos. Havendo empate entre alternativas à escolha em relação ao atributo mais importante, o desempate será feito através do atributo que estiver em segundo lugar na escala de hierarquia, e assim por diante. Este é, segundo MOLDAU (1985), o esquema que o homem segue ao satisfazer suas necessidades; esse comportamento de escolha deve-se, conforme o autor, à existência de objetivos irredutíveis. Entre outros casos identificados como processos de escolha lexicográfica, ALEXIS et al. apud FOERSTER (1979), curiosamente, identifica o caso da escolha de roupas por uma mulher. A figura a seguir apresenta a estrutura de decisão subjacente a esse modelo.

ALEXIS, M.; HAINES, G.H.; SIMON, I. Consumer information processing: the case of women's clothing. In: KING, R. L., ed. *Marketing and the new science of planning.* Chicago, American Marketing, p.197-205, 1968 apud FOERSTER (1979) p.20.

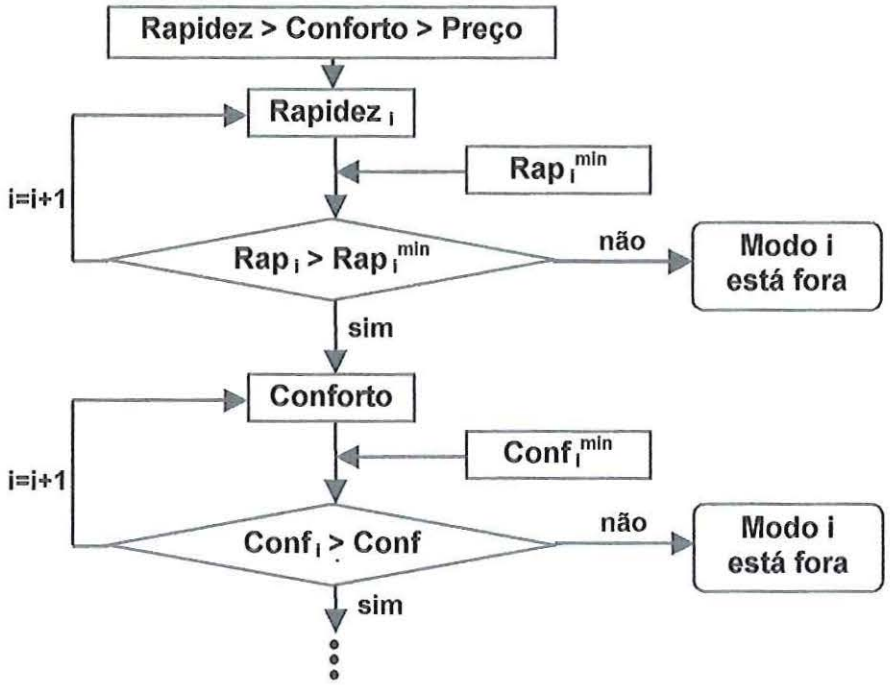


Figura 2.2: Estrutura do processo de decisão do modelo não-compensatório de Escolha Lexicográfica.

A estrutura de decisão acima mostra que os atributos foram hierarquicamente ordenados sendo que o mais importante atributo posiciona-se à esquerda. Então, é verificado se a alternativa satisfaz a condição mínima para este atributo. Se satisfizer, então passa-se à verificação em relação ao segundo atributo mais relevante e, caso contrário, a alternativa já está descartada.

Modelos Conjuntivos procuram representar o processo de escolha em que o tomador de decisão estabelece o nível mínimo para cada um dos atributos. Conseqüentemente, as alternativas para serem consideradas como tais devem possuir todos seus atributos mais interessantes do que aquele mínimo. Esse processo decisório, também chamado por Foerster como "chamada à rejeição", rejeita qualquer alternativa que não satisfaça o nível mínimo de aceitabilidade considerando-se todos os atributos envolvidos. Desta forma, este modelo não está propriamente orientado para escolher uma alternativa, e sim para definir um

conjunto de alternativas viáveis. GOLDBERG apud FOERSTER (1979) encontrou que diagnósticos de clínica médica freqüentemente são baseados nesse processo conjuntivo.

Já os Modelos Disjuntivos, ao contrário dos Conjuntivos, aceitam quaisquer alternativas que possuam pelo menos um atributo que ultrapasse o valor mínimo estabelecido como satisfatório. Esse modelo também está orientado para a definição de não apenas uma, porém de um conjunto de alternativas viáveis, a menos que apenas uma das alternativas satisfaça a condição mínima. Segundo KAWAMOTO (1987), em transporte é comum encontrar casos em que a escolha é sujeita a restrições, quer seja de horário, como por exemplo o caso do trabalhador que pretenda almoçar em casa. Para que isso seja possível é necessário que a rapidez do modo de viagem seja maior que um determinado valor.

Há ainda outros tipos de modelos não-compensatórios como por exemplo, os Modelos de Dominância, os Modelos Maximax e os Maximin. Para maiores detalhes a respeito destes modelos ver FOERSTER (1979).

Em transporte, a utilização desses modelos, não-compensatórios, em cidades de grande porte que apresentam um grande número de modos disponíveis, incluindo-se os modos combinados, implica uma considerável restrição, uma vez que tal processo nem sempre converge para um único modo, dada a possibilidade de ocorrência de modos com características semelhantes. Após aplicarem o modelo não-compensatório, no contexto de transporte, RECKER & GOLOB (1979) acreditam que possivelmente a maioria dos processos de decisão contêm ambos os elementos, compensatório e não-compensatório. Esses autores sugerem que seria utilizado o paradigma não-compensatório nas fases iniciais do processo de decisão para a redução da lista de alternativas, para então ser aplicado o compensatório. Como comentou KAWAMOTO (1987), embora essa idéia pareça interessante, não considera o fato de que a discrepância entre a realidade e o resultado de previsão, obtido através de modelos compensatórios, foi grande também nos casos onde existiam apenas duas alternativas (ver GAUDRY, 1981).

GOLDBERG, L.R. Five models of clinical judgement: an empirical comparison between linear and non-linear representations of human inference processes. *Organizational Behavior and Human Performance,* v.6, p.458-479, 1971 apud FOERSTER (1979) p.21.

Os modelos não-compensatórios não consideram o fato de que cada usuário estabelece o gasto máximo com o item transporte, de acordo com a sua possibilidade financeira, a qual depende, entre outros, da sua renda. Caso esse fato fosse considerado no modelo, ele deixaria de ser um modelo não-compensatório, uma vez que se estaria admitindo a possibilidade de compensação entre o gasto máximo com o item transporte e a renda.

2.1.3 MODELO SEMI-COMPENSATÓRIO

É um modelo proposto por Kawamoto (1987) após reconhecer as limitações dos modelos compensatórios e não-compensatórios.

O pressuposto fundamental do modelo semi-compensatório é que dentre um conjunto de modos de viagem disponíveis, o usuário deseja viajar de modo intrinsecamente mais útil, isto é, de alternativa mais rápida, mais confortável e mais segura. É a manifestação da natureza hedonista do homem, isto é, a tendência do homem em buscar o bem estar, no sentido amplo da palavra, afastando-se do sofrimento. Esse comportamento independe da situação sócio-econômica do indivíduo. Assim, nesse modelo o termo utilidade ou a utilidade intrínseca do modo de viagem tem o sentido de qualidade que uma alternativa de transporte possui para atender à ansiedade natural de uma pessoa de chegar ao destino, da maneira mais rápida e mais confortável possível.

Por outro lado, com exceção de caminhada e carona, o uso dos modos geralmente está associado a um custo monetário. O dinheiro que se gasta pelo serviço de transporte tem, por sua vez, um valor subjetivo, uma utilidade como objeto de troca para satisfazer a necessidade de transposição de uma distância ou um desejo de chegada ao destino. Assim, o princípio da racionalidade humana está presente nesse modelo na medida em que seu processo de escolha supõe que o homem seja capaz de perceber quais alternativas são mais úteis que as respectivas quantias de dinheiro que seriam desembolsadas pelo seu uso, e supõe também que o homem saiba discernir qual alternativa proporciona o maior bem estar.

O modelo semi-compensatório explora exatamente estes dois pontos. Em primeiro lugar estabelece que a utilidade percebida da modalidade de viagem num

dado contexto depende da rapidez e do conforto físico e segurança que ele oferece, e que todas as pessoas percebem essa utilidade de forma semelhante, conforme um padrão representado por uma dada expressão matemática. Em segundo lugar propõe que a utilidade atribuída pelo homem ao custo monetário percebido, associado ao uso do modo de viagem, é uma função de sua situação sócio-econômica. Têm-se aí duas utilidades: o modo de viagem rápido e confortável, e o dinheiro, que talvez possa ser aplicado em algo mais útil do que transporte. Caso a utilidade do modo de viagem seja maior do que a utilidade da quantia de dinheiro correspondente ao uso do modo, então esta modalidade de viagem será viável.

Pode-se supor, por exemplo, que a utilidade intrínseca da alternativa de modo de viagem j seja quantificada da seguinte forma:

$$U^{I}{}_{j} = a \cdot \text{DISTÂNCIA}^{b} \cdot \text{RAPIDEZ}^{c} \cdot \text{CONFORTO}^{d} \tag{2.8}$$

onde a, b, c e d são constantes determinadas; e que o índice de utilidade atribuída por um indivíduo ao dinheiro seja estimado através da seguinte expressão:

$$U^{\$}{}_{j} = \alpha \cdot \text{PREÇO}^{\beta} \cdot \text{RENDA}^{\gamma} \cdot \text{N°_DEPENDENTES}^{\delta} \tag{2.9}$$

onde α, β, γ e δ são constantes determinadas. De acordo com a estrutura de decisão subjacente ao modelo semi-compensatório, esboçada abaixo na Figura 2.3, um indivíduo inicia o processo de escolha modal comparando as utilidades intrínsecas dos diversos modos alternativos que lhe estão disponíveis, procurando identificar o modo de maior utilidade intrínseca, como por exemplo o modo 1. Em seguida, procura verificar a viabilidade de uso do modo 1 comparando a utilidade intrínseca deste serviço com a utilidade do dinheiro a ser desembolsado ao usá-lo. Caso a utilidade do modo seja maior do que a do dinheiro, o modo será considerado viável e consequentemente escolhido. Caso contrário, significa que a troca do dinheiro pelo serviço não é vantajosa, ou ainda, que o modo será percebido como caro e não será usado. Conseqüentemente, o indivíduo passará a analisar o modo 2, e assim por diante.

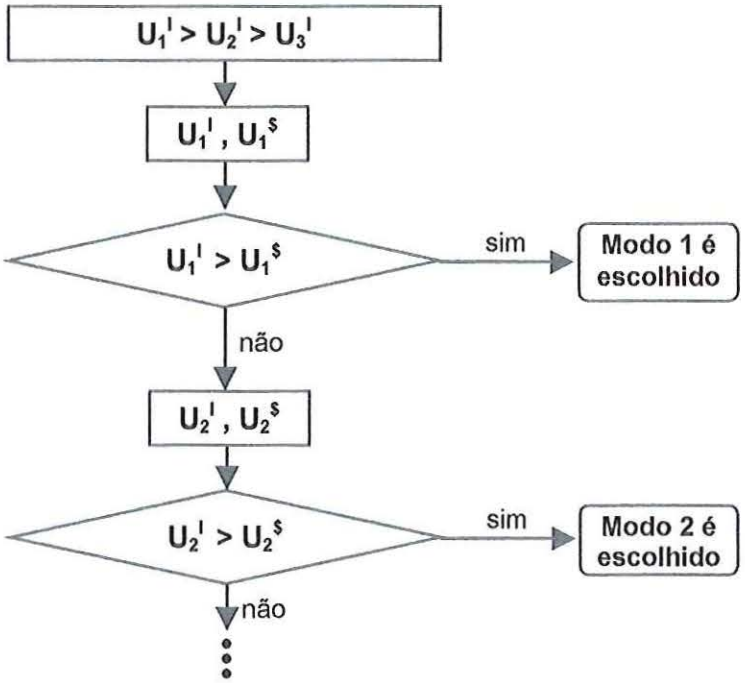


Figura 2.3: Estrutura do processo de decisão do modelo semi-compensatório

Segundo KAWAMOTO (1987), a grande diferença entre as suposições implícitas nos modelos de escolha convencionais e o modelo semi-compensatório é que enquanto aquelas supõem que a percepção da utilidade dos modos de viagem mudam conforme a condição sócio-econômica, este supõe que todas as pessoas percebem a utilidade de forma essencialmente igual. Imagina-se que as pessoas pobres viajam de ônibus ou a pé, não porque sentem menos o sofrimento do que os mais abastados ou porque o desejo de viajar de modo rápido e confortável seja mais tênue nas pessoas mais pobres. Todos gostariam de viajar de modo mais rápido e mais confortável. A questão é se a sua situação sócio-econômica permite a utilização do modo desejado.

A concepção do modelo semi-compensatório afasta-se da idéia de que o usuário pode sempre considerar, por exemplo, o tempo e o custo monetário compensatórios entre si, isto é, que o aumento no custo pode sempre ser sentido como compensado pela diminuição no tempo de viagem. Sabe-se que para a população de baixa renda, um aumento considerável no preço da passagem de ônibus pode torná-lo proibitivo, e que não haveria diminuição no tempo de viagem

que pudesse resgatar a viabilidade do modo ônibus para ela. Portanto, o modelo semi-compensatório descarta a compensatoriedade generalizada entre atributos de viagem.

Ademais, a abordagem desse modelo permite a explicação qualitativa de muitos comportamentos do usuário de transporte, inclusive daqueles não explicáveis pelos modelos convencionais, como por exemplo o comportamento não compensatório e descontínuo. Como é o entendimento de alguns autores, o processo convencional descreve o comportamento de escolha do indivíduo através da probabilidade, o que significa que um indivíduo com dadas condições sócio-econômicas necessariamente utilizaria, em maior ou menor grau, todas as modalidades de transporte disponíveis.

Maiores detalhes sobre esse modelo podem ser encontrados em KAWAMOTO (1987), KAWAMOTO (1989) e KAWAMOTO (1994). Um comparativo de desempenho entre o modelo convencional padrão (MNL) e o modelo semi-compensatório encontra-se descrito em KAWAMOTO (1990), e um procedimento de calibração em KAWAMOTO & SETTI (1992).

2.2 CONSIDERAÇÕES FINAIS

Inúmeras aplicações em demanda por transporte podem ser encontradas na literatura envolvendo os modelos aqui apresentados. Em particular, a utilização dos modelos compensatórios, baseados na teoria da utilidade aleatória, difundiu-se além dos limites do transporte para outras áreas nas quais era desejável a previsão do comportamento de escolha do indivíduo, tais como pesquisa de mercado para bens de consumo (BERKOVIC, 1985), previsão de demanda por eletricidade (MCFADDEN, 1986), estudo de demanda para investimento em serviço de telefonia (BEN-AKIVA, MCFADDEN & TRAIN, 1987) entre outras. Devido ao pioneirismo no desenvolvimento desses modelos de utilidade aleatória (por ex.: o modelo MNL), Daniel L. McFadden recebeu da *Royal Swedish Academy of Sciences* o prêmio Nobel em economia no ano 2000, citado pela Academia em merecimento por "teorias e métodos vastamente usados na análise estatística do comportamento individual" (HIRSCH, 2000).

Embora esses modelos dominem o estado da prática, há um consenso entre pesquisadores de que ainda não se alcançou um estado da arte na modelagem comportamental (MCFADDEN, 2000). A esse respeito, o autor afirma que o desenvolvimento de um modelo comportamental unificado de demanda por transporte não foi completado e que os modelos atuais apoiam-se pesadamente em métodos de calibração de engenharia e não representam um refinamento cumulativo de estudos comportamentais. De acordo com PIPKIN (1995), os modelos de escolha baseados na teoria da utilidade contém fortes e não bem examinadas suposições de que pessoas são calculadores egoístas retratados em teoria econômica elementar. BARNES & SHEPPARD (1992) apontam que a dominância nos estudos em transporte dos modelos de escolha baseados na teoria da utilidade pode ser vista como uma enorme inconsciente atitude ideológica que pratica um fetiche de racionalidade (!).

Uma abordagem alternativa para a modelagem da escolha em transporte desenvolveu-se enfocando o padrão de viagens do domicílio como um todo. Esta abordagem é conhecida como Modelos Baseados em Atividades[1] (*Activity Based Models*) e não será apresentada neste texto. Para maiores detalhes sobre essa forma de modelagem ver RECKER et al. (1986a, 1986b), KURANI & LEE-GOSSELIN (1997) e KITAMURA (1997).

A pesquisa em modelagem de escolha, alimentada por novas pesquisas de teoria de decisão comportamental[2], continua a desenvolver modelos mais refinados, tais como o Modelo de Variáveis Latentes que combina dados reais observados e hipotéticos na tentativa de incorporar construtos psicológicos não observáveis, como por exemplo atitude, percepção e motivação, para melhoria de performance predictiva (MCFADDEN, 1986; MORIKAWA & MCFADDEN, 1986; BEN-AKIVA et al., 1999 e MCFADDEN, 2000).

Posteriormente à apresentação do modelo semi-compensatório (KAWAMOTO, 1987) como uma abordagem ao processo de escolha com tratamento baseado na psicofísica multidimensional, KAWAMOTO (1994) chama novamente a atenção para a desigualdade entre o hedonismo e o racionalismo no

[1] Também conhecida como *Constraint-oriented Approach* ou *Complex Travel Behaviors*.
[2] Veja BEN-AKIVA etal. (1999) para exemplos de pesquisas em teoria da decisão (Behavioral Decision Theory – BDT).

processo de escolha, desta vez derivando o modelo semi-compensatório a partir da teoria microeconômica tradicional, baseando-se na hipótese da inclusão do conceito de bens inferiores e normais na formulação do comportamento do usuário do sistema de transporte. Pela ocasião o autor afirmava que no estágio atual do desenvolvimento da teoria do comportamento de consumidor e de usuários de transporte, ainda não estava claro quais e como os atributos influenciavam na percepção.

Para ilustrar a situação do caminho percorrido até meados da década passada, PIPKIN (1995) afirma que o estudo sobre a escolha do modo de viagem permanece importante para o transporte nos dias atuais e que a questão do paradigma de escolha ainda está viva.

3 MINERAÇÃO E DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS

Nas últimas duas décadas houve um aumento explosivo da capacidade de gerar e coletar dados. Avanços na coleta de dados para uso científico, a vasta utilização de códigos de barras para produtos comerciais, a computadorização de negócios (ex.: compras com cartão de crédito) e de transações governamentais (ex.: recolhimento de impostos) geraram uma avalanche de dados que em associação a avanços na tecnologia de armazenamento criaram "montanhas" de dados armazenados.

Nos dias atuais, bancos de dados de vários *gigabytes* ou *terabytes* de informação são comuns (HAND, 1998). Exemplificando, a empresa norte-americana de telefonia AT&T concentra mais de 100 milhões de clientes e registra mais de 200 milhões de telefonemas num único dia (CORTES & PREGIBON apud HAND, 1998). A também norte-americana e varejista Wal-Mart executa 20 milhões de transações diariamente só no território de seu país (BABCOCK, 1994). Segundo FAYYAD et al. (1996) o sistema de observação da Terra da NASA gera 50 gigabytes por hora desde o ano 2000.

Tais volumes de dados armazenados suplantaram as técnicas estatísticas convencionais de análise de dados, como menciona HAND (1998). A necessidade de uma nova geração de técnicas e ferramentas, capazes de auxiliar a análise dessas montanhas de dados à procura de um conhecimento útil, fez emergir um novo campo de pesquisa chamado Descoberta de Conhecimento em Banco de Dados ou *Knowledge Discovery Databases* (KDD). Esse termo, KDD, foi cunhado em 1989 para referir-se ao amplo processo de busca de conhecimento em dados. Dentro desse contexto, o termo *data mining* (DM), ou mineração de dados, refere-se a um único passo dentro do processo de KDD. Essa interpretação foi

formalizada no primeiro *International Conference on KDD* realizado em 1995 em Montreal e é adotada por pesquisadores em Inteligência Artificial e Aprendizado de Máquina (FAYYAD et al., 1996). Desta forma, KDD é o processo de extração de informação útil a partir de dados e que envolve várias etapas, entre elas a de *data mining* que por sua vez é o conjunto de técnicas usadas em uma abordagem automatizada para explorar exaustivamente e trazer à superfície relações complexas e escondidas ou não suspeitas em dados (MOXON, 1996). É comum encontrar na imprensa popular e também em Estatística a expressão *data mining* empregada com o significado do processo de KDD, fato esse resultante do aumento de vendas de pacotes de software que incluem algoritmos de *data mining* (CABENA et al., 1997).

Em geral, o processo de KDD é constituído dos passos ilustrados na Figura 3.1.

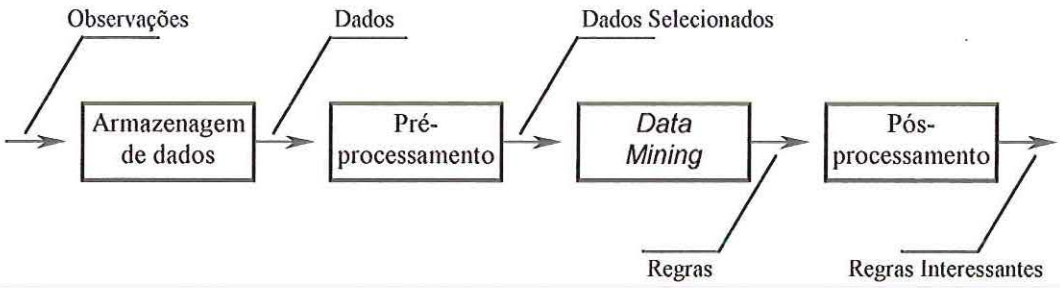


Figura 3.1 Uma visão geral do processo de KDD

Observações provindas de diferentes fontes e em diferentes formatos são coletadas e armazenadas como dados. Um pré-processamento envolve a limpeza nos dados como por exemplo a remoção de ruídos ou dados incompletos, a seleção dos dados e a formatação de acordo com a aplicação. O passo seguinte, *data mining*, inclui a escolha do algoritmo apropriado e regras (conceitos) são aprendidas (aqui regras como termo genérico representando o que é minerado e descoberto). Em um pós-processamento as regras são avaliadas e interpretadas em termos de sua aplicabilidade e compreensão, uma vez que freqüentemente o

CORTES, C.; PREGIBON, D. *Mega-monitoring*. Washington, University of Washington / Microsoft Summer Research / Institute on Data Mining, 1997 apud HAND (1998) p.112.

número de regras é grande e algumas são mais "interessantes"[1] do que outras (LIU & MOTODA, 2000).

Em termos de tempo e esforço consumidos, nem todos os passos desse processo são iguais. A preparação dos dados envolvendo a seleção, limpeza e formatação pode demandar mais de 60% do tempo utilizado em todo o processo, enquanto a parcela de tempo gasto no passo de mineração tipicamente constitui apenas 10% do tempo utilizado em todo o processo(CABENA et al., 1997). Para maiores detalhes sobre as etapas de um processo de KDD ver BRACHMAN & ANAND (1996) e FAYYAD et al. (1996).

A tarefa da etapa de *data mining* nesse processo é conhecida como análise *secundária* de dados. Entende-se por análise *primária* de dados a coleta de dados com uma questão particular ou um conjunto de questões particulares *a priori* que podem ser traduzidas em forma de hipóteses a serem testadas – planejamento de experimentos e amostragem são dois exemplos que tiveram seu desenvolvimento voltado para esse tipo de análise. Por outro lado, a análise secundária procura relações não suspeitas implícitas nos dados. Assim, enquanto a análise estatística convencional tem como base um processo hipotético-dedutivo, a abordagem dirigida a dados das técnicas de *data mining* se configura como um processo indutivo (HAND apud HAND, 1998).

Entre as várias técnicas de *data mining* estão a regressão e classificação, análise de associação, análise de cluster, indução de árvores de decisão, visualização e redes neurais artificiais. Na seção seguinte será abordada a determinada técnica de redes neurais artificiais a ser utilizada como motor minerador neste trabalho.

3.1 REDES NEURAIS ARTIFICIAIS

Existem vários diferentes tipos (modelos) de redes neurais artificiais (RNA), cada um indicado para uma diferente tarefa. O tipo de RNA empregado neste estudo vem a sem o mais indicado na literatura para a tarefa de classificação,

[1] No capítulo 4, subitem 4.7.1 "Interpretação e Análise", encontra-se a definição do vem a ser "interessante" para fins deste estudo.
HAND, D.J. (s.d.). Statistics and the science method. To appear in *enciclopedia of biostatistics.* Chiscester, Wiley. /no prelo/ apud HAND (1998) p.114.

sendo é conhecido como feed-forward *multilayer perceptron* (MLP). Segue-se uma breve revisão de suas características. Foge ao foco deste estudo um detalhamento da teoria de redes neurais artificiais uma vez que inúmeros títulos na literatura tratam desse assunto, para citar alguns: HERTZ et al. (1991), BEALE & JACKSON (1992), CARVALHO et al. (1999) e HAYKIN (1999)[2].

Como ilustrado na Figura 3.2, a rede neural do tipo MLP é composta de unidades de processamento ("neurônios") ou nodos dispostos em camadas e interligadas por conexões ("sinapses") associadas a pesos que ponderam a entrada recebida por cada unidade da rede. O estado de uma unidade em um dado tempo é representado por sua ativação, que é um valor numérico real. Na camada de entrada, a ativação das unidades representam os valores recebidos dos atributos de entrada do domínio ao qual a rede neural está sendo aplicada. As unidades da camada de saída representam as decisões tomadas pela rede. Posicionadas entre as unidades da camada de entrada e as de saída pode haver um determinado número de camadas intermediárias de unidades. A tarefa das unidades intermediárias é transformar o espaço de entrada em outro espaço no qual seja mais apropriado para as unidades de saída fazerem discriminações (CRAVEN, 1996). VERLEYSEN (1998) interpreta as unidades intermediárias como sendo

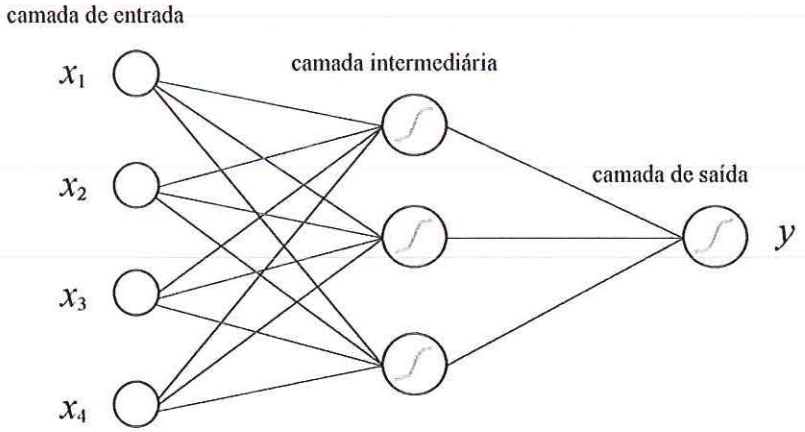


Figura 3.2: Uma rede neural MLP com 4 unidades de entrada, 3 unidades intermediárias e uma unidade de saída (4-3-1).

[2] Redes neurais artificiais (RNA) está neste texto sempre se referindo ao tipo de rede MLP, exceto quando for discriminado outro tipo.

"graus de liberdade" adicionados à estrutura do modelo de maneira a se alcançar uma solução.

O computar nessa MLP se inicia com a ativação das unidades de entrada por uma dada instância do domínio em questão. Essa ativação de entrada alimenta à diante, através das conexões ponderadas por pesos, as unidades da camada intermediária e então as unidades da camada de saída. A resposta fornecida pela rede é determinada pela ativação resultante nas unidades de saída. Por exemplo, a soma ponderada de entrada (*net*) para uma unidade em uma camada intermediária ou de saída é dada por:

$$net_i = \sum_j w_{ij} a_j + \theta_i \tag{3.1}$$

onde w_{ij} é o peso da conexão entre a unidade à montante j para a unidade i; a_j é a ativação provinda da unidade j; e θ_i é o parâmetro ajustável chamado *bias* da unidade i que pode ser interpretado como a predisposição de uma unidade em relação à ativação. A ativação de uma unidade intermediária ou de saída é determinada pelo passar de sua soma ponderada de entrada através de uma função de transferência, às vezes chamada de função de ativação. Usualmente, a função de transferência utilizada em redes MLP é a função logística (Equação 3.2) que conforma a *net* da unidade para um valor de ativação possível dentro do intervalo [0,1].

$$a_i = \frac{1}{1 + e^{-net_i}} \tag{3.2}$$

O conhecimento adquirido (aprendido) sobre o domínio pela rede MLP está, portanto, armazenado na forma com que as unidades estão conectadas, em sua arquitetura, e em seus parâmetros ajustáveis, pesos de conexões e *bias* de unidades. Na rede MLP utilizada neste trabalho, o conhecimento é adquirido por meio de um processo de treinamento supervisionado. No treinamento supervisionado é apresentado para a RNA um conjunto de instâncias na forma $\langle \vec{x}, y \rangle$ onde $\vec{x}$ representa o vetor dos valores dos atributos de entrada e y é o

correspondente valor desejado da variável de saída. A saída desejada é comparada com a saída que a rede forneceu (calculou) e a informação sobre o erro é usada para o ajuste dos parâmetros da rede. A minimização desse erro no processo de treinamento é feita através de um método de otimização de gradiente descendente. De maneira geral, tal método de otimização executa iterativamente dois passos: calcula o gradiente da função de erro a ser minimizada, em relação aos parâmetros ajustáveis da rede, e ajusta esses parâmetros na direção sugerida pelo gradiente. A soma dos erros quadráticos (*SSE*) de todas as saídas fornecidas pela rede (Equação 3.3) é normalmente utilizada como medida de desempenho da rede e também como função de erro a ser minimizada pelo algoritmo de treinamento (CARVALHO et al., 1999 e ZHANG et al., 1998).

$$SSE = \frac{1}{2}\sum_{i=1}^{n}(y_i - o_i)^2 \tag{3.3}$$

onde n é o número de instâncias do conjunto de treinamento; y_i é o valor desejado de saída; o_i é o valor de ativação da unidade de saída; e 1/2 é incluído para simplificar a expressão das derivadas computadas no algoritmo de treinamento.

Vários algoritmos de aprendizagem (treinamento) podem ser usados para minimizar a função de custo (erro) objetivo. Um dos mais populares para o uso em redes MLP é o chamado *backpropagation* (RUMELHART et al., 1986) com atualização *on-line*. Segundo WHITE (1989) o *on-line backpropagation* pode ser visto como uma forma estocástica do gradiente descendente em que as mudanças dos parâmetros ajustáveis não seguem o gradiente da função custo calculado para todo o conjunto de treinamento, mas sim o gradiente calculado para cada instância do conjunto individualmente.

Trocando em miúdos, as redes neurais aprendem através de exemplos e o aprendizado ocorre quando a rede atinge uma solução generalizada para um dado problema. Diz-se então que a rede está treinada. Para atingir essa solução generalizada o treinamento é parado antes que um mínimo local na função custo seja alcançado. O motivo por trás desta técnica de parada do treinamento é evitar que a rede "decore" (*over-fitting*[3]) o conjunto de dados de treinamento. Uma forma

[3] Também referenciado na literatura como *over-training* (CARVALHO et al., 1999).

de estimar o ponto de parada do treinamento é usar um outro conjunto de dados, diferente daquele de treinamento, chamado conjunto de validação para monitoramento da acurácia predictiva da rede enquanto ela está sendo treinada. Assim, ao invés de salvar os parâmetros da rede correspondentes ao mínimo da função custo, esse procedimento salva-os na iteração do método de otimização que resulta a mais elevada acurácia para o conjunto de validação.

3.1.1 Usando Redes Neurais para *Data Mining*

Como já mencionado, o foco central da tarefa de mineração de dados, *data mining*, é obter informação útil, "*insights*", a respeito de um conjunto de dados. Freqüentemente, o atingir desse objetivo envolve a aplicação de métodos de aprendizado de máquina para indutivamente construir-se modelos a partir dos dados disponíveis. Entre tais métodos está a RNA.

Considerando os métodos concorrentes, as redes neurais artificiais provêm o "*bias*[4] indutivo" mais apropriado (CRAVEN & SHAVLIK, 1997). Brevemente falando sobre o significado do termo *bias* indutivo: dado um determinado conjunto de dados, existem infinitos modelos que poderiam adequar-se (ajustar-se) a esses dados e cada algoritmo de aprendizado tem seu *bias* indutivo que determina o modelo resultante dessa adequação. Segundo CRAVEN & SHAVLIK (1997) existem dois aspectos em relação ao *bias* indutivo de um algoritmo: sua parcialidade do espaço de hipóteses restritas e sua parcialidade da preferência. A parcialidade do espaço de hipóteses restritas refere-se às restrições que um algoritmo de aprendizado impõe sobre as hipóteses que ele é capaz de construir. Por exemplo, o espaço de hipóteses de uma rede neural composta de uma única unidade de processamento que implementa uma função de ativação do tipo limiar (função passo), conhecida como *perceptron* (CARVALHO et al., 1999), é limitado às funções discriminantes lineares. Já as redes neurais do tipo MLP com duas camadas intermediárias contendo unidades de processamento que implementam funções de ativação não-lineares (por ex.: função sigmoidal logística) podem aproximar qualquer função (CYBENKO, 1988). Por sua vez, a parcialidade da preferência refere-se à ordem de preferência que um algoritmo de aprendizado

determina sobre o modelo em desenvolvimento considerando seu espaço de hipóteses. Por exemplo, os algoritmos de aprendizado, em sua maioria, inicialmente tentam adequar uma hipótese simples para um determinado conjunto de dados de treinamento e então exploram progressivamente hipóteses mais complexas até encontrar uma aceitável adequação.

Para alguns casos, redes neurais possuem uma parcialidade do espaço de hipóteses mais apropriado que outros algoritmos de aprendizado, como por exemplo a classe de problemas de previsão temporal ou seqüencial. Um tipo de RNA conhecido como redes neurais recorrentes (HOPFIELD, 1982) que é freqüentemente aplicado a essa classe de problemas, é capaz de manter o estado de informação de um passo de tempo ao seguinte.

Em outros casos, prefere-se redes neurais não por causa da classe de hipóteses que elas são capazes de representar, mas simplesmente porque elas induzem hipóteses que generalizam melhor. Muitos estudos empíricos têm apontado que existem domínios em que as redes neurais provém acurácia predictiva superior aos comumente usados algoritmos de aprendizado simbólico (SHAVLIK et al., 1991). A esse respeito, APTÉ & WEISS (1997) também afirmam que redes neurais possuem melhor performance preditiva que algoritmos de indução de árvores de decisão a partir de dados.

Além do fato de possuírem, portanto, um apropriado *bias* indutivo para a tarefa de *data mining* em uma grande variedade de problemas, nas redes neurais todos os atributos (variáveis) de entrada de uma instância contribuem ao mesmo tempo (em cada época de treinamento) para a modificação dos parâmetros ajustáveis. Em outras palavras, nas RNAs os atributos são usados na adequação do modelo em uma forma multivariada (WHITE, 1989; RIPLEY, 1993; ZHANG et al. 1998 e LIU & MOTODA, 2000). Esse fato colabora para a não necessidade do conhecimento *a priori* sobre as relações entre as variáveis de entrada e saída do problema e diferencia as RNAs de algoritmos de indução de árvore de decisão, como por exemplo o algoritmo de indução de árvores C4.5 (QUINLAN, 1993), e de classificadores como o *Naive Bayesian* (DOMINGOS & PAZZANI, 1996), que são

[4] Esta palavra mantida do original em inglês pode ser traduzida como: predisposição ou parcialidade (OXFORD, 2000).

abordagens univariadas no que concerne em como os atributos são usados para a adequação (construção) do modelo.

3.2 EXTRAÇÃO DE CONHECIMENTO DE RNA

Em geral existe uma hierarquia entre dados, informação e conhecimento. No momento em que se atribui um determinado significado aos dados, estes tornam-se uma informação. Quando especialistas no domínio do problema elaboram uma norma ou regra, a interpretação de confronto entre a informação (fato) e essa regra constitui-se um conhecimento (BRAGA et al., 1998).

Apesar de as redes neurais possuírem algumas características favoráveis à utilização em *data mining*, como anteriormente descrito, uma das principais críticas a seu respeito é a sua incapacidade de explicar como e porque uma determinada rede neural treinada gera suas respostas, sendo assim comumente chamada de "caixa preta" (VERLEYSEN, 1998; ZHANG et al., 1998; HAYKIN, 1999; CARVALHO et al., 1999 e TOWELL & SHAVLIK, 1993). Isto porque ela revela que descobriu algo mas não diz o quê. Seria conveniente como ferramenta de *data mining* que a capacidade de explanação se tornasse parte integrante da funcionalidade das redes neurais, explicitando o conhecimento adquirido sobre seu domínio de aplicação.

Uma hipótese ou o conhecimento aprendido por uma RNA que está definido por sua arquitetura, funções de ativação usadas pelas unidades de processamento e parâmetros ajustáveis é difícil de ser compreendido por várias razões. Uma rede típica tem centenas e até milhares de parâmetros em valores numéricos reais. Esses parâmetros codificam as relações entre os atributos de entrada, $\vec{x}$, e a saída alvo, $\vec{y}$, e embora não seja difícil de entender a relação codificada em um único parâmetro, a típica grande quantidade desses é que dificulta a compreensão. Além do mais, em uma MLP esses parâmetros podem representar relações não-lineares não-monotônicas entre os atributos de entrada e a(s) saída(s) alvo. Assim, usualmente não é possível determinar isoladamente o efeito de um dado atributo de entrada sobre a saída, uma vez que esse efeito pode ser mediado pelos valores de outros atributos.

Em uma rede, essas relações não-lineares são representadas pelas unidades de processamento intermediárias, as quais combinam as entradas de múltiplos atributos, permitindo assim que o modelo tire vantagem das dependências entre os atributos. Em última análise, as unidades intermediárias podem ser vistas como representando atributos derivados. De acordo com HINTON (1986) apud CRAVEN (1996), entender as unidades intermediárias individualmente é uma tarefa difícil pelo fato de elas aprenderem representações distribuídas. Em uma representação distribuída, unidades intermediárias individuais não correspondem aos conhecidos atributos do domínio. Ao invés disso, os atributos, que venham a demonstrar-se importantes no domínio do problema, são codificados por padrões de ativação através de várias unidades intermediárias. Similarmente, cada unidade intermediária executa uma parte no representar de vários atributos derivados.

Uma abordagem para o entendimento do conhecimento representado por uma RNA treinada é traduzi-lo em uma linguagem mais compreensível. Os métodos que usam esse paradigma se diferem em três diferentes dimensões: a linguagem de representação, os requisitos à rede e a estratégia de extração (CRAVEN & SHAVLIK, 1997).

A linguagem de representação é a forma utilizada pelo método de extração para a descrição do modelo aprendido pela rede neural. Algumas linguagens utilizadas por vários métodos incluem regras de inferência conjuntiva (se-então), regras m-de-n e árvores de decisão, entre outras.

Os requisitos à rede referem-se às restrições em termos de arquitetura e treinamento impostas pelo método, ou seja, referem-se ao intervalo de redes ao qual o método é aplicável.

A estratégia de extração refere-se à maneira pela qual o método explora um espaço de descrições candidatas e qual o nível de descrição ele usa para caracterizar uma dada RNA, separando os métodos em duas classes: local e global. A abordagem local inclui os métodos que extraem regras que descrevem cada unidade intermediária e de saída em termos das conexões com unidades antecessoras, eventualmente decompondo a rede MLP num conjunto de redes de uma única camada. A vantagem dessa abordagem, segundo BRAGA et al. (1998), é que ela produz descrições mais simples, porém com a desvantagem de as regras

extraídas não fornecerem uma representação muito apurada da rede. Entre os algoritmos propostos para a extração de conhecimento de RNA por esta abordagem local, estão o SUBSET (TOWELL & SHAVLIK, 1993), o M-of-N (TOWELL & SHAVLIK,1993) e o FERNN (SETIONO & LEOW, 2000). Esses algoritmos possuem algumas limitações significativas, tais como baixa generalização e escalabilidade, por requererem um procedimento de treinamento especial ou pela imposição de alguma restrição na arquitetura da rede, ou ainda por restringir o uso de funções de ativação nas unidades da rede. Por sua vez, a abordagem global extrai um conjunto de regras que descrevem o comportamento das unidades de saída diretamente em termos das unidades de entrada. Os algoritmos dessa abordagem global são o VIA (TRHUN, 1995) e o Trepan (Craven & SHAVLIK, 1996).

Uma revisão detalhada de vários tipos de métodos de extração pode ser encontrada em CRAVEN (1996) e outras referências a esse respeito em BRAGA et al. (1998). Na seção seguinte é apresentado o método utilizado neste trabalho, o algoritmo Trepan.

3.2.1 Algoritmo Trepan

O algoritmo Trepan (Trees *parroting network*) (CRAVEN & SHAVLIK, 1996) foi desenvolvido para resolver o problema de extrair uma hipótese compreensível a partir de uma rede neural treinada como uma tarefa de aprendizado indutivo. Nessa tarefa, o conceito alvo é a função representada por uma rede treinada e a hipótese produzida pelo algoritmo de aprendizado é a árvore de decisão que aproxima essa rede.

Alguns métodos de extração de regras de redes neurais possuem aplicabilidade limitada por requererem processos de treinamento especiais como é o caso do algoritmo M-of-N (TOWELL & SHAVLIK,1993). Outros métodos são limitados na aplicabilidade porque impõem restrições sobre a arquitetura de rede, como por exemplo os métodos RULEX (ANDREWS & GEVA, 1995) e FERNN (SETIONO & LEOW, 2000), ou restrições sobre a função de ativação (transferência), como por exemplo o método KT (FU, 1994).

O Trepan difere-se de outros métodos de extração de regras por não testar diretamente regras hipotéticas contra a rede, e por não traduzir unidades de processamento individuais, intermediárias ou de saída, em regras. Entretanto, o processo de extração do Trepan envolve o refinar progressivo de um modelo a respeito de toda a rede (estratégia de extração global). Neste caso o modelo é a árvore de decisão que é construída através de uma busca heurística do melhor caso[5].

O algoritmo Trepan constrói árvores de decisão pelo particionar recursivo do espaço do conjunto de dados de entrada. Em tais árvores de decisão (Figura 3.3), cada nó interno representa um critério de divisão que particiona o espaço de entrada de maneira a melhorar a separação das instâncias de diferentes classes e cada folha, por sua vez, representa uma classe predita.

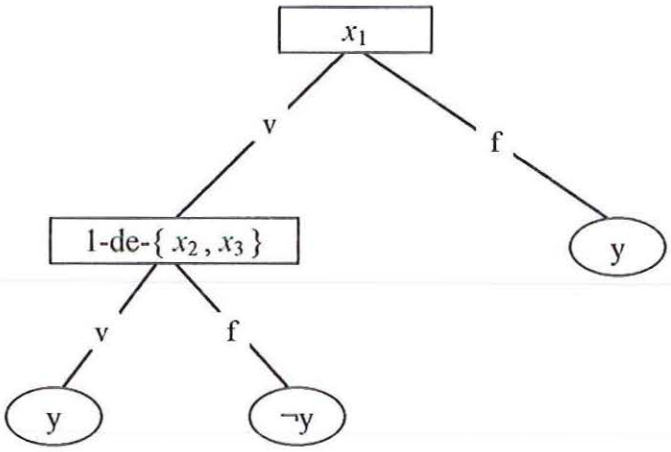


Figura 3.3: Exemplo de árvore de decisão de nível de profundidade igual a três. O nó chamado "raiz" é o localizado no alto e neste caso contém apenas o atributo x_1. Para se alcançar a folha da classe y localizada no extremo esquerdo dessa árvore sujeita-se às restrições: $(x_1 = \text{verdadeiro}) \wedge (x_2 = \text{verdadeiro} \vee x_3 = \text{verdadeiro})$.

[5] Essa busca heurística conhecida por *best-first search* expande o espaço de busca, camada por camada, avaliando todas as novas opções, ou neste caso nós, geradas e escolhe a melhor opção em cada camada para ser expandida, repetindo este processo até que não haja mais expansão possível (LIU & MOTODA, 2000).

Durante a construção de uma árvore, o Trepan mantém em memória uma fila de folhas que são expandidas em sub-árvores quando removidas da fila. Em cada nó da fila, o Trepan armazena (i) um subconjunto das instâncias de treinamento, (ii) um conjunto de instâncias complementares, e (iii) um conjunto de restrições. O subconjunto de instâncias de treinamento armazenado consiste simplesmente dos exemplos que alcançam aquele nó. As instâncias complementares são usadas em conjunto com as de treinamento para selecionar o teste de divisão no caso de um nó interno ou para determinar a classe no caso de uma folha. O conjunto de restrições descreve as condições que instâncias devem satisfazer de maneira a alcançar o nó; esta informação sobre as instâncias que alcançam um nó é usada para determinar um conjunto de instâncias complementares para um novo nó criado.

Encontra-se abaixo uma breve descrição dos aspectos que compõem o algoritmo Trepan, e que pode ser encontrada em detalhes em CRAVEN (1996). Avaliações empíricas do desempenho do Trepan, aplicado a diferentes domínios, frente a outros algoritmos de indução de árvores, tais como ID2-of-3, CN2 e C4.5, podem ser encontradas em MARTINELI (1999), BRAGA et al. (1998) e CRAVEN (1996).

3.2.1.1 Perguntas de associação e o Oráculo

Quando induz uma árvore de decisão, o Trepan utiliza perguntas de associação que são feitas para um oráculo envolvendo instâncias de entrada. Dada uma pergunta de associação, a tarefa do oráculo é retornar a classe à qual uma instância pertence. Uma vez que se busca aprender a função representada pela rede treinada, esta constitui-se o oráculo e responder uma pergunta de associação trata-se simplesmente de classificar uma dada instância.

As instâncias usadas pelo Trepan para as perguntas de associação provém de duas fontes. Uma delas é o conjunto de dados usados para o treinamento da rede. Como outra fonte, o Trepan usa o conjunto de treinamento para construir modelos que representem a distribuição subjacente aos dados e então utiliza esses modelos para gerar novas instâncias, as instâncias complementares empregadas nas perguntas de associação. A habilidade de fazer perguntas de associação e utilizar instâncias complementares possibilita ao Trepan selecionar tanto um teste

de divisão para um nó interno quanto uma classe para um nó folha da árvore apoiando sua decisão em um grande número de dados, ou instâncias.

3.2.1.2 Expansão da árvore

Diferentemente de outros algoritmos de indução que constróem a árvore de decisão utilizando busca exaustiva em profundidade primeiro (depth-first), como por exemplo o algoritmo C4.5, a expansão da árvore feita pelo Trepan utiliza busca heurística por melhor caso primeiro (best-first). A noção do melhor caso aqui refere-se àquele nó que se adicionado à árvore traduz o maior ganho em fidelidade da árvore extraída em relação à rede. Por fidelidade entende-se a extensão a qual a árvore concorda com a rede na classificação. A expansão de uma árvore em uma forma "o melhor primeiro" proporciona um controle fino sobre o tamanho da árvore gerada, uma vez que o processo pode ser parado em qualquer ponto de profundidade desejado.

3.2.1.3 Testes de divisão

A tarefa de selecionar um teste de divisão de um nó envolve a decisão de como particionar o espaço de instâncias que alcançam esse nó. Da mesma forma que o algoritmo ID2-of-3, o Trepan emprega um processo de busca heurística para construir os testes de divisão e expressões Booleanas m-de-n para produzir descrições extraídas mais compactas. Uma expressão m-de-n é satisfeita quando pelo menos m de n condições são satisfeitas, sendo m e n números inteiros. Por exemplo, suponha-se que se tenha os atributos x_1, x_2 e x_3, a expressão m-de-n do tipo 2-de-$\{x_1, \neg x_2, x_3\}$ é sob a ótica da lógica formal equivalente a $(x_1 \wedge \neg x_2) \vee (x_1 \wedge x_3) \vee (\neg x_2 \wedge x_3)$.

3.2.1.4 Critério de parada

O Trepan emprega três critérios para decisão da parada do crescimento da árvore extraída. Primeiro, usa um teste estatístico para decidir se, com alta probabilidade, um nó cobre somente instâncias de uma mesma classe. Se o faz, então este nó não é mais expandido. Segundo, emprega um parâmetro que permite ao usuário determinar um limite de tamanho da árvore a ser gerada. De certa

forma, esse parâmetro, que é especificado em termos de profundidade alcançada por nós internos, permite controlar a compreensibilidade da árvore produzida por limitar seu tamanho máximo aceitável. E por último, o Trepan pode usar um conjunto de validação, associado ao parâmetro tamanho-limite, para decidir a respeito da árvore resultante. Uma vez que o Trepan gera árvores em uma maneira melhor-primeiro, pode-se entender que ele produz uma seqüência aninhada de árvores na qual cada árvore na seqüência difere da sua predecessora somente pela sub-árvore que corresponde ao nó expandido por último. Através de um conjunto de validação, o Trepan mede a fidelidade de cada árvore nessa seqüência, e então retorna a árvore que possui o maior nível de fidelidade para a rede.

As principais vantagens da abordagem do método Trepan, em comparação a outros métodos de extração de conhecimento de redes neurais são duas. Como primeira, o Trepan pode ser aplicado a uma vasta classe de redes neurais. Essa generalidade deriva do fato de que sua interação com a rede consiste somente de perguntas de associação. Dado que perguntas de associação envolvem apenas classificar uma instância, o Trepan não requer nenhuma arquitetura especial nem tampouco qualquer restrição ao processo de treinamento da rede. A outra vantagem principal é que o algoritmo Trepan possibilita um controle sobre a complexidade das hipóteses retornadas pelo processo de extração. Essa capacidade é proporcionada pelo fato de o algoritmo representar suas hipótese extraídas usando árvores de decisão e expandindo-as em uma forma melhor-primeiro. Primeiramente o Trepan extrai uma descrição muito simples da rede treinada, ou seja, um nó apenas, e então sucessivamente refina essa descrição para melhorar sua fidelidade para a rede. Dessa forma, o algoritmo explora crescentemente mais complexas, porém mais fiéis, descrições de uma dada rede treinada.

3.3 REDES NEURAIS NO DOMÍNIO DA ESCOLHA MODAL EM TRANSPORTES

Na literatura consultada, os exemplos de utilização de RNA no domínio da escolha do modo de viagem são poucos e relativamente recentes. Na maioria dos casos encontrados utilizou-se redes neurais como ferramenta para modelagem e comparação de performance contra o modelo convencional, de Utilidade Aleatória.

Somente um estudo foi encontrado onde as redes neurais foram aplicadas com o intuito de investigação sobre o domínio do comportamento de viagem. A seguir, apresenta-se uma breve discussão que relaciona cada um desses estudos. Revisões sobre a aplicação de redes neurais artificiais abrangendo várias sub-áreas da Engenharia de Transportes podem ser encontradas em DOUGHERTY (1995 & 1997), HIMANEN et al. (1998), FURTADO (1998) e RAIA Jr. (2000).

NIJKAMP et al. (1996) analisaram a divisão modal interurbana na Itália para dois modos de viagem, ferroviário e rodoviário, comparando os resultados obtidos com o modelo convencional (Logit) e o modelo de redes neurais artificiais tipo MLP. Para o ajuste dos modelos foi utilizado um conjunto de 1396 dados obtido do censo italiano do ano de 1987, referente a 67 divisões do território daquele país. Os atributos de entrada considerados foram a distância, o tempo e o custo de viagem e a variável de saída foi a probabilidade de escolha de um dos modos. Relatou-se que a arquitetura de RNA com melhor desempenho foi a 6-3-1, ou seja, uma rede com seis unidades de entrada que correspondem aos três atributos de entrada para os dois modos, uma camada intermediária com três unidades processadoras e uma unidade de saída. Os autores concluíram que o desempenho dos dois modelos foram próximos entre si, com uma pequena vantagem para o desempenho do modelo neural.

Um estudo conduzido por DE CARVALHO (1999) comparou a performance do modelo de RNA tipo MLP com o modelo convencional (Logit) expresso em duas formas, uma com função de utilidade linear e outra não-linear. Ela utilizou um conjunto de dados sintético com três atributos de entrada e três alternativas disponíveis à escolha de modo de viagem. Foi constatado que o modelo neural produziu melhores resultados que ambos os modelos Logit. A melhor arquitetura de rede relatada foi uma sem camada intermediária, à qual a autora se refere como rede 9-3, isto é, uma rede com nove unidades de entrada que correspondem aos três atributos de entrada para os três modos e com três unidades de saída que se referem às três alternativas de modo de viagem.

GONZÁLEZ-ESCRIVÁ & SÁNCHEZ (2000) compararam também o desempenho de um modelo Logit e de uma RNA, tipo MLP, para a previsão de escolha entre dois modos de viagem no Corredor Norte da Espanha. Um modelo

Logit com função de utilidade linear e uma única arquitetura de rede proposta com uma camada intermediária com 10 unidades processadoras foram calibrados/treinados utilizando dois conjuntos de dados provindos de pesquisas com coletas de dois tipos, preferência revelada e declarada[6], que incluíam dois e três atributos de entrada respectivamente e duas alternativas de escolha de modo em ambas. Para ambos os conjuntos de dados o modelo de RNA superou o modelo Logit na previsão de escolha para um conjunto de dados de teste não utilizados na fase de ajuste dos modelos.

Um outro estudo semelhante foi desenvolvido por DAPENA & MEDINA (2000) no corredor Madrid-Barcelona com dados a respeito da escolha entre 5 modos alternativos para viagens por motivo trabalho, que continham informações a respeito de quatro atributos de entrada. Para a comparação de desempenho entre modelos, foram utilizados um modelo Logit de função de utilidade linear, uma rede MLP com 16-12-5 unidades dispostas em camadas de entrada, intermediária e de saída, respectivamente, e outra rede MLP de arquitetura curiosamente não declarada no referido artigo. Relatou-se que as redes tiveram desempenho superior para a previsão agregada de escolha e que o modelo Logit teria superado os modelos neurais na previsão individual (desagregada). Entretanto não estão apresentados nesse artigo os valores do desempenho desagregado atingido pelas redes, mas somente ambos os desempenhos, agregado e desagregado, alcançados pelo modelo Logit.

SHMUELI et al. (1996) utilizou redes neurais para explorar uma comparação dos padrões de viagens de homens e mulheres em Israel. Embora o foco do estudo de SHMUELI não tenha sido a divisão modal, dado que o modo de viagem participou como atributo de entrada, trata-se do estudo que mais se aproxima do ora proposto, pelo caráter investigativo buscando conhecer a conexão entre variáveis de entrada, atributos sócio-econômicos e de viagem, e as saídas, padrões de viagem observados, utilizando uma abordagem livre de suposições *a priori*, redes neurais tipo MLP. Os dados utilizados foram coletados por uma pesquisa de entrevista domiciliar de preferência revelada e referem-se às viagens realizadas em dois dias cotidianos nos entornos de Jerusalém, Tel Aviv e Haifa. Esses dados

[6] No capítulo 5, subtítulo 5.2 "Pesquisa Origem-Destino", deste texto, encontra-se a

continham 16 atributos de entrada e a saída se referia ao número de viagens realizadas. Foi proposto que algumas redes teriam todos os atributos como entrada, totalizando 83 unidades de entrada e outras teriam um subconjunto destes, e que algumas redes teriam quatro unidades de saída relativas ao número de viagens para cada determinado motivo de viagem e outras teriam apenas uma unidade de saída para estimar somente o número total de viagens realizadas. Um subconjunto de atributos de entrada seria definido através da utilização do algoritmo de indução de árvores CART que, segundo a autora, identificaria os atributos significativos. A forma utilizada para verificar os propósitos do estudo foi limitar as conexões entre unidades da rede e fazer ligações chamadas "*by pass*", ou seja , ligar unidades de entrada diretamente a unidades de uma segunda camada intermediária, ultrapassando uma primeira camada intermediária sem se conectar a ela. Essas ligações que restringiam os neurônios das duas camadas intermediárias empregadas a apenas alguns neurônios de entrada foram justificadas pela autora como baseadas em "intuição". E essas unidades (neurônios) das camadas intermediárias com conexões restringidas foram chamadas de conceitos de segunda e terceira ordem. Os autores comentam que além das arquiteturas com conexões restringidas foram também testadas arquiteturas convencionais totalmente conectadas, porém não são apresentados os resultados dos desempenhos destas diferentes propostas. Os autores concluem entre outras coisas que os resultados de experimentos de inversão com as redes treinadas sugerem que o sexo é um atributo determinante no padrão de viagem. E que a utilização do CART para definição dos atributos relevantes para construção de propostas de redes não resultou em significativa melhoria de resultados.

definição desses dois métodos de coleta.

4 DESCRIÇÃO DO MÉTODO

Neste capítulo estão apresentadas as fases que constituem o método utilizado para que os objetivos desta pesquisa de mestrado fossem alcançados. Foram elaboradas seis fases de procedimentos. A Figura 4.1 mostra a estrutura do método, bem como o ordenamento das fases que o constituem. É interessante notar que a Figura 4.1 pode ser vista como uma expansão em detalhamento de tarefas da Figura 3.1, apresentada no capítulo anterior, a qual ilustra os passos que constituem em geral um processo de KDD.

Abaixo segue a descrição das fases e procedimentos realizados neste estudo. Esta descrição objetiva uma visualização geral do método, deixando os detalhes de execução de cada procedimento para serem relatados oportunamente ao longo do texto desta dissertação.

4.1 BASE DIGITAL DE DADOS GEO-REFERENCIADOS

O dados disponíveis para este estudo foram coletados na cidade de Bauru, no estado de São Paulo, através de uma pesquisa de entrevista domiciliar de coleta de dados de *seção cruzada* e *preferência revelada* (ORTÚZAR & WILLUMSEN, 1994), realizada por uma empresa privada sob supervisão da Empresa Municipal de Desenvolvimento Urbano e Rural de Bauru – EMDURB, durante o ano de 1997. A pesquisa realizada em 4.000 domicílios, entrevistou aproximadamente 16.000 pessoas, coletando informações relacionadas a viagens realizadas no dia anterior, como por exemplo: endereço de origem, endereço de destino, modo de viagem utilizado, hora de saída, hora de chegada; e informações sócio-econômicas do viajante tais como: idade, grau de escolaridade, número de membros do domicílio, renda, etc. Uma base digital de dados geo-referenciados contendo 8.984 registros de viagem, adaptada de RAIA Jr. (2000), compõe a parcela útil, resultante da

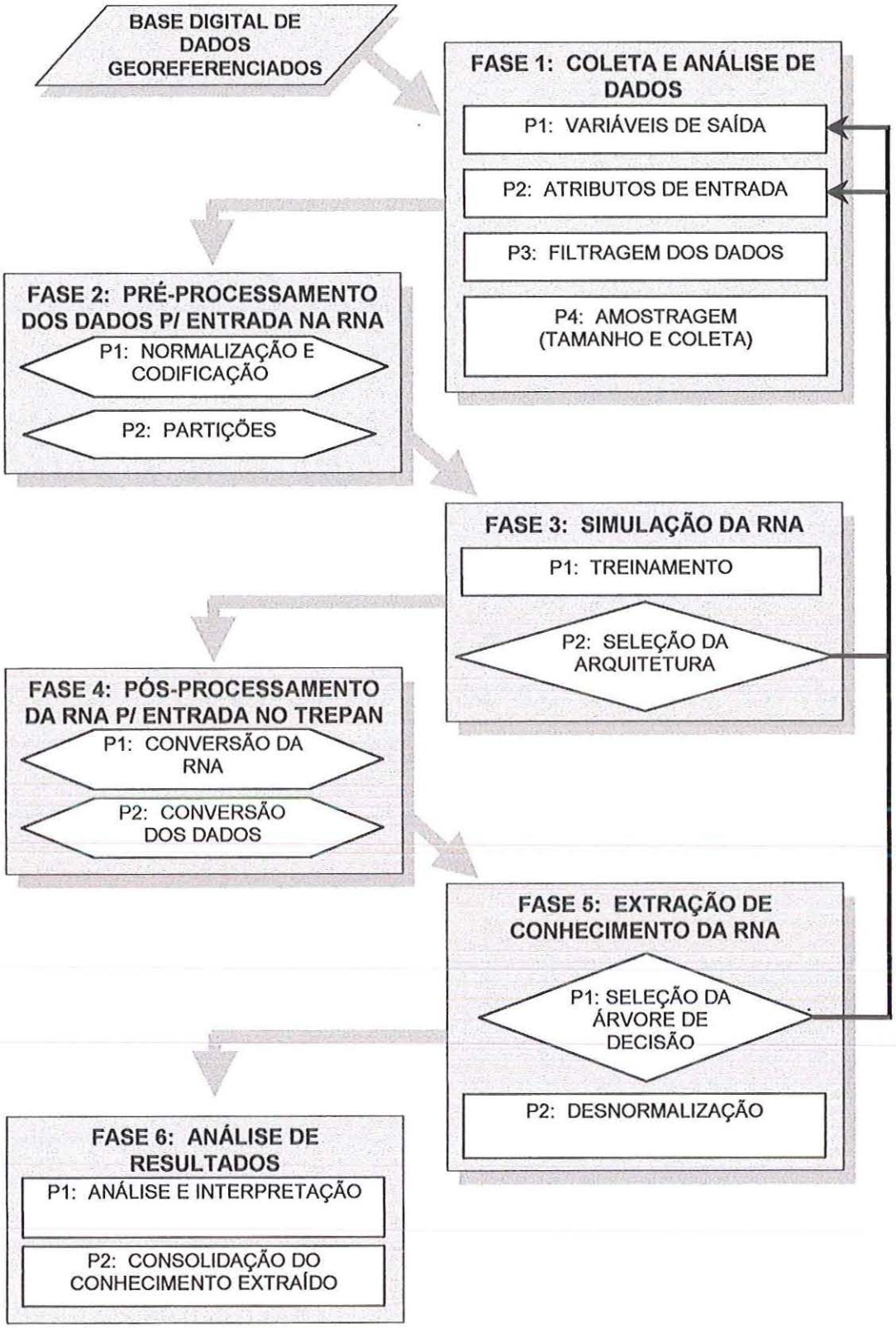


Figura 4.1.: Estrutura do método proposto.

pesquisa domiciliar, para fins deste estudo. Informações mais detalhadas sobre o município, a pesquisa domiciliar realizada e a base digital de dados encontram-se à frente, no capítulo de coleta de dados.

4.2 FASE 1: COLETA E ANÁLISE DE DADOS

Esta fase tem como objetivo a formação da amostra, a partir do disponível conjunto de dados, a ser empregada durante o desenvolver deste trabalho.

Apesar de os dados, nesta fase, já se encontrarem dispostos em uma base de SIG, ainda estão em forma bruta, tendo passado por limpeza apenas nos campos de endereços de origem e destino de viagem, para possibilitar a localização dos mesmos. Portanto, além de definir a amostra a ser utilizada, o pré-processamento de dados objetiva a melhoria da qualidade dos dados fornecidos como entrada para um algoritmo de aprendizado de máquina (FAMILI et al., 1997). A definição de quais informações ou atributos serão as variáveis de entrada, quais modos de viagem serão as variáveis de saída, o tamanho da amostra, e uma filtragem para limpeza dos dados devem ser procedimentos executados anteriormente à fase de modelagem no domínio da escolha modal. De uma forma geral o pré-processamento pode ser dividido em dois grupos de tarefas: dependentes e independentes do domínio. Tarefas dependentes do domínio podem ser solucionadas através de técnicas específicas utilizando conhecimento do domínio, como por exemplo o método de separação dos modos de viagem executado dentro do procedimento "Variáveis de Saída", a frente descrito. Por outro lado, tarefas independentes do domínio, isto é de aplicação genérica, referem-se, por exemplo, à redução de ruído nos dados ou ao tratamento de registros com valores inexistentes.

A seguir, descrevem-se os procedimentos desta fase, mostrando-se de forma objetiva e sucinta, os passos a serem seguidos durante sua execução.

4.2.1 Variáveis de Saída

Por tratar-se da modelagem do comportamento de escolha modal, as variáveis de saída do modelo são os modos de viagem disponíveis à escolha do

indivíduo para realização de uma viagem. Ou seja, a partir dos dados de entrada o modelo indica como saída a classe pertencente ou modo escolhido. Definem-se neste procedimento quais modos de viagem serão tratados pelo modelo proposto.

A pesquisa de entrevista domiciliar coletou informações de viagens referentes a quatro alternativas de modo viagem. Dentre estas alternativas, três delas agregam dois diferentes modos em uma única alternativa para a resposta do entrevistado. Com o intuito de enfocar-se o estudo em cada modo distintamente, aplicou-se um método de separação para dois modos, agrupados em uma das alternativas, baseado nas características inerentes a estes. Embora fosse desejada a separação de todos os modos que se apresentam agrupados em alternativas, somente em uma delas encontrou-se informações da pesquisa domiciliar para tanto.

O método baseia-se na análise de tabelas de freqüência resultantes do cruzamento de três informações, ou atributos, referentes às viagens feitas por determinada alternativa de modo, como por exemplo a alternativa de modo não motorizado. Nesta alternativa deseja-se identificar quais viagens foram feitas por um modo e quais foram pelo outro, dentre os dois modos que esta alternativa representa, por exemplo, os modos bicicleta e caminhada no caso de alternativa de modo não motorizado. O cruzamento de atributos foi feito com o auxílio do software estatístico *STATISTICA*[1]. Após a distinção entre estes dois modos, um deles, e todos o registros pertencentes a ele, foi excluído da base de dados e não estará sob o foco deste estudo. Assim, a alternativa de modo submetida a este método de distinção passa a representar viagens exercidas apenas por um modo de viagem.

O fato de a pesquisa domiciliar não ter colhido informações sobre quais modos de viagem estavam disponíveis para o entrevistado implica na suposição de que todos os modos eram disponíveis em todas as viagens coletadas. Inicialmente, este estudo procurou enfocar quatro modos de viagem e na fase de simulação RNA, esta quantidade foi reduzida para três modos de viagem de acordo com justificativa oportunamente apresentada.

[1] StatSoft, Inc. (1999). STATISTICA for Windows [Computer program manual]. StatSoft, Inc., 2300 East 14th Street, Tulsa, OK 74104. http://www.statsoft.com

Definidos quais modos de viagem serão tratados, os procedimentos seguintes definem outras duas dimensões do problema: quais atributos de entrada e quais instâncias serão usados na modelagem proposta.

4.2.2 Atributos de entrada

A partir do conjunto de atributos disponíveis, procurou-se selecionar quais atributos formariam um conjunto ótimo de entrada para a modelagem. A seleção do conjunto dos atributos de entrada, empregados na modelagem proposta neste estudo, busca a melhoria de performance em termos da precisão preditiva da RNA treinada e da simplicidade ou compreensibilidade da árvore de decisão gerada. Esta seleção segue um modelo de seleção de atributos para aprendizado de máquina conhecido como *wrapper model* (LIU & MOTODA, 2000). Este modelo consiste em variar os atributos de entrada, formando-se diferentes subconjuntos de atributos, e observar a performance do próprio classificador, neste caso a RNA, para a determinação de qual subconjunto é o ótimo, à tarefa de classificação.

4.2.3 Filtragem dos dados

Este procedimento tem como objetivo a limpeza dos dados tanto pela remoção de registros com informações nulas, absurdas, inexistentes ou incompletas, bem como pela remoção de dados considerados não desejados, como por exemplo, viagens menores que 150 metros, viagens feitas por pessoas com menos de 5 anos de idade, etc. Uma filtragem por inspeção em toda a base de dados deve detectar os registros a serem removidos, conforme a primeira situação descrita, e uma filtragem dirigida deve identificar os registros considerados não desejados, a serem removidos também.

4.2.4 Amostragem

Este procedimento de amostragem objetiva a definição de um subconjunto da base de dados que a represente e que seja adequado para o treinamento de RNAs.

O processo de extração da amostra a partir da base de dados deve ser o de *amostragem aleatória*. O tamanho da amostra deve ser definido através da observação, na base de dados, da *distribuição amostral das médias* para atributos contínuos e da *distribuição amostral das proporções* para atributos discretos (SPIEGEL, 1984), considerando-se o caso mais crítico para o grau de confiabilidade e o erro de estimativa escolhidos (SMITH, 1979).

Será feita uma segmentação da base de dados para a formação de agrupamentos homogêneos de registros considerando-se todos os atributos, sócio-ecônomicos e de viagem, bem como o modo escolhido. Serão removidos os registros repetidos dentro de cada agrupamento e, então, será extraída a amostra por sorteio aleatório a partir dos dados exclusivos restantes na base de dados. Esta seleção de registros exclusivos visa a maximização da diversidade contida na amostra, de tamanho definido, para o aprendizado de máquina e está descrita em detalhes, adiante, neste texto.

Deve ser observado na composição da amostra um balanceamento entre as quantidades de dados pertencentes a cada classe de modo de viagem. Um desbalanceamento entre classes no conjunto de dados para treinamento da RNA pode fazer com que o aprendizado favoreça a classe com maior quantidade de instâncias[2], ou dados (JAPKOWICZ, 2000). Um primeiro sorteio aleatório deve satisfazer o tamanho mínimo determinado da amostra. Em seguida serão sorteados dados, também de forma aleatória, para as classes minoritárias até que se estabeleça a igualdade no número de instâncias entre todas as classes.

Uma vez definidos quais registros, ou viagens, compõem a amostra, serão coletadas informações de viagem para a alternativa de modo ônibus, tais como a distância de acesso e egresso à linha de ônibus, a distância viajada a bordo e a quantidade estimada de tarifas pagas. Estas informações, diferentemente de outras, não foram possíveis de ser obtidas por processo automatizado pelo software de SIG utilizado, *TransCAD*[3] versão 3.2, e serão obtidas de forma manual, para cada uma das viagens da amostra, com o auxílio de ferramentas do software.

[2] Este texto não faz distinção entre as palavras: casos, dados ou exemplos de treinamento, e instâncias.
[3] Caliper Corporation. 1172 Beacon Street, Newton, Massachusetts, USA.

4.3 FASE 2: PRÉ-PROCESSAMENTO DOS DADOS PARA ENTRADA NA SIMULAÇÃO DE REDE NEURAL ARTIFICIAL

Esta fase é de caráter preparatório para a subseqüente fase de simulação de RNAs. Trata-se apenas da formatação dos dados amostrados tanto para a adequação necessária para o modelo de RNA escolhido, que implica em normalização, codificação e divisão dos dados em partições, quanto para a adequação ao formato de leitura exigido pelo simulador de RNA escolhido.

4.3.1 Normalização e Codificação

Devem ser identificados, no conjunto de dados para a simulação, quais atributos são numéricos, ou contínuos, tais como idade e distância, e quais atributos são categóricos, ou discretos, como por exemplo sexo ou motivo de viagem.

Os dados de entrada pertencentes a atributos discretos nominais, ou seja, sem ordenação, devem ser codificados em vetores de valores binários, onde somente um elemento do vetor é igual a 1 e os demais são iguais a 0. Um exemplo deste caso pode ser o atributo sexo. O sexo feminino pode ser representado pelo vetor (0,1) e o masculino pelo vetor (1,0).

Os dados de entrada pertencentes a atributos discretos ordinais, com ordenação presente nos dados, devem ser codificados em vetores binários em "tons de cinza", de forma a representar tal ordenação. Como exemplo de características em que existe ordenação nos dados, pode ser citado o atributo grau de instrução, ou de escolaridade de um indivíduo. Os graus de instrução 1°, 2° e 3° podem ser codificados, por exemplo, para os vetores binários (0,0), (0,1) e (1,1) respectivamente.

Todos os dados pertencentes a atributos numéricos devem ser normalizados para um mesmo intervalo, por exemplo o intervalo 0 e 1. Dados com atributos em escalas diferentes podem confundir a rede fazendo com que esta dê mais importância para aqueles com maiores valores (MARTINELI, 1999). Entretanto cada atributo numérico deve ser normalizado individualmente, a partir dos próprios valores limites, superior e inferior, observados entre seus dados. A normalização de cada atributo numérico deverá ser feita através da Equação 4.1.

$$y_i = \left(\frac{(x_i - x_{mín})(y_{máx} - y_{mín})}{x_{máx} - x_{mín}} \right) + y_{mín} \qquad ,para\ i=1..n \qquad (4.1)$$

onde: y_i é o valor normalizado e x_i é o valor a ser normalizado; $x_{mín}$ e $x_{máx}$ são o menor e o maior valor do intervalo origem, respectivamente; $y_{mín}$ e $y_{máx}$ são o menor e o maior valor do intervalo de destino da normalização, respectivamente; e n é a quantidade de valores do intervalo de origem a serem normalizados.

Para os dados correspondentes às classes de saída, alternativas de modo de viagem, deve ser feita codificação para vetores de valores binários sem ordenação.

4.3.2 Formação das Partições: Treinamento, Validação e Teste

Tanto para o processo de aprendizado por simulação da RNA quanto para o processo de extração do conhecimento da RNA, os dados devem ser divididos em partição de treinamento, de validação e de teste. Segundo as regras apresentadas em *PROBEN1* por PRECHELT (1994), estas partições devem conter 50%, 25% e 25% de todos os dados da amostra, respectivamente. A seleção de dados para composição de cada partição deve ser feita de maneira aleatória e sem reposição. Cada partição deve ser um subconjunto mutuamente exclusivo da amostra, ou seja, uma instância da amostra pode estar somente em uma das partições.

Como método de "reamostragem" (*resampling*), para uma estimativa mais precisa da taxa de acerto de classificação do modelo de RNA, devem ser feitos três diferentes conjuntos de dados para simulação, cada qual contendo partições de treinamento, validação e teste. Ou seja, estes três conjuntos de dados para simulação nada mais são que três diferentes formas de arranjar os dados da amostra em partições de treinamento, validação e teste.

4.4 FASE 3: SIMULAÇÃO DE REDES NEURAIS ARTIFICIAIS

Nesta fase apresentam-se os procedimentos de treinamento e seleção de arquitetura de RNA, considerados nos experimentos para obtenção das RNAs treinadas de melhor desempenho. A sua finalidade é definir o tipo de RNA utilizado, definir o algoritmo e seus parâmetros de aprendizagem, bem como o critério de

parada do treinamento, propor as diferentes arquiteturas de RNA para treinamento e definir os critérios de avaliação do desempenho e de escolha da RNA treinada. Como produto desta fase, tem-se as RNAs treinadas que apresentam melhor desempenho e os respectivos índices de desempenho.

4.4.1 Treinamento de RNA

Para o processo de treinamento das redes MLP será utilizado o algoritmo de aprendizado supervisionado *backpropagation* (RUMELHART et al, 1986), com atualização de pesos por padrão (*on-line*). O software empregado para a simulação das redes será o *SNNS – Stuttgart Neural Network Simulator* versão 4.1 (ZELL et al, 1995). Será utilizada tanto a interface gráfica do simulador quanto seu recurso de automatização para treinamento de redes, chamado *Batchman*. A interface gráfica do *SNNS* proporciona a visualização dos erros de treinamento e validação por ciclo, durante o treinamento, além do fácil acesso a troca de parâmetros de aprendizagem, o que a torna ferramenta indispensável para investigação do processo de treinamento. O recurso *Batchman* deverá ser empregado na busca extensiva por arquiteturas de melhor desempenho pelo fato de executar o processo de treinamento e retornar os resultados deste de maneira automática.

Para execução de todos os treinamentos, a inicialização dos pesos da rede a ser treinada será feita segundo a forma padrão oferecida pelo simulador: aleatória com valores possíveis dentro do intervalo de –1 a 1. As instâncias de treinamento e de validação deverão ser apresentadas à rede durante a simulação, em ordem aleatória. A apresentação de instâncias de validação, para cálculo do erro de validação durante o treinamento, deverá ser feita a cada vinte ciclos de treinamento. A taxa de aprendizagem deverá ser decrescente durante o processo de treinamento, ou seja, o valor numérico atribuído a esse parâmetro no início do processo deverá sofrer redução durante o treinamento de uma rede. Essa redução da taxa de aprendizagem procura dirimir problemas de instabilidade de treinamento e melhor aproximação do algoritmo de aprendizado à solução mínima. Para maiores detalhes consultar CARVALHO et al (1999).

O critério de parada do treinamento deverá ser a elevação do erro de validação durante o treinamento. Assim, foi escolhido, para cálculo do erro de

validação, o erro médio quadrático MSE (Mean Squared Error) disponível no simulador *SNNS*, calculado de acordo com a Equação 4.2.

$$MSE = \frac{\sum_{p \in padrões} \sum_{j \in saída} (y_{pj} - o_{pj})^2}{n} \tag{4.2}$$

onde y_{pj} é a saída desejada do neurônio de saída j para o padrão p, o_{pj} é a saída obtida (ou calculada) para este mesmo padrão, e n é o número de padrões.

Neste trabalho, um processo de treinamento que empregue o recurso de automatização *Batchman* deverá ser composto por três eventos de treinamento envolvendo três diferentes conjuntos de dados para simulação, cada conjunto sendo formado por partições de treinamento, de validação e teste, como definido no procedimento de formação das partições. Em cada evento de treinamento ocorrem seis treinamentos, cada um inicializado com um diferente valor para o parâmetro taxa de aprendizagem. Desta maneira, cada arquitetura de rede, submetida a simulação automatizada, deverá passar dezoito vezes por treinamento, com validação, e teste.

A execução de treinamentos através da interface gráfica do simulador não estará sujeita ao rigor das dezoito repetições acima descritas uma vez que será empregada apenas a caráter de estudo e investigação de processos de treinamento.

4.4.2 Seleção da Arquitetura

O espaço de busca entre as redes válidas é muito grande. Além disso, ele é deceptivo e multimodal, como mostram MILLER et al. (1989). Deceptivo porque duas arquiteturas similares podem apresentar desempenhos muito diferentes. E multimodal porque duas arquiteturas muito distintas podem ter desempenhos semelhantes.

À procura da rede de melhor desempenho, serão propostas várias diferentes arquiteturas, com uma e duas camadas intermediárias de neurônios, para serem submetidas simultaneamente ao processo de treinamento. A proposição das arquiteturas deve observar as quantidades, a serem modeladas, de atributos de entrada e de alternativas de modo de viagem, o que define, respectivamente, a

quantidade de neurônios das camadas de entrada e de saída da rede. As arquiteturas propostas devem possuir variadas quantidades de neurônios nas camadas intermediárias. Iniciando-se com propostas com poucos neurônios, aumenta-se de maneira linear esta quantidade, com incremento pequeno no início e maior após atingido um determinado tamanho de camada, até propostas com uma quantidade razoavelmente grande de neurônios, em uma e em duas camadas intermediárias.

O desempenho de uma arquitetura deverá ser medido através da média dos três melhores desempenhos obtidos em cada evento de treinamento. Deverá ser observado também o valor do desvio padrão para essa média. A escolha da arquitetura de melhor desempenho envolve o dilema da melhor, ou maior média, contra o melhor, ou menor desvio padrão observados. Uma vez escolhida a arquitetura de melhor desempenho, a rede treinada que a representará deverá ser a de maior desempenho dentre as dezoito redes desta mesma arquitetura, considerando novamente o desvio padrão, porém agora para desempenho entre as classes de saída da rede.

A medida do desempenho deverá ser a porcentagem de casos, ou instâncias, corretamente classificados para a partição de validação, obtida através de uma ferramenta do simulador que utiliza a técnica "*winner takes all*", para análise de classificação da rede. Essa técnica, quando aplicada à análise dos valores de saída de uma rede treinada, indica o neurônio da camada de saída que fornece o maior valor de saída como sendo a classe correspondente para determinada instância apresentada à classificação. Assim, uma instância é considerada corretamente classificada quando a classe indicada (calculada) corresponde à real classe informada (desejada). Deverá ser observado tanto o percentual de acerto médio de classificação de uma rede, para as instâncias da partição de validação, bem como o percentual de acerto de classificação para cada classe de saída desta rede. Procura-se, dessa maneira, evitar atribuir-se um alto desempenho a uma rede que eventualmente possua um percentual de acerto médio elevado mas que execute bem a classificação para apenas uma das classes.

4.5 FASE 4: PÓS-PROCESSAMENTO DA RNA PARA ENTRADA NO ALGORITMO TREPAN

O algoritmo *Trepan* não teve seu desenvolvimento visando um acoplamento direto de funcionamento a qualquer simulador de redes específico. Por isso, o *Trepan* demanda formatos próprios de entrada, para rede treinada e dados de simulação.

Meramente preparatória, essa fase faz a interface entre as etapas de simulação de RNA e de extração de conhecimento, através da conversão dos formatos de rede treinada e dados utilizados no simulador *SNNS* para os formatos de leitura exigidos pelo algoritmo *Trepan*.

4.5.1 Conversão da RNA

A conversão do formato da rede treinada, e salva em arquivo de saída do SNNS, para o formato de leitura do Trepan, deverá ser feita através de um software criado em linguagem científica C. Essa conversão não envolve qualquer tipo de cálculo mas apenas a disposição e a tabulação, em outro formato, de todas as informações que concernem uma rede treinada, tais como pesos das ligações, número de camadas, número de neurônios por camada, etc.

4.5.2 Conversão dos Dados

O algoritmo Trepan emprega em seu funcionamento todas as partições, de treinamento, de validação e de teste, usadas na fase de simulação. Assim como no procedimento anterior, deverá ser utilizado um software, criado em linguagem científica C, para a conversão do formato dos arquivos das partições para o formato de leitura do Trepan. Nessa conversão os atributos contínuos mantém-se na forma normalizada, porém os atributos discretos são convertidos da forma vetorial para o nome da categoria representada. Por exemplo, em uma dada instância, o vetor (0,0,1) como representante da categoria 1° grau, para o atributo grau de escolaridade, deve ser convertido para o nome "primário", ou "1grau". No caso dos atributos discretos "booleanos", é feita a conversão da forma vetorial para a forma "booleana", por exemplo o vetor (0,1) para "falso" e o vetor (1,0) para "verdadeiro".

4.6 FASE 5: EXTRAÇÃO DE CONHECIMENTO

Nesta fase apresentam-se os procedimentos de seleção da árvore de decisão e decodificação dos valores normalizados e codificados dos atributos, dispostos nos nós das árvores. Esta fase objetiva encontrar a árvore com a profundidade de melhor desempenho e que mais se aproxima, na tarefa de classificação, de sua original rede treinada.

4.6.1 Seleção da Árvore de Decisão

Dispondo-se dos dados preparados na fase anterior, para cada rede treinada e selecionada na fase de simulação deverá ser gerada uma árvore com nível de profundidade igual a 100. Deverá ser, então, analisado o arquivo de saída que guarda, para cada nível de profundidade atingido, os índices de acerto de classificação, bem como os índices de fidelidade ou proximidade da árvore em relação à rede, para que se determine a profundidade ótima, ou seja, a profundidade onde se encontram os maiores índices tanto de classificação como de proximidade. Uma árvore deverá ser gerada para a profundidade ótima de cada rede. Havendo profundidades diferentes com índices de acerto próximos, deverá ser gerada uma árvore para cada profundidade. Porém considera-se a melhor árvore, para determinada rede, aquela com menor profundidade dentre as de índices de acerto semelhantes.

O método do algoritmo Trepan a ser utilizado é o método disjuntivo, por ser, como concluiu CRAVEN (1996), o de maior compreensibilidade de árvore gerada, dentre os disponíveis no algoritmo, pelo fato de indicar a escolha de apenas um atributo por nó.

4.6.2 Desnormalização

A árvore gerada pelo Trepan traz, eventualmente, em seus nós, atributos contínuos relacionados a valores normalizados, o que dificulta a interpretação da árvore. Esses valores, presentes nos nós da árvore, deverão ser trazidos de volta aos valores originais de cada atributo para possibilitar a montagem de uma regra

correspondente às restrições impostas pelos nós. Assim, se um nó apresentar por exemplo a relação "idade>0,206", esta deverá ser transformada para "idade>22anos", o que suscita direta compreensão. Situação semelhante ocorre com atributos discretos e do tipo "booleano", como é o caso do atributo sexo que apresenta-se na árvore recebendo valores "verdadeiro" ou "falso". Para facilitar a compreensão nestes casos, o atributo "booleano" deverá ter seus valores transformados de volta à situação inicial adotada, como por exemplo "sexo=verdadeiro" substitui-se por "homen", e assim por diante. Para uma árvore, quando aplicável, uma planilha será confeccionada de forma que suas linhas correspondam ao caminho de restrições o qual se sujeita para que a partir do nó raiz se atinja uma folha.

4.7 FASE 6: ANÁLISE DE RESULTADOS

Na fase final deste método, analisam-se os resultados, em termos de conhecimento extraído das RNAs treinadas. Os procedimentos desta fase, abaixo descritos, determinam o quanto os resultados são interessantes mediante critérios apresentados por KLÖSGEN (1996) para a definição do grau de *interestingness*[4] para as saídas, conhecimento, ou achados, de um processo de KDD.

4.7.1 Análise e Interpretação

A discussão a respeito da análise e da interpretação dos resultados segue alguns critérios que compõem o caráter multiface do "interessante", para a avaliação do conhecimento obtido:

- *Evidência*: diz respeito a significância de um achado medido por um critério estatístico, neste caso o índice de acerto de classificação da árvore de decisão gerada;
- *Redundância*: refere-se a quantidade de similaridades observadas entre as árvores geradas;

[4] Foi mantido o termo original por não haver um correspondente direto na língua portuguesa. Todavia, pode-se entender por grau de *interestingness*, o grau de algo ser interessante.

- *Utilidade*: relaciona um achado aos objetivos buscados por este estudo;
- *Novidade*: inclui o desvio do conhecimento prévio sobre o domínio enfocado;
- *Simplicidade*: refere-se a complexidade sintática da apresentação do achado.

4.7.2 Consolidação do Conhecimento Extraído

A consolidação do conhecimento extraído concerne o critério de generalização alcançada. Sendo assim, o conjunto total da base de dados será submetido à árvore de decisão gerada, e escolhida como a melhor opção, para que seja calculado o índice de acerto de classificação. Procura-se através deste índice de acerto, calculado para a "população", inferir-se sobre a capacidade de generalização do conhecimento obtido, representado na árvore de decisão.

4.8 CONSIDERAÇÕES FINAIS

O trabalho empregado na adaptação da base de dados em SIG não foi apresentado como uma fase pertencente ao método, com o intuito de que o método fosse configurado apenas por fases diretamente referentes a modelagem executada, visando a possível replicação deste, a despeito de qualquer forma a qual os dados venham a estar dispostos.

Os procedimentos de seleção da arquitetura de RNA e da árvore de decisão fazem retroalimentação aos procedimentos de variáveis de saída e de entrada como mostra a Figura 3.1. Esta interação permite que a partir dos resultados, obtidos nos procedimentos de seleção, possa-se redefinir o conjunto de variáveis, ou atributos, de entrada à procura da melhoria da performance de classificação. Ainda permite que mediante verificação da performance de classificação para cada uma das classes, ou modos de viagem, possa-se detectar eventual desbalanceamento de performance de classificação entre os modos. Tal desbalanceamento pode indicar um mal ajuste do modelo de RNA ou a provável falta de informação necessária à definição do conjunto de padrões pertencentes a um modo de viagem, no espaço de busca disponível à RNA.

5 COLETA E ANÁLISE DE DADOS

Este capítulo contém a descrição pormenorizada dos procedimentos da fase de aquisição do conjunto de dados empregados neste estudo. Esta fase de um processo de KDD é a mais dispendiosa em termos de esforços consumidos para sua execução, como menciona CABENA et al. (1997), e como tal, gerou a criação de um anexo para disposição de uma parte das tarefas executadas nesta fase, fora do corpo principal deste texto.

Ainda que houvesse um grande volume de informações a serem relatadas nesse capítulo, decidiu-se pela inclusão de uma breve descrição da cidade em estudo e da pesquisa de entrevista domiciliar, com o intuito de apresentar o ambiente de onde provém os dados e a forma com a qual foram coletados.

5.1 CARACTERIZAÇÃO DA CIDADE EM ESTUDO

O município de Bauru está localizado na região centro-oeste do estado de São Paulo, a 325 km da Capital, e sua população é pouco superior a 310 mil habitantes. As suas maiores fontes de renda são os setores de comércio e de serviços, que têm amplitude regional, e em menor escala, o setor industrial (RAIA Jr.,2000).

A taxa de urbanização do município é alta, 98,2% (SEADE, 2001), quando comparada com a média brasileira, 79% em 1995 (RAIA JR, 2000). O tamanho da frota de veículos em Bauru era cerca de 125 mil veículos, em 1999, com 10 mil novos veículos colocados nas ruas por ano (LIMA, 1999). Isto representa 1 veículo para cada grupo de 2,4 moradores, um valor bem maior que a média nacional, que em 1997 era de 1 veículo para cada grupo de 9,4 habitantes (ANFAVEA, 1999).

Próximo à data em que foi realizada a pesquisa de entrevista domiciliar em Bauru, 1997, o sistema de transporte coletivo por ônibus era operado por três empresas que transportavam cerca de 4,2 milhões de passageiros/mês (NÚMERO, 1999). Hoje ainda são três empresas operadoras, porém a demanda caiu para cerca de 3.1 milhões de passageiros/mês, o que resulta uma média de pouco mais de 100 mil passageiros/dia – dados médios anuais do ano de 2000 (EMDURB, 2001). O sistema é composto por 85 linhas, sendo 54 diametrais, 22 radiais, 3 circulares, 1 distrital e 5 noturnas. A quilometragem média mensal percorrida é de 1,7 milhões de quilômetros, com quase 100% de atendimento à área urbanizada, considerando uma faixa de atendimento de 300m medidos a partir de cada lado da linha (EMDURB, 2001).

5.2 PESQUISA ORIGEM-DESTINO

Os dados utilizados neste trabalho provém de uma pesquisa origem-destino realizada pela EMDURB (Empresa Municipal de Desenvolvimento Urbano e Rural de Bauru) que faz parte do projeto "Desenvolvimento de Modelos de Demanda", financiado pelo FINEP (Financiadora de Estudos e Projetos), no programa RECOPE (Rede Cooperativa de Pesquisa), em que o município de Bauru tem como parceiro o Departamento de Transportes da EESC.

A pesquisa origem-destino contou com o método de coleta de entrevista domiciliar, ou seja, cada domicílio da amostra foi visitado por um entrevistador que coletou informações sobre as viagens realizadas no dia anterior por todos os membros daquele domicílio. A informação sobre a real escolha feita e relatada pelos entrevistados é conhecida como informação, ou dado, de *preferência revelada*. Esse tipo de coleta é também conhecido como *seção cruzada*, ou seja, procura-se aproximar com a amostra de um único dia as variações prováveis de comportamento que hajam durante o decorrer dos dias próximos àquele pesquisado. Raros são os exemplos na literatura de coleta do tipo *seção*

longitudinal, no qual o domicílio é entrevistado por vários dias consecutivos ou até semanas consecutivas[1], dado o elevado custo de coleta.

A coleta foi realizada durante o ano de 1997 considerando o calendário escolar para que fosse observada a situação típica cotidiana, e abrangeu 4 mil domicílios. Os domicílios foram sorteados de forma aleatória através do cadastro da companhia de luz da cidade. Na época, a cidade possuía pouco mais de 300 mil habitantes. Adotando-se o número médio de 3,34 pessoas por família obtido da própria pesquisa, obtém-se aproximadamente 90 mil famílias ou domicílios. Atualmente a cidade possui 108 mil domicílios (SEADE, 2001). Assim sendo, o número de domicílios abordados representa 4,5% do total de domicílios do município e está bem acima do mínimo admissível sugerido por ORTÚZAR & WILLUMSEN (1994) que, para uma população entre 150 mil e 300 mil habitantes, é de 2,9%.

A pesquisa domiciliar já utilizada em outros estudos, por LÓPEZ-REYES (1999) e RAIA Jr. (2000), catalogou 23.314 deslocamentos, ou viagens intraurbanas. É importante destacar que essa pesquisa domiciliar realizada na cidade de Bauru não tinha como objetivo servir para a calibração de algum modelo de escolha modal, nem tampouco servir para aprendizado de uma RNA para fins de escolha modal, mas sim para estimar a matriz O/D (origem-destino) de viagem. Assim sendo, não foram coletadas informações sobre os modos de viagem disponíveis ao usuário, no ato da entrevista. Também ficaram excluídas outras informações a respeito das alternativas de modo que não foram escolhidas, como por exemplo a disponibilidade do vale transporte ou o tempo de viagem para os modos alternativos, não escolhidos.

Os modos de viagens foram divididos em quatro categorias no questionário da entrevista domiciliar da seguinte forma:

- *Modo 1: automóvel ou moto como motorista;*
- *Modo 2: automóvel ou motocicleta como passageiro (carona);*
- *Modo 3: bicicleta ou caminhada (modos não motorizados) e*
- *Modo 4: ônibus.*

[1] Para exemplos desse tipo de coleta ver MARBLE et al. (1972), GOLOB & MEURS (1986) e

Pode-se notar que essa divisão de modos de viagem fez com que os modos 1, 2 e 3 aglutinassem diferentes modos, o que acarreta uma perda de informação já na fase de coleta. No caso dos modos 1 e 2, os modos automóvel e motocicleta, lá aglutinados, certamente traduzem diferentes interferências no sistema de transporte e provavelmente a escolha dos mesmos é caracterizada por diferentes aspectos ou atributos. O modo 3, por sua vez, aglutina os modos caminhada e bicicleta, o que desfavoreceria, à primeira vista, o reconhecimento dos padrões de viagem destes modos, e de suas características, distintamente.

5.3 PREPARAÇÃO E ANÁLISE DOS DADOS

5.3.1 Utilização de SIG

Uma base digital de dados da cidade de Bauru foi construída em um Sistema de Informações Geográficas (SIG), utilizando o software *TransCAD* versão 3.0, por RAIA Jr. (2000) e gentilmente cedida para uso neste estudo. Essa base geo-referenciada é composta por eixos de ruas, quadras, nomes das ruas, número das quadras, trajetos das linhas de ônibus, zonas de tráfego, rios e córregos, traçado ferroviário, etc. A partir dessa base foram localizados os endereços de origem e os de destino das viagens e posicionados nas esquinas, ou cruzamentos de vias, mais próximos, dados estes coletados na pesquisa domiciliar. Essa localização de endereços possibilita o cálculo das distâncias de viagem necessárias a este estudo, através de ferramentas disponíveis no software utilizado. Para maiores informações sobre os procedimentos utilizados para cálculo da distância de viagem, consultar WERMERSCH & SILVA (2001).

Dentre as 23.314 viagens coletadas pela pesquisa em Bauru, a grande maioria, pouco mais de 22.500 das viagens, era do tipo pendular, ou seja, do tipo ida e volta. Apenas algumas poucas viagens eram sequenciadas. Para fins deste estudo foram somente consideradas as viagens de ida. Assim foram localizadas, no SIG, um total de 8.984 viagens de ida, dentre aproximadamente 11.000 disponíveis. Essa diferença ocorreu pela perda de alguns endereços de origem e/ou de destino, dada à qualidade da informação proveniente da pesquisa domiciliar: nomes de

KITAMURA & BOVY (1987).

logradouros incompletos, inexistentes, etc. Desta forma, este trabalho emprega toda a amostra útil, no que se refere a viagens de ida, disponível na pesquisa origem-destino realizada em Bauru, ou seja, os 8.984 registros ou viagens, acima mencionados.

5.3.2 Definição das Alternativas de Escolha Modal

A divisão dos modos de viagem feita no momento da entrevista domiciliar desfavorece o estudo ou uma melhor caracterização do comportamento em relação a cada modo individualmente, porque, aglutina diferentes modos em uma única opção de resposta para o entrevistado, como é o caso nos modos 1, 2 e 3. A partir das informações colhidas pela pesquisa domiciliar, acredita-se não ser possível a separação dos modos automóvel e motocicleta incluídos nas opções de modos 1 e 2. Porém no caso do modo 3, onde inclui-se o modo bicicleta e caminhada, foi aplicado um método indireto para a separação destes modos.

Há importância na separação dos modos bicicleta e caminhada pois as características da viagem, como por exemplo rapidez e distância percorrida, para estes modos podem ser muito diferentes. Uma viagem por bicicleta na área urbana de uma cidade, que passe pelo centro comercial na hora pico de trânsito, pode ser mais rápida que por modo carro e ainda pode percorrer uma distância tão longa quanto por modo ônibus ou mesmo carro. Estas diferentes características das viagens por bicicleta misturadas às viagens por caminhada num único modo, dificultariam uma precisa caracterização deste modo pela modelagem proposta, dada a heterogeneidade deste modo misto.

O método aplicado busca distinguir as viagens realizadas por bicicleta dentre as viagens do modo 3. Para isto, procurou-se por critérios que possibilitassem um processo de filtragem nos dados, identificando e removendo os correspondentes às viagens por modo bicicleta. A aplicação dessa filtragem no conjunto de viagens por modo 3 (2.670 viagens), identificou 289 prováveis viagens por modo bicicleta as quais foram excluídas deste conjunto. Essas 289 viagens por bicicleta representam 3,2% de todas as viagens coletadas em Bauru e disponíveis para este trabalho, 8984 viagens. Um estudo realizado por GEIPOT (1985) na cidade de Bauru estimou, naquela época, que cerca de 2% da população estaria

utilizando o modo "ciclo" que incluía bicicletas e ciclomotores. Embora o método aqui aplicado tenha identificado um percentual maior de prováveis viajantes por modo bicicleta que o estudo mais antigo realizado pelo GEIPOT, admite-se que este método tenha atingido seu objetivo de apenas eliminar distorções grosseiras no conjunto de dados do modo 3. Destinou-se o anexo C deste texto para a apresentação detalhada desse método aplicado.

Uma vez excluídas as viagens dadas como sendo por modo bicicleta, considera-se que o restante das viagens do modo 3 (2.381 viagens) sejam por modo caminhada, e como tal, resulta em 8.695 viagens agora disponíveis para este trabalho.

Os modos de viagem foram, então, considerados como sendo quatro da seguinte maneira:

- modo 1: automóvel ou motocicleta como motorista;
- modo 2: carona (automóvel ou motocicleta como passageiro);
- modo 3: a pé (caminhada);
- modo 4: ônibus.

Pelo fato da pesquisa domiciliar não ter coletado informação sobre quais modos de viagem o entrevistado realmente dispunha, considera-se *a priori* que todos os quatro modos, considerados neste estudo, eram as alternativas disponíveis à escolha do viajante.

5.3.3 Definição dos Atributos ou Variáveis de Entrada

A definição das variáveis de entrada baseia-se no princípio de que, em qualquer modelo de escolha discreta, os seguintes aspectos devem ser observados: as características sócio-econômicas do indivíduo que exerce a escolha e as características das alternativas de escolha (ORTÚZAR & WILLUMSEN, 1994). As características sócio-econômicas identificam o indivíduo e representam seus atributos. As características das alternativas sintetizam a relativa atratividade de cada alternativa de escolha, neste caso, escolha do modo de viagem. Assume-se, então, que as variáveis de entrada podem ser de dois tipos: atributos sócio-econômicos que correspondem ao conjunto de informações a respeito do indivíduo,

e atributos de viagem que, por sua vez, correspondem ao conjunto de informações sobre os disponíveis modos de realizar uma determinada viagem.

Alguns atributos de entrada foram obtidos diretamente da base de dados (tais) como os atributos sócio-econômicos: idade, sexo, grau de instrução; ou o atributo de viagem: motivo. Outros demandaram obtenção de forma indireta como por exemplo os atributos de viagem: custo por modo automóvel e tempo por modo caminhada. Ainda houve o caso da obtenção de atributos que demandaram uma amostragem, justificada pelo alto custo de aquisição de informação, como foi o caso para os atributos de viagem referentes ao modo ônibus.

Em seguida descrevem-se os atributos utilizados neste trabalho, com os respectivos nomes associados, destacados entre parênteses, que os representam nas estruturas das árvores de decisão, à diante neste texto apresentadas.

5.3.3.1 Atributos sócio-econômicos

Idade do viajante (**n_idade**). A idade apresenta-se como atributo contínuo com unidade expressa em anos

Sexo do viajante (**sexo**). Este atributo discreto recebe valor 1, ou "*true*", para o sexo masculino e valor 0 ou ,"*false*", para o sexo feminino.

Situação familiar (**sit_fam**). É a situação em que se encontra o viajante na hierarquia familiar. Este atributo discreto possui as seguintes categorias: (1) chefe de família; (2) cônjuge; (3) filhos; (4) parentes; (5) agregados; e (6) empregado ou residente.

Grau de instrução (**gr_inst**). Este atributo discreto determina o grau de escolaridade do viajante e divide-se nas categorias: (1) primário incompleto; (2) primário completo ou ginásio incompleto; (3) primeiro grau completo ou segundo grau incompleto; (4) segundo grau completo ou terceiro grau incompleto; (5) terceiro grau completo; e (6) analfabeto ou criança que não estuda. A divisão nestas categorias está em acordo com a antiga Lei Federal de Diretrizes e Bases da Educação n° 5692/71, vigente na data em que a pesquisa domiciliar foi realizada.

Carteira de motorista (**cart_mot**). Este atributo discreto determina se o viajante possui a carteira nacional de habilitação para dirigir veículo motorizado ou não.

Tamanho da família (**tam_fam**). É o número de pessoas pertencentes à família do domicílio. As cinco categorias deste atributo representam famílias de: (1) uma pessoa; (2) duas pessoas; (3) três pessoas; (4) quatro pessoas; (5) cinco pessoas ou mais.

Renda domiciliar (**n_salar**). É a soma das rendas, ou receitas, mensais dos membros da família no domicílio, declaradas pelo entrevistado na pesquisa domiciliar. Este atributo contínuo era disponível em valores da moeda vigente e foi convertido para número de salários mínimos, considerando o valor do salário mínimo da época (R$ 112,00).

Número de automóveis no domicílio (**n_autos**). Este atributo foi dividido em quatro categorias: (1) nenhum automóvel no domicílio; (2) um automóvel; (3) dois automóveis; (4) três ou mais automóveis.

5.3.3.2 Atributos de viagem

Motivo de viagem (**motivo**). Este atributo discreto indica o que motivou a realização de uma viagem e recebe seis possíveis argumentos: (1) trabalho, (2) escola; (3) compras; (4) saúde; (5) lazer; e (6) outros.

Distância de caminho mínimo (**dist**). Trata-se da menor distância entre o endereço de origem e o de destino da viagem, obtida através de determinada ferramenta do software *TransCAD v.3.2* que considera a rede viária da cidade para o cálculo desta distância. O valor de distância calculado pelo software de SIG está sujeito a um erro de 100 metros como já mencionado anteriormente. Embora este atributo não tenha sido utilizado diretamente como atributo de entrada para a modelagem, foi utilizado para o cálculo do tempo de viagem para os modos, como descrito a seguir.

5.3.3.2.1 O atributo tempo de viagem

Como já mencionado anteriormente, a pesquisa origem-destino não coletou informações a respeito das alternativas de modo de viagem disponíveis. Isto implica

em ter-se somente a informação do tempo de viagem, declarado pelo entrevistado, por modo caminhada, por exemplo, quando a viagem tiver sido realizada por este modo. A falta de informação do tempo de viagem para os modos alternativos, e a imprecisão da informação do tempo gasto em uma viagem, declarada pelo entrevistado, determinou a obtenção do atributo tempo de viagem de forma indireta.

Através do atributo distância de viagem, medida razoavelmente precisa por ser calculada em SIG a partir dos endereços de origem e de destino declarados pelo entrevistado, e considerando suposições a respeito de características médias inerentes a cada modo de viagem, por exemplo a velocidade média de caminhada e a velocidade média do ônibus em trecho urbano, torna-se possível a obtenção do atributo tempo de viagem para todos os modos. Se por um lado, essa forma indireta para o cálculo do tempo de viagem transpõe o problema de obtenção deste atributo, por outro prejudica o perfil desagregado da modelagem proposta, dado que atribui ao indivíduo uma propriedade provinda do agregado, a velocidade média. Um entrevistado, na pesquisa domiciliar, ao viajar por modo caminhada pode não ter desenvolvido a determinada velocidade média associada a todas as viagens por este modo e, por conseguinte, gastou um tempo diferente do calculado, por esta forma indireta, para a realização de sua viagem.

Abaixo segue a descrição das suposições e da formulação para cálculo do tempo para cada modo de viagem.

Tempo de viagem por modo caminhada (**t_ape**). É obtido dividindo-se a distância de caminho mínimo por uma velocidade média uniforme, que foi assumida como sendo 4,5 km/h, para quem caminha, e será utilizado na unidade horas (h). A suposição de que todos os viajantes caminham com mesma velocidade constante implica em dizer que o tempo de viagem por este modo é igualmente proporcional à distância de caminho mínimo para todos os viajantes. Vide Equação 5.1:

$$v = \frac{dist}{t} \quad \Rightarrow \quad t = \frac{1}{v} \times dist \quad \Rightarrow \quad \frac{1}{v} = a \quad \Rightarrow$$

$$t = a \times dist \tag{5.1}$$

onde v é a velocidade uniforme de caminhada; $dist$ é a distância de caminho mínimo; t é o tempo de viagem por modo caminhada; e a é uma constante.

Por ser um atributo contínuo, a normalização dos valores do atributo tempo de caminhada, para a simulação de RNA, torna-o indistinguível do atributo distância de caminho mínimo. Ou seja, como esses atributos são proporcionais, a rede recebe como entrada um valor normalizado que tanto poderia corresponder a um atributo ou a outro, já que a normalização do atributo tempo de caminhada desassocia o valor nominal, em termos de tempo de viagem, dos dados, guardando apenas a proporção entre seus valores e distribuindo-os dentro do intervalo 0 e 1. Sendo assim, a rede não "enxergaria" a diferença entre distância de caminho mínimo e tempo de viagem por caminhada, caso ambos fossem apresentados como atributos de entrada. Isto faz com que se possa "traduzir" o atributo tempo de caminhada para a distância viajada de caminho mínimo, na análise da árvore de decisão extraída de RNA treinada, quando se desejar.

Tempo de viagem por modo automóvel (**t_auto**). O tempo de viagem por este modo foi calculado a partir da distância de caminho mínimo (dist) e a velocidade média uniforme adotada de 30 km/h. Houve o acréscimo de 2,5 minutos ao cálculo do tempo para o automóvel, em consideração a desvios de trajeto ocasionados por sentido das vias urbanas e tempo gasto para estacionar o veículo.

Tempo de viagem por modo ônibus (**t_bus**). Para o cálculo do tempo de viagem por modo ônibus foi coletada a distância de viagem a bordo do ônibus, bem como as distâncias de acesso (do endereço de origem à linha de ônibus) e as distâncias de egresso (da linha de ônibus ao endereço de destino). As considerações a respeito de como foi feita a coleta destas distâncias encontram-se à frente descritas nesse texto. A soma das distâncias de acesso e egresso foi acrescida de 100 metros devido à imprecisão de localização de endereços no SIG. Para o cálculo da parcela de tempo a bordo do ônibus, a velocidade uniforme adotada para o ônibus foi de 15 km/h e para a parcela referente aos percursos de acesso e egresso, adotou-se novamente a velocidade uniforme de caminhada de 4,5 km/h. Foi acrescentada uma parcela de 10 minutos, de tempo de espera em ponto de parada, para todas as viagens. A Equação 5.2 expressa como foi calculado o tempo de viagem por ônibus:

$$\text{t_bus} = \frac{(d_acesso + d_egresso + 0{,}1)}{4{,}5} + \frac{d_abordo}{15} + \frac{10}{60} \qquad (5.2)$$

onde t_bus é o tempo de viagem por ônibus em horas; d_acesso é a distância de acesso; d_egresso é a distância de egresso; e d_abordo é a distância que o viajante percorreu viajando ônibus (distâncias em quilômetros).

Custo de viagem por modo automóvel (**c_carro**). Adotou-se a mesma formulação proposta por LOPEZ-REYES (1999), para se estimar o custo de viagem por automóvel (Equação 5.3):

$$\text{c_carro} = \left(\frac{(dist + 0{,}4)}{7}\right) \cdot 0{,}85 \tag{5.3}$$

onde c_carro é o custo monetário (R$); dist é a distância de caminho mínimo (km); 0,4 corresponde ao acréscimo devido ao sentido das vias urbanas; 0,85 (R$) era o preço do litro de combustível (gasolina), vigente no período da pesquisa domiciliar; e 7 corresponde ao consumo médio adotado para automóveis, expresso em km/l. Considera-se esta formulação como sendo uma aproximação do custo percebido pelo usuário de automóvel, pois, como menciona LOPEZ-REYES (1999), torna-se difícil adotar uma formulação exata para explicar os custos de viagem dos usuários do automóvel.

Custo de viagem por modo ônibus (**c_bus**). A coleta de dados a respeito do modo ônibus, feita para uma amostra e adiante descrita, obteve a informação de quantos ônibus o viajante tomaria, ou seja, quantas tarifas ele pagaria para realizar determinada viagem por este modo. A partir dessa informação foram consideradas as duas diferentes situações de cálculo do custo de viagem por ônibus, para cada viajante: os viajantes estudantes menores de 18 anos têm desconto de 30%; e as pessoas com idade superior a 65 anos viajam gratuitamente.

Tarifa dupla (**tarifa**). Esse atributo categórico recebe valor 1 quando haveria um transbordo de linha de ônibus, ou seja, haveria o pagamento de duas passagens de ônibus para a realização de uma viagem.

5.3.4 Filtragem dos Dados

Esta filtragem procurou remover o ruído nos dados que "distraem" um algoritmo de aprendizado, guiando-o erroneamente, em seu aprendizado da função

objetivo do problema (LIU & MOTODA, 2000), bem como excluir dados considerados inconvenientes ao enfoque do problema.

Toda a base de dados foi inspecionada à procura de registros com campos incompletos e foram excluídos tais registros. Nessa inspeção foram encontrados e excluídos também os registros que possuíam o campo renda domiciliar igual a zero.

Foram desprezados os registros de viagem com distâncias de menor caminho (dist) menores que 150 metros, devido à imprecisão do processo de localização de endereços, empregado no SIG, que pode errar a localização em torno de 50 metros para cada endereço de origem e de destino.

Foram desprezados, também, os registros de viagem referentes a indivíduos com idade inferior a cinco anos, por considerar-se que tais indivíduos não exercem escolha sobre o modo de viagem.

Por razão de insuficiência de número de casos foram desprezados os registros de viagem que possuíam o atributo grau de instrução (gr_inst) na categoria 6 (analfabeto ou criança que não estuda) e o atributo situação familiar (sit_fam) na categoria 7 (visitante não residente no domicílio).

Enfim, a filtragem dos dados reduziu a quantidade de registros disponíveis da base de dados de 8.695 para 7.608 viagens.

5.3.5 Amostragem

A vistoria executada na base de dados, descrita no item anterior, havia indicado que a distância de viagem por ônibus (a bordo), bem como as distâncias de acesso e egresso relativas a este modo de viagem, obtidas por processo automático no SIG, não correspondiam às respectivas reais distâncias viajadas obtidas manualmente, com o intuito de aferição, através do SIG. A imprecisão observada nestas medidas de distância relativas ao modo ônibus motivou uma obtenção não automática que garantisse a qualidade da informação obtida.

Pretendia-se inicialmente usar a base de dados em sua totalidade de registros para a modelagem proposta. Porém, a obtenção das medidas de distância do modo ônibus por um processo não automático, ou seja, manual para cada registro de viagem individual, se por um lado garantia a qualidade da medida, por

outro lado inviabilizava a utilização de todos os 7.608 registros de viagens disponíveis, devido ao enorme dispêndio de tempo que seria necessário à obtenção das medidas de distância para tal quantidade de viagens.

A solução encontrada foi a extração de uma amostra da base de dados para fins de utilização na modelagem proposta neste estudo. Assim, a quantidade reduzida de viagens tornaria possível a obtenção das medidas de distância por ônibus com qualidade, de maneira não automática, mas manual, ainda que com o auxílio do ambiente de SIG.

5.3.5.1 Tamanho da amostra

O procedimento utilizado para a determinação do tamanho da amostra apoia-se no Teorema do Limite Central, o qual postula que a estimativa da média de uma amostra tende a ser distribuído segundo a distribuição Normal, com o aumento do tamanho da amostra. Este procedimento encontra-se descrito em SMITH (1979) e em ORTÚZAR & WILLUMSEN (1994).

Com efeito, a base de dados com 7608 registros de viagem foi considerada como sendo a população a qual deseja-se inferir com a amostra. Para o cálculo do tamanho da amostra foram considerados: a variância dos atributos, utilizados na modelagem; o erro de estimativa e o intervalo de confiança desejados. As Equações 5.4 e 5.5 formulam, considerando população infinita, o cálculo do tamanho da amostra para atributos contínuos e atributos discretos, respectivamente:

$$n_o = z^2 \frac{S^2}{E^2} \tag{5.4}$$

$$n_o = z^2 \frac{p(1-p)}{E^2} \tag{5.5}$$

onde n_o é o tamanho da amostra; z é o coeficiente da distribuição Normal para o intervalo de confiança; S é o desvio padrão e S^2 é a variância do atributo contínuo; E é o erro de estimativa; e $p(1-p)$ é a variância da distribuição de Bernoulli, para atributos discretos.

A Tabela 5.7 apresenta os tamanhos de amostra calculados para o intervalo de confiança escolhido de 95% e para dois diferentes erros de estimativa, 10% e 5%.

Tabela 5.7: Tamanhos necessários de amostra.

Atributos	Média / p	Variância	Tamanho da amostra n_0	
			p/ erro de 10%	p/ erro de 5%
n_idade	32,757	314,444	113	450
sexo	0,496	0,250	96	384
sit_fam	0,390	0,238	91	365
gr_inst	0,289	0,206	79	316
cart_mot	0,464	0,249	96	382
tam_fam	0,324	0,219	84	337
n_salar	13,073	161,421	363	1451
n_autos	0,466	0,249	96	382
motivo	0,421	0,244	94	374
t_ape	0,650	0,214	195	780
t_auto	0,137	0,005	98	392
c_auto	0,404	0,064	151	604

De acordo com a tabela acima, o caso mais crítico é observado para o atributo renda domiciliar (n_salar), que determinou um tamanho de amostra de 1451 viagens, para um erro de estimativa de 5%.

Embora neste trabalho tenha-se decidido pelo emprego de uma amostra de 1451 viagens, não se deve considerar o erro de estimativa de 5% como o verdadeiro erro de estimativa para esta amostra, tendo em vista a modelagem por RNA.

A decomposição da amostra em partições, conforme já mencionado neste texto, para a simulação de RNA, implica que seja apresentada à fase de aprendizagem da RNA apenas a metade do tamanho da amostra, aproximadamente 730 viagens. Entretanto, observa-se na tabela acima que um tamanho de amostra de 730 viagens ultrapassaria a necessidade para todos os atributos, considerando um erro de estimativa de 10%. Além disto, vale notar que para uma amostra de 1451 viagens, as partições de validação e teste, com um

quarto do tamanho da amostra cada, aproximadamente 380 viagens, ultrapassam a necessidade para todos o atributos, também para um erro de estimativa de 10%.

Sendo assim, acredita-se que o erro de estimativa para a amostra de 1451 viagens está em algum lugar entre 5% e 10%, o que se considera satisfatório para a proposta deste trabalho.

5.3.5.2 Segmentação da base de dados e seleção de registros

Anteriormente à extração da amostra a partir da base de dados, por amostragem simples, randômica, onde a probabilidade de escolha de um registro da base de dados para participar da amostra é igual para todos os registros desta, foi feita uma segmentação da base de dados e seleção de registros exclusivos. A segmentação da base de dados é a divisão da base total (7.608 viagens) heterogênea em segmentos ou subconjuntos menores que sejam homogêneos nas características sócio-econômicas e de viagem relacionadas com o modo de viagem escolhido. A seleção de registros exclusivos é a formação de um subconjunto de registros, da base de dados já segmentada, que contenha um único registro representante de cada segmento.

Para a execução das tarefas de segmentação e seleção, foi atribuído um código a cada registro da base de dados que o representasse diante de todos os atributos, sócio-econômicos e de viagem, bem como diante do modo de viagem escolhido (Figura 5.4). Identificados os códigos iguais, que por sua vez representavam um segmento com características semelhantes, tanto para atributos quanto para modo escolhido, selecionou-se, de forma aleatória, apenas um registro para representar aquele segmento.

Essa seleção por registros exclusivos justifica-se pelo alto custo de coleta de informações para cada registro constituinte da amostra, e pela característica particular do processo de aprendizagem de máquina, do modelo de RNA escolhido. Tal processo de aprendizagem, o treinamento da RNA, compõe-se de subseqüentes unidades chamadas épocas, ou ciclos. Em cada época ou ciclo de treinamento todas as instâncias de um conjunto de treinamento são apresentadas uma vez para a rede. A apresentação de uma instância de treinamento repetida não

agregaria informação ao processo de aprendizagem da rede, porém contribuiria para a redução da diversificação de informação contida na amostra.

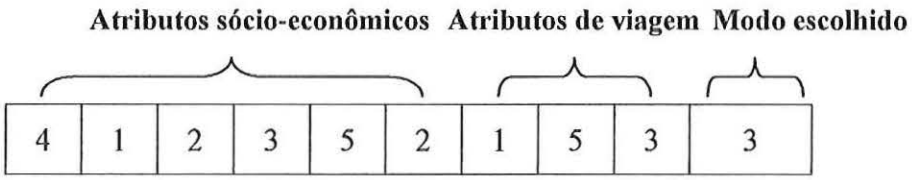


Figura 5.4: Exemplo de código para um registro de viagem.

No código, exemplificado na figura acima, cada atributo está representado em uma posição e o algarismo, dentro da posição, associa a qual categoria, daquele atributo, um registro de viagem pertence. Mesmo os atributos contínuos foram subdivididos em categorias para que fosse feita a representação de um registro de viagem por um código. Caso os atributos contínuos fossem representados por seus valores reais, em suas devidas posições no código, provavelmente não haveriam códigos iguais, o que não selecionaria registros participantes de um segmento com características semelhantes.

A representação dos atributos contínuos em subcategorias, ou seja, de forma discreta, está relacionada com a quantidade de registros considerados iguais. Se o intervalo de valores reais de um atributo contínuo fosse dividido somente entre duas categorias, ou ainda, todo locado em uma única categoria, seria possível dizer que , em relação àquele atributo, muitos registros são iguais, ou ainda, todos os registros são iguais. Para ilustrar a relevância da escolha das subcategorias para atributos, nesta seleção, pode-se compará-la à escolha da malha de uma peneira para o peneiramento de materiais sólidos, onde o espaçamento dos fios entrançados da malha determinam quais sólidos passam e quais são retidos no peneiramento.

O código utilizado compunha-se de 11 posições que se referiam a combinação de 10 atributos e o modo de viagem escolhido. No apêndice deste texto encontra-se uma tabela que apresenta os atributos usados para a composição do código de seleção, bem como as correspondentes subcategorias definidas.

O selecionar de um único registro para representar um segmento, da base de dados, com características semelhantes, faz com que a informação sobre

eventuais modas (tendências) de comportamento agregado, que possam estar ocorrendo na população, não cheguem à fase de treinamento da RNA, ou seja, tal informação não é "enxergada" ou mapeada pela RNA em sua aproximação da função objetivo. A Figura 5.5 ilustra a situação mencionada, para a seleção dos registros. Este estudo não se preocupou com a perda de informação a respeito do comportamento agregado, uma vez que o enfoque é dado sobre o comportamento individual, ou desagregado, de escolha. Todavia, o assunto recebe novamente atenção nas discussões e considerações finais deste trabalho.

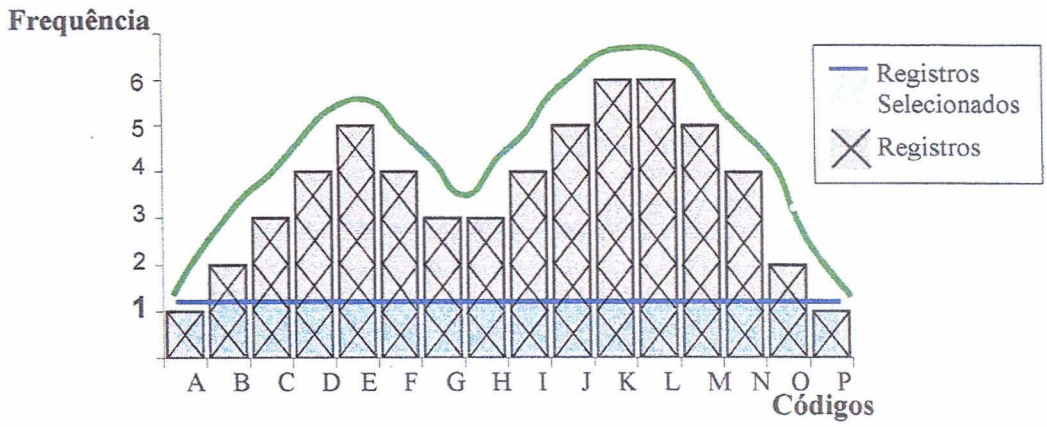


Figura 5.5: Gráfico ilustrativo da seleção de registros codificados.

Como resultado da seleção na base de dados de 7.608 registros de viagens obteve-se um subconjunto de 6.100 registros exclusivos. A partir desse conjunto de registros exclusivos procedeu-se à extração de uma amostra de 1.451 registros de viagem, por amostragem simples. Sorteou-se, de maneira aleatória, uma quantidade próxima do tamanho desejado da amostra, e subseqüentes sorteios aleatórios completaram quantidades equivalentes de registros para cada modo de viagem. Sendo assim, a amostra apresenta-se balanceada para a quantidade de dados pertencente a cada classe de modo de viagem, que se distribuem da seguinte forma: 362 registros de viagem por modo automóvel, 348 por modo carona, 367 por modo caminhada e 374 por modo ônibus. Por interação gráfica através das Figuras 5.6 e 5.7 geradas em SIG, pode-se examinar a distribuição espacial dos dados amostrados em relação a toda a base de dados.

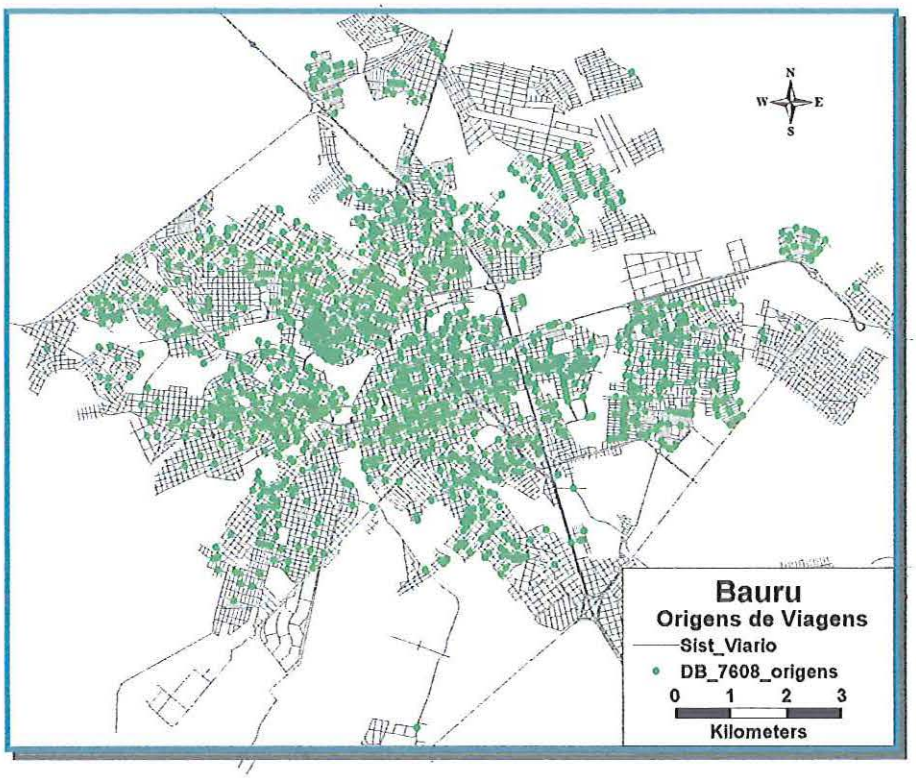


Figura 5.6: Localização das origens das viagens de toda a base de dados.

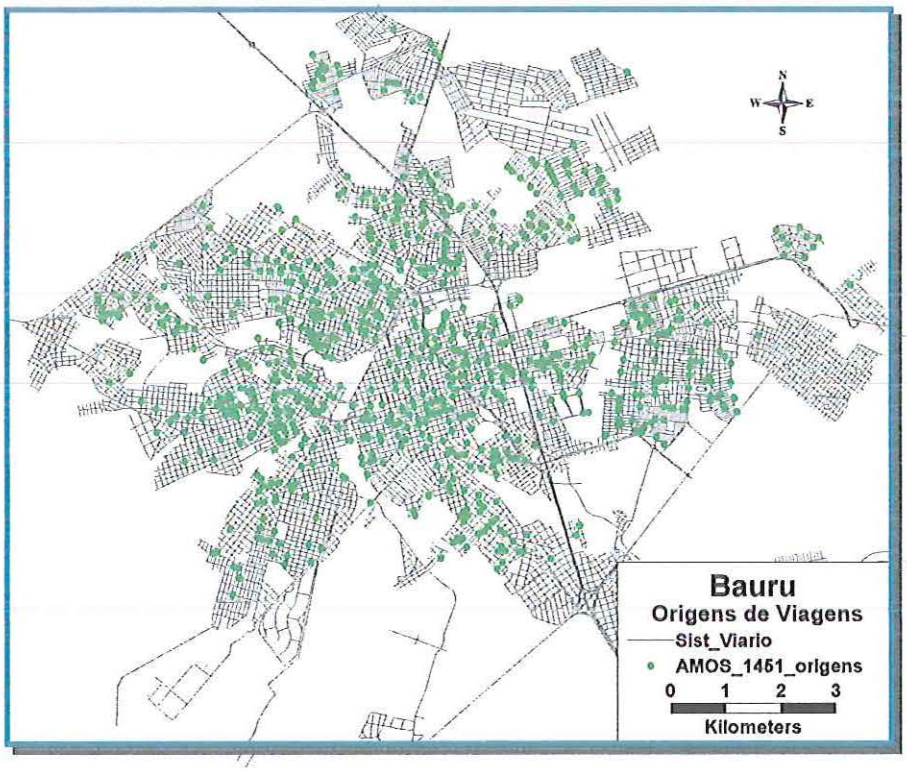


Figura 5.7: Localização das origens das viagens da amostra.

5.3.5.3 Coleta de distâncias de viagem por modo ônibus

Como já mencionado acima no item sobre amostragem, pretendia-se inicialmente utilizar todos os 7.608 registros de viagem, disponíveis da base de dados, para a modelagem por RNA. No entanto, um problema na obtenção das distâncias de viagem para o modo ônibus demandou a utilização de uma amostra de tamanho menor (1.451 viagens).

As medidas de distância de viagem por modo ônibus foram inicialmente obtidas para todas as viagens (7.608) por um processo automático no SIG que considerou uma rede formada pelos trechos de vias nos quais circula pelo menos uma rota de transporte público (Figura 5.8). Neste processo, não foram impostas restrições de nenhuma natureza para transferências entre rotas, o que não reflete exatamente a realidade, uma vez que as transferências, ou transbordos, que ocorrem no mundo real são limitadas por uma série de fatores (ver MARQUES, 1998). Além do fato desse processo automático não ser capaz de obter a quantidade de tarifas pagas, referentes a eventuais transbordos, uma vistoria por sorteio aleatório para aferição, que comparou algumas distâncias estimadas por tal processo contra distâncias tomadas de forma manual no SIG, indicou imprecisão em relação a todas as distâncias: a bordo, de acesso e de egresso.

Decidiu-se, então, fazer a coleta de informações para modo ônibus manualmente, com o auxílio do SIG, para uma quantidade reduzida de viagens, o tamanho da amostra (1.451 viagens). Esse procedimento manual procurou obedecer, tanto quanto possível, a situação real de utilização do sistema de transporte público da cidade em estudo. Para cada viagem definia-se no mapa viário da cidade os pontos de origem e de destino. Em seguida, testava-se uma a uma as 85 linhas de ônibus do sistema à busca de uma linha que melhor se

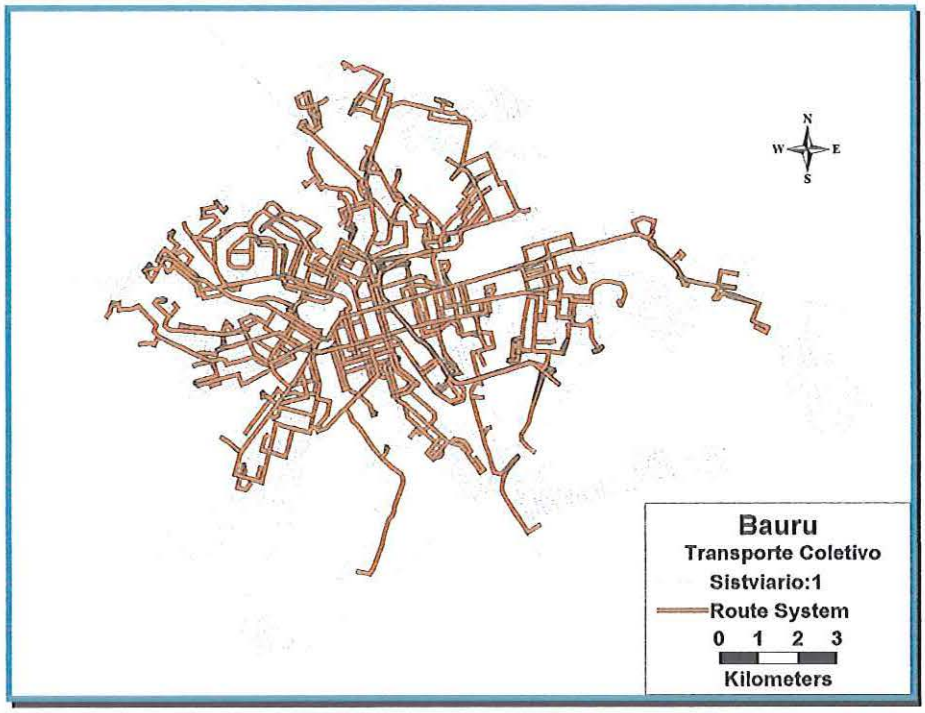


Figura 5.8: Rede de transporte público formada por sobreposição das 85 rotas (linhas) de ônibus existentes, em destaque sobre a rede do sistema viário. Fonte RAIA Jr. (2000)

ajustasse ao atendimento daqueles pontos. Uma vez encontrada uma linha que servisse àquela viagem (Figura 5.9): media-se a menor distância de acesso do ponto de origem até a linha, considerando a rede viária; da mesma forma media-se a menor distância de egresso da linha até o ponto de destino; tomava-se a medida de distância a bordo, considerando a linha; e guardava-se a informação de que uma tarifa havia sido paga. Caso fosse constatado a necessidade da utilização de duas linhas de ônibus para atendimento de uma viagem (Figura 5.10): da mesma forma descrita acima, mediam-se às distâncias de acesso e egresso; ao medir-se a distância a bordo neste caso, a transferência de linha era considerada somente na posição onde houvesse a interferência da próxima linha, ou por sobreposição ou por cruzamento. O procedimento acima descrito foi repetido aproximadamente mil e quinhentas vezes, ou seja, para todos as viagens da amostra.

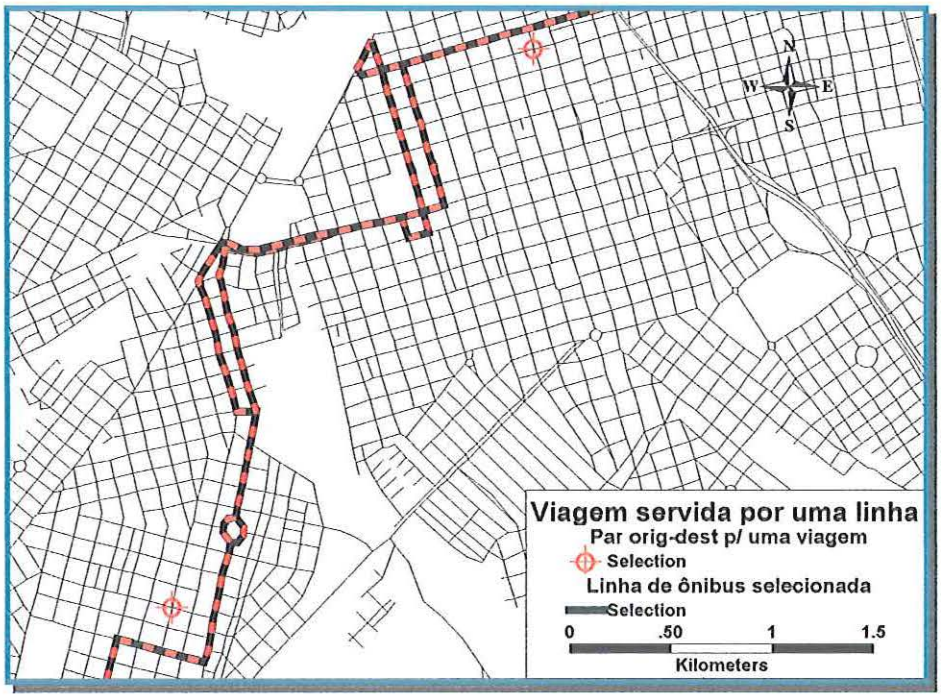


Figura 5.9: Exemplo de viagem realizada utilizando uma linha de ônibus.

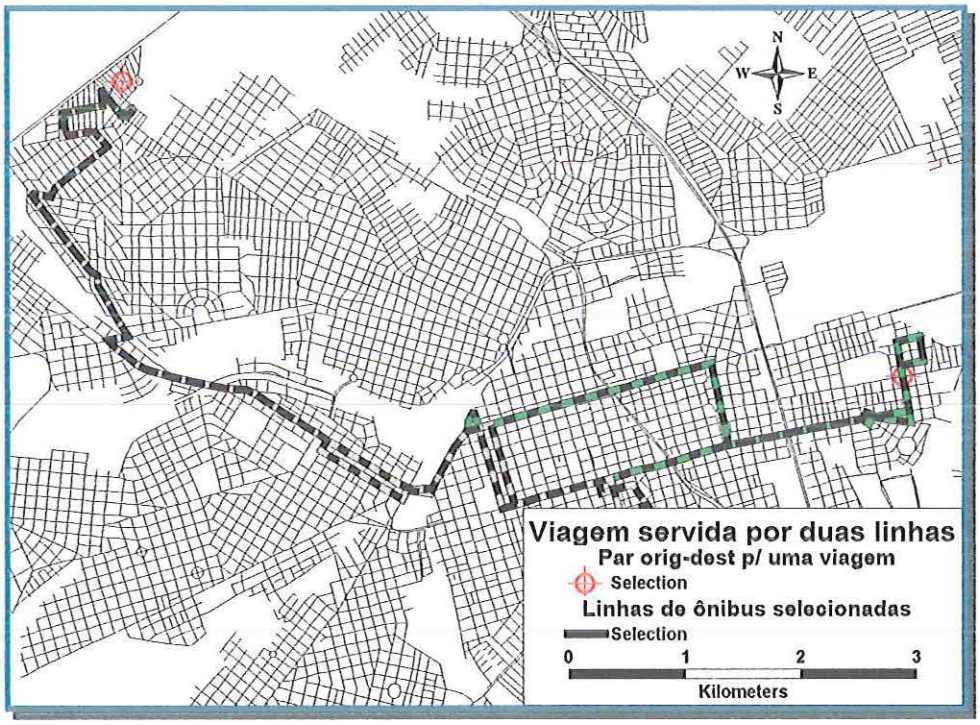


Figura 5.10: Exemplo de viagem realizada utilizando duas linhas de ônibus.

Apesar das rotas, ou linhas, do sistema de transporte público implementadas no SIG não conterem informação a respeito do sentido de itinerário, foram consultados os itinerários disponíveis em EMDURB (2001) para orientação do sentido das linhas utilizadas na coleta de distâncias de viagem por ônibus, em favor da melhoria de acurácia das medidas de distância coletadas.

Uma vez coletadas as distâncias de viagem por modo ônibus para os 1.451 registros da amostra, tornou-se possível o cálculo dos dois atributos referentes a este modo e conferiu-se o erro de estimativa para estes. A tabela abaixo mostra que os erros de estimativa calculados estão bem abaixo do valor de 10%, considerado como erro de estimativa satisfatório para a amostra.

Tabela 5.8: Erros de estimativa para intervalo de confiança de 95% e tamanho de amostra de 1.451 elementos.

Atributos	Média	Variância	Erro de estimativa %
t_bus	0,520	0,0623	2,423
c_bus	0,668	0,2218	2,470

Para maiores detalhes deste problema de obtenção de distâncias de viagem por transporte público urbano, bem como uma análise comparativa entre as distâncias obtidas manualmente e aquelas obtidas por processo automático do SIG, ver WERMERSCH & SILVA (2001).

6 MINERAÇÃO: OBTENÇÃO E ANÁLISE DE RESULTADOS

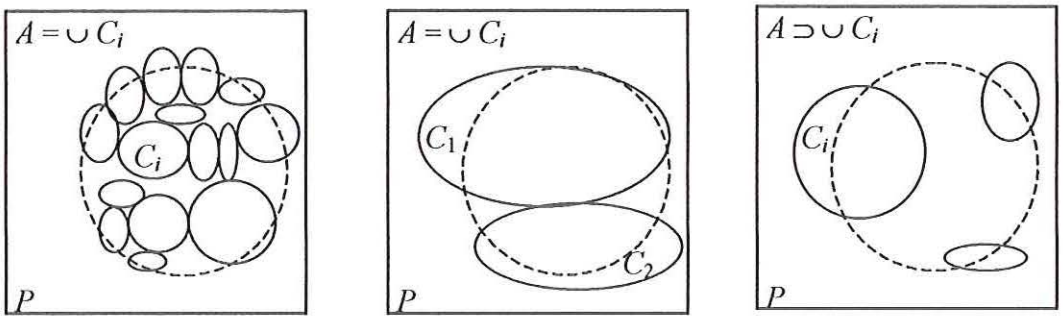
Este capítulo se refere à etapa de *data mining* (Fig.: 3.1) do trabalho ou mais especificamente às fases de simulação de rede neural e extração da árvore de decisão da rede (Fig.:4.1). A análise de resultados aqui limita-se à verificação da acurácia preditiva das redes e correspondentes árvores extraídas, bem como da simplicidade da estrutura dessas árvores. Encontra-se no capítulo seguinte a análise e interpretação do conhecimento encontrado, como resultado final deste trabalho.

6.1 DIRETRIZ DE BUSCA DE PARADIGMAS

A busca pela rede treinada e sua correspondente árvore extraída visou não apenas a performance de classificação mas principalmente a compreensibilidade da estrutura de conhecimento extraído. A esse respeito QUINLAN apud GAINES (1996) identifica expressões de conceito, resultantes de um processo de KDD, como sendo aquelas descrições efetivamente computáveis e corretas que podem também ser assimiladas e usadas por um ser humano. Como um exemplo, o autor cita um caso onde o algoritmo de indução de árvores de decisão ID3 (QUILAN, 1986), antecessor de mesmo autor do algoritmo C4.5, derivou uma árvore para o domínio do jogo de xadrez a qual estava absolutamente correta e computacionalmente eficiente, mas completamente incompreensível para os especialistas em jogos de xadrez.

No sentido de procurar por árvores de decisão de melhor compreensibilidade foi utilizada a profundidade da árvore como critério de medida, abordando assim, apenas a complexidade sintática do achado. Busca-se uma

estrutura de conhecimento representativa e pequena, com poucos e abrangentes conceitos, que descreva com razoável acurácia uma estrutura subjacente ao espaço de instâncias e ao mesmo tempo que seja de fácil compreensão para um analista. A Figura 6.1 ilustra essa situação para, por exemplo, um caso binário, onde somente a classe positiva A (conjunto alvo) para uma população P deveria ser coberta pelo conjunto de conceitos C_i.



(a)Classificação/Previsão (b)Descrição (c)Fragmentos

Figura 6.1: Situações típicas de conceitos descobertos para uma classe alvo. Fonte: KLÖSGEN (1996).

Através da análise do tamanho da árvore gerada, o que determina a quantidade de conceitos obtidos e sua correspondente acurácia preditiva, procurou-se uma aproximação da situação (b) da figura acima, afastando-se da situação (a) e evitando-se a situação (c).

6.2 A BUSCA

Os experimentos realizados para obtenção das redes treinadas utilizam três conjuntos de dados, cada um contendo três partições, treinamento, validação e teste, conforme definido no capítulo 4 deste texto. Um conjunto de dados para os quatro modos de viagem possui aproximadamente 750 instâncias na partição de treinamento e 350 instâncias em cada partição de validação e teste. Pelo fato de haver balanceamento entre as quantidades de instâncias pertencentes a cada modo, a remoção de um modo de viagem dentre os quatro existentes, e

---

QUINLAN, J.R. Foreword. In: *Knowledge Discovery in Databases*, XI-XII. Cambridge, MA, MIT Press, 1991 apud GAINES (1996), p.205.

três modos de viagem possua aproximadamente 560 e 260 instâncias nas partições de treinamento e validação/teste respectivamente.

Em cada evento de treinamento que utiliza um desses conjuntos de dados, uma determinada arquitetura passa por seis treinamentos independentes, cada um inicializado com uma diferente taxa de aprendizagem. A taxa de aprendizagem para o primeiro treinamento de uma arquitetura foi definida como 0,1, para o segundo como 0,2 e assim sucessivamente até a maior delas, que é 0,6. Esses valores de taxa de aprendizagem foram escolhidos pela observação das curvas de erro, de treinamento e validação em vários eventos de treinamento realizados através da interface gráfica do simulador.

Nos experimentos realizados, uma rede é nomeada de acordo com a disposição das unidades de processamento nas camadas, ou seja, de acordo com sua arquitetura. Por exemplo a rede 52-60-4 significa que possui 52 unidades de entrada, 60 unidades intermediárias e 4 unidades de saída.

O percentual médio de acerto à classificação de uma arquitetura de rede, mostrado nas figuras abaixo, se refere à média entre os melhores desempenhos observados para cada um dos três conjunto de dados utilizados no processo de treinamento.

6.2.1 Resultados preliminares

Conforme já mencionado, inicialmente foram incluídos os quatro modos de viagem, carro, carona, a pé e ônibus, para a tarefa de classificação pelo modelo de RNA. Porém a baixa performance de todas as arquiteturas de RNA testadas em relação ao modo carona, mostrada a seguir, conduziu a decisão de excluí-lo da classificação.

O primeiro experimento foi conduzido com todos os atributos de entrada disponíveis, relacionados no capítulo 5, alistados na Tabela D.1 do anexo D, e envolveu sete diferentes propostas de arquitetura de rede submetidas a dezoito treinamentos cada, o que gerou 126 simulações. O gráfico apresentado na Figura 6.2 mostra os resultantes níveis de acerto médio entre classes e para cada uma delas distintamente, nas diferentes arquiteturas.

6.2 mostra os resultantes níveis de acerto médio entre classes e para cada uma delas distintamente, nas diferentes arquiteturas.

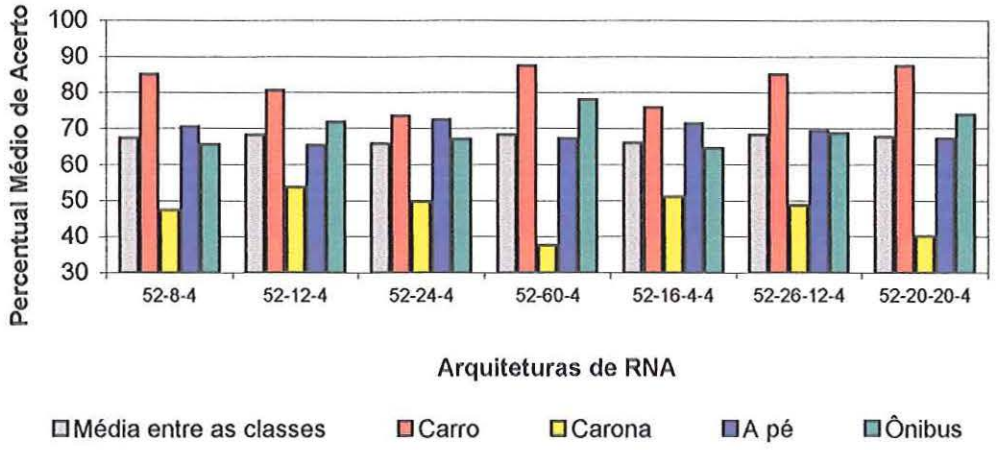


Figura 6.2: Gráfico de resultados do experimento 1 para 4 modos de viagem.

Na figura acima, o desempenho médio apresentado pelas arquiteturas de rede foi de 67,4%, destacando-se no quesito desempenho-desvio a rede 52-12-4 com nível de acerto médio de 68,3% e desvio padrão entre classes de 9,3%. Observa-se em todas as redes que a classe carona possui um baixo nível de acerto à classificação em relação às outras classes.

À procura de uma melhoria no nível de acerto da classe carona outros experimentos foram realizados. Para o experimento seguinte foi criado um novo atributo que relacionasse características observadas na maioria dos viajantes do modo carona. Esse atributo sintético e categórico, chamado "caronista", recebe valor 1 quando o viajante perfaz simultaneamente duas características, não possui carteira de motorista e é do sexo feminino, e valor 0 em caso contrário. A partir deste experimento e em todos os seguintes não se inclui mais o atributo "tarifa dupla" por considerar-se que esta informação já está contida no atributo "custo de viagem por ônibus". A Figura 6.3 mostra os resultados para este segundo experimento conduzido com os atributos de entrada mostrados na Tabela D.2.

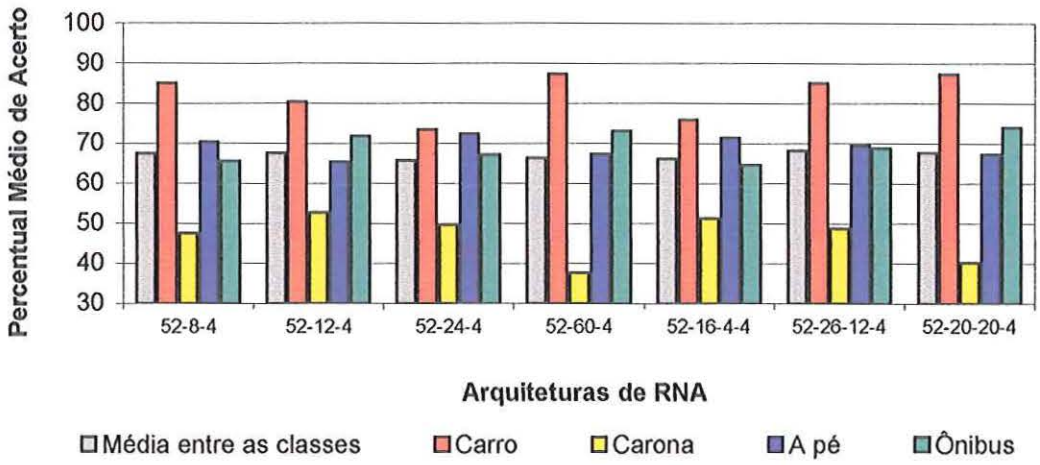


Figura 6.3: Gráfico dos resultados do experimento 2 para 4 modos de viagem.

Os resultados do gráfico acima mostram que as redes tiveram desempenho similar ao experimento anterior, com desempenho médio de 65,2% sendo novamente a arquitetura de melhores índices de desempenho a 52-12-4, com 67,6% de acerto e desvio padrão de 9,3%. Tal fato sugere que a remoção do atributo tarifa não interferiu, e tampouco a inclusão do atributo caronista contribuiu para a melhoria de performance das redes para a classe carona ou para as classes em geral.

Para um seguinte experimento foram removidos os atributos motivo de viagem e distância e mantidos todos os outros atributos que participaram do experimento anterior (vide Tabela D.3). A Figura 6.4 mostra o gráfico com os resultados para esse terceiro experimento que envolveu 11 arquiteturas diferentes de RNA.

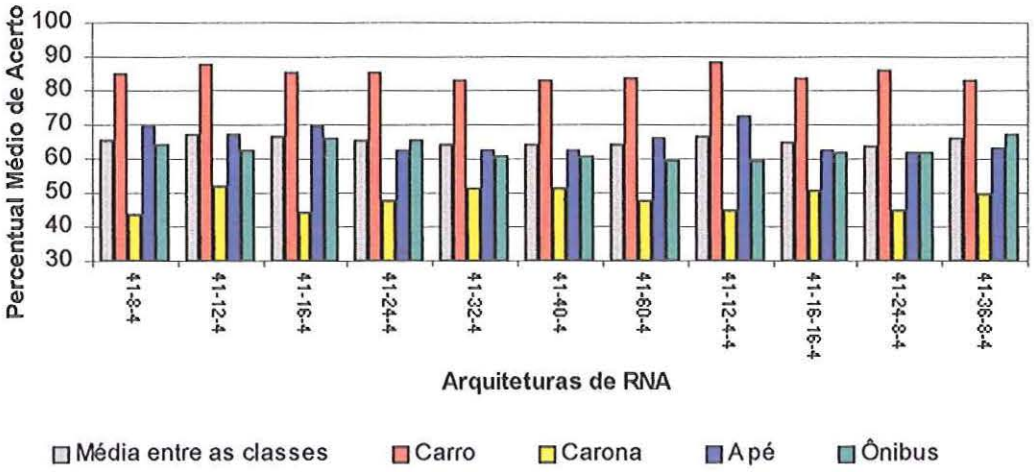


Figura 6.4: Gráfico dos resultados do experimento 3 para 4 modos de viagem.

Nesse terceiro experimento o desempenho médio ficou em 65,3% e a vencedora entre as arquiteturas foi a 41-12-4 com percentual de acerto de 67,6% e desvio padrão entre classes de 14,3%. Apesar da remoção dos dois atributos o desempenho das redes investigadas não aumentou tanto em termos médios quanto para a classe carona, que manteve os mais baixos percentuais de acerto entre as classes.

Os atributos discretos ordinais utilizados nesses três primeiros experimentos não apresentavam codificação que correspondesse à ordenação presente entre suas categorias. Então foi preparado um novo experimento com conjuntos de dados com codificação para atributos discretos ordinais conforme definido no capítulo 4, ítem 4.3.1 "Normalização e Codificação", deste texto. Para este quarto experimento foram considerados novamente todos os atributos com exceção apenas do atributo tarifa, conforme mostra a Tabela D.4. O gráfico da Figura 6.5 mostra os resultados deste experimento com 11 arquiteturas diferentes e 198 simulações.

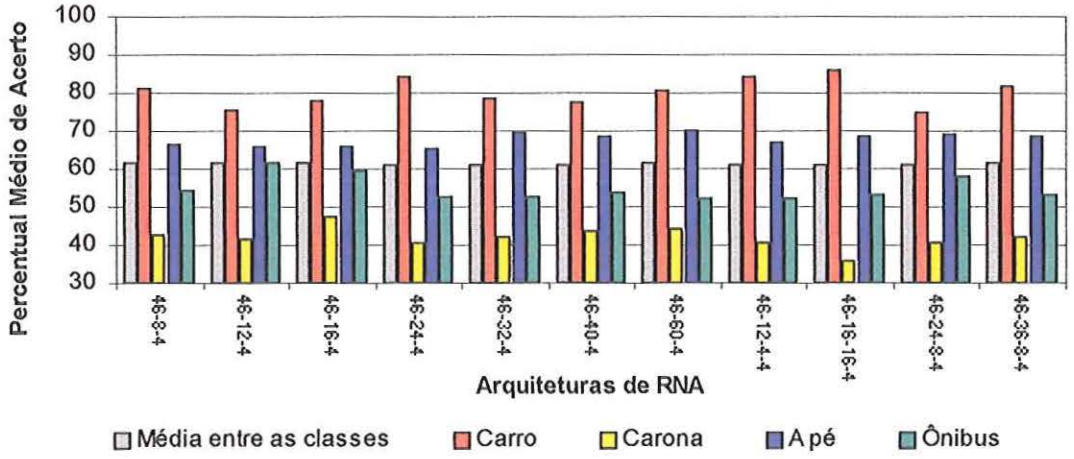


Figura 6.5: Gráfico de resultados do experimento 4 para 4 modos de viagem.

Houve, como mostra o gráfico acima, uma pequena piora no desempenho médio das redes, que ficou em 61,3% percentuais de acerto à classificação. Desta vez, para a classe carona, observa-se que nenhuma rede sequer atingiu o nível de acerto de 50%.

Uma vez que nos quatro experimentos realizados alguns atributos contínuos estavam representados em categorias, como por exemplo o atributo renda domiciliar, decidiu-se então transformar tais atributos de acordo com o definido no sub-ítem 5.3.3.1 deste texto. Tal transformação reduziu o número de unidades de entrada demandado nas redes para 31, sendo considerados todos os atributos de entrada com exceção dos atributos tarifa e distância de caminho mínimo (vide Tabela D.5). Seguiu-se a execução de uma busca extensiva à procura de arquiteturas de redes que obtivessem melhores resultados, principalmente para a classe carona. Esta busca extensiva configura um quinto experimento que foi realizado envolvendo 49 arquiteturas de rede e 882 simulações. As Figuras 6.6, 6.7, 6.8, 6.9 e 6.10 apresentam os resultados deste experimento.

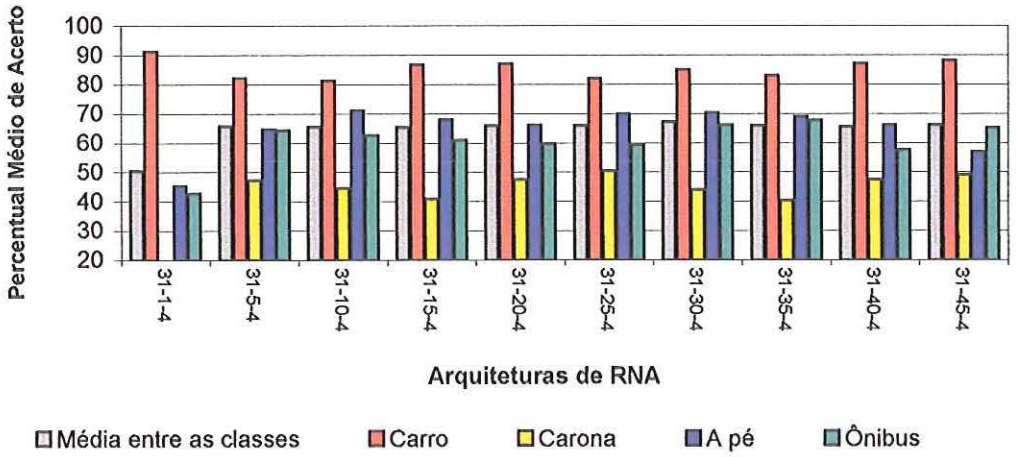


Figura 6.6: Gráfico dos resultados do experimento 5 para 4 modos de viagem.

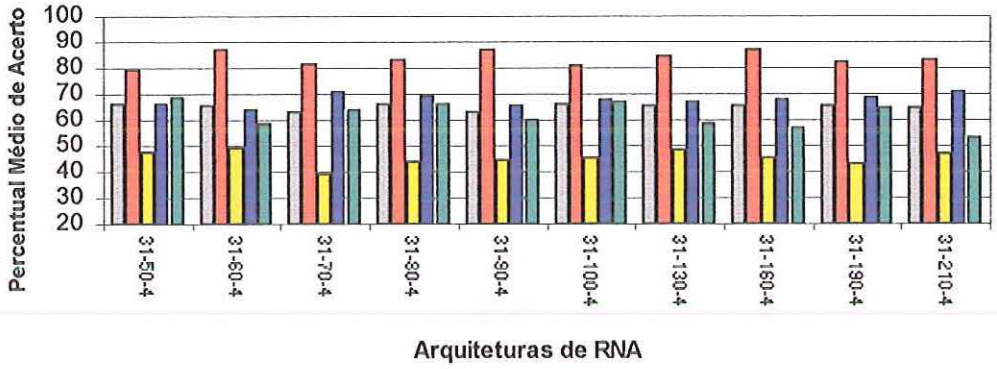


Figura 6.7: Continuação do gráfico da Figura 6.6.

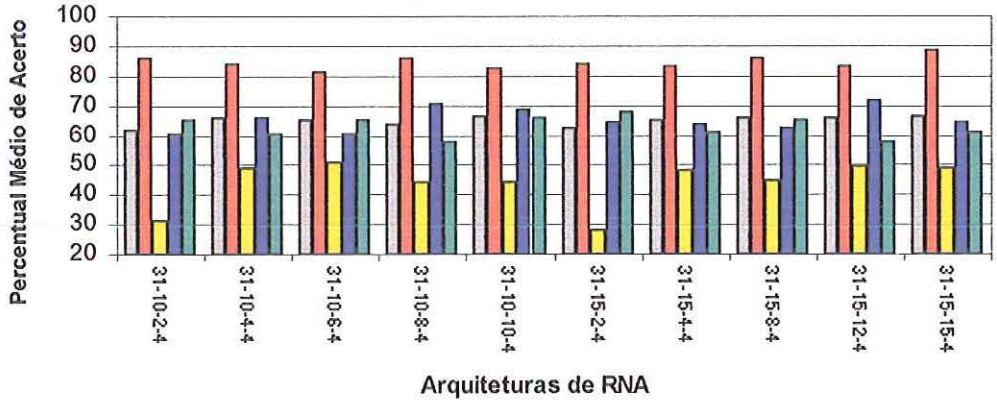


Figura 6.8: Continuação do gráfico da Figura 6.6.

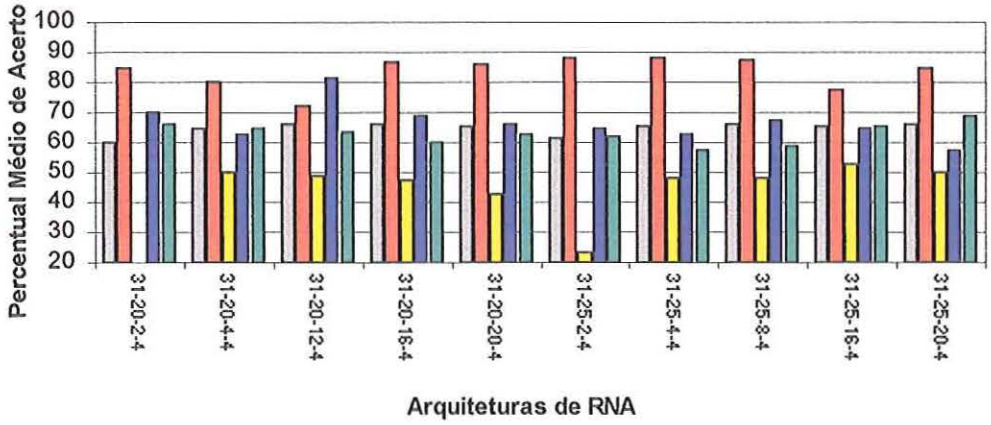


Figura 6.9: Continuação do gráfico da Figura 6.6.

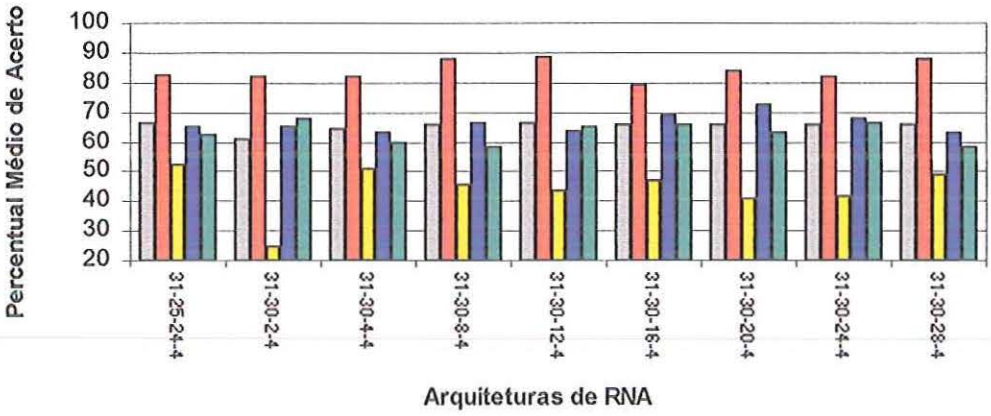


Figura 6.10: Continuação do gráfico da Figura 6.6.

O desempenho médio para as arquiteturas com uma camada intermediária foi de 64,75%, sendo a vencedora a rede 31-30-4 com acerto médio de 67,2% e desvio padrão entre classes de 14,9%. Entre as arquiteturas de duas camadas o desempenho médio foi de 64,8% e a vencedora, a arquitetura 31-25-24-4, obteve percentual de acerto de 66,1% e desvio entre classes de 11,9%. Esses números revelam que não houve mudança significativa no desempenho das redes e os gráficos resultantes desse quinto experimento mostram que também não houve, em relação aos experimentos anteriores, uma melhora na classificação da classe carona por nenhuma rede investigada. Como já se havia observado em outros

experimentos, somente em algumas arquiteturas o percentual de acerto médio para a classe carona apenas ultrapassa em pouco o nível de 50%.

Na busca por alguma informação que auxiliasse a seleção dos atributos de entrada das redes para uma maior eficácia na tarefa de classificação, lançou-se mão de outro software simulador de RNA chamado *EasyNN*[1] versão 7.5, que fornece um gráfico de importância relativa das entradas de uma rede. Esse gráfico é gerado pelo simulador através de um método com abordagem local que embora não seja divulgado pelo fabricante, uma vez que o software tem distribuição comercial apenas, foi considerado similar ao método publicado por GARSON (1991) para interpretação dos pesos das conexões de uma rede neural, num estudo comparativo conduzido por SILVA et al. (2000).

Como menciona VERLEYSEN (1998), a informação provinda da interpretação dos parâmetros de um modelo de rede neural deve ser vista como um expediente suplementar de informação sobre a rede, o que concorda com o emprego desse expediente, neste estudo, como apenas à procura de uma "dica" sobre a seleção dos atributos de entrada do modelo neural.

Uma rede de arquitetura 31-10-4 foi submetida a três processos de treinamento utilizando os conjuntos de dados do quinto experimento, anteriormente referido, no simulador EasyNN conforme configuração disponível neste, como por exemplo taxa de aprendizado decrescente e crescimento do número de unidades da camada intermediária a partir da arquitetura inicial. Os três treinamentos resultaram em uma arquitetura 31-11-4 com gráficos de importância de entradas semelhantes, com diferença apenas nas últimas posições de importância. A Figura 6.11 mostra o gráfico para um desses treinamentos.

[1] WOLSTENHOLME, S. (1999). *EasyNN version 7.5 – neural network simulator.* England.

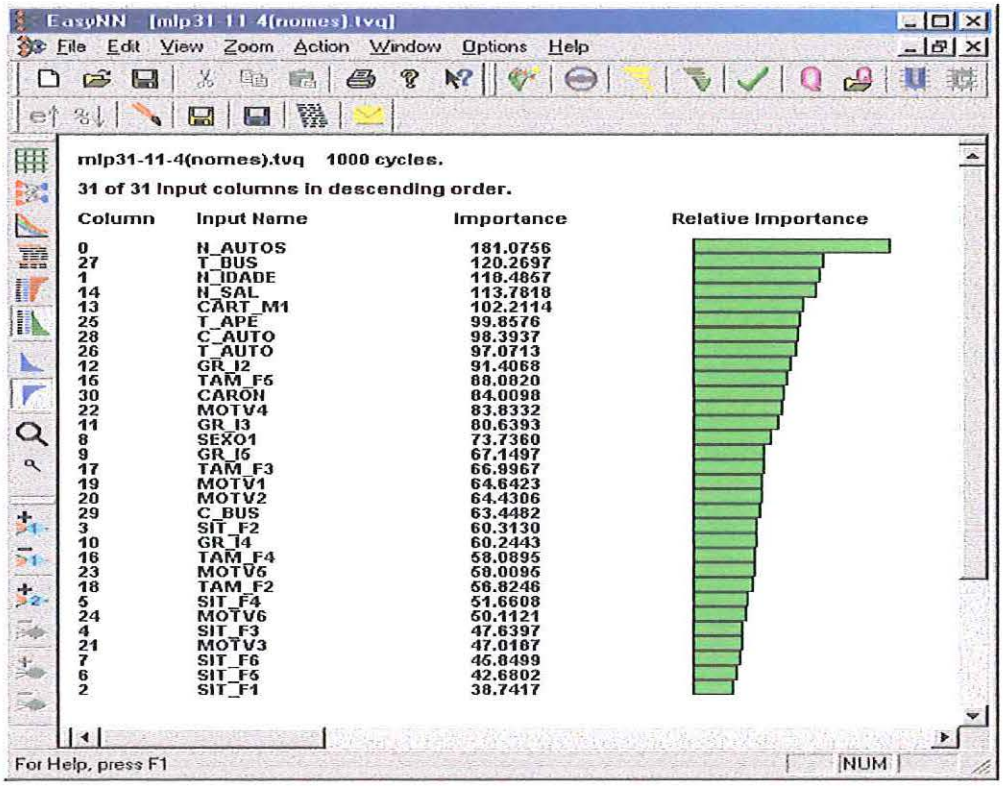


Figura 6.11: Gráfico de importância relativa das entradas da rede 31-11-4.

O gráfico de importância relativa entre os atributos de entrada acima mostra que os atributos de viagem situação familiar e motivo de viagem podem não estar contribuindo significativamente para a classificação dos modos de viagem pela rede neural. Decidiu-se então pela realização de mais três experimentos.

Um sexto experimento envolveu todas os atributos presentes no quinto experimento com exceção do atributo de viagem situação familiar (vide Tabela D.6), o que demandou 25 unidades de entrada às arquiteturas de rede. A Figura 6.12 mostra o gráfico com os resultados para as 9 arquiteturas de rede investigadas neste experimento.

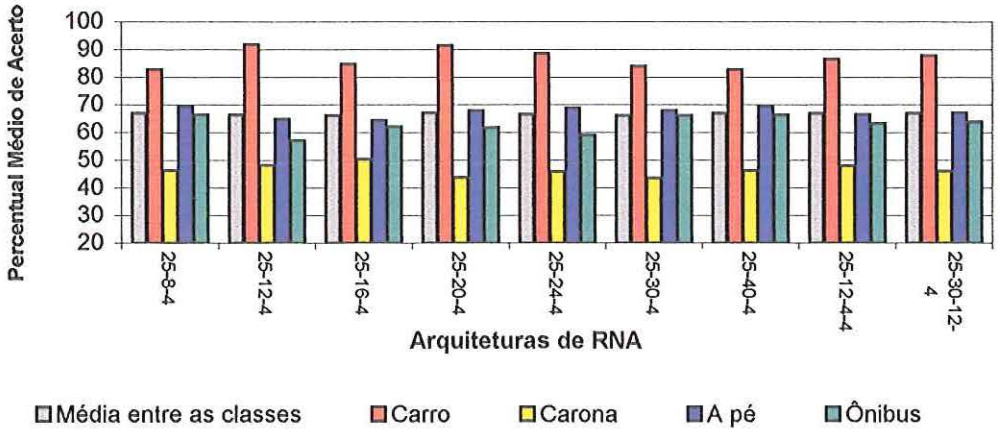


Figura 6.12: Gráfico dos resultados do experimento 6 para 4 modos de viagem.

O desempenho médio das redes neste sexto experimento ficou em 66,7% com desvio padrão entre classes 15,0%. Em relação ao experimento anterior, apesar do resultado ter sido ligeiramente superior, não houve nenhuma rede com melhoria significativa para o desempenho médio, inclusive para a classe carona.

Um sétimo experimento contou com dois novos atributos além daqueles presentes no quinto experimento (vide Tabela D.7). Os atributos sócio-econômicos trabalha e estuda referem-se ao tipo de ocupação do viajante. Cada um destes atributos discretos nominais recebe valor 1 quando o viajante possuir a ocupação correspondente ao atributo, e valor 0 em caso contrário. O gráfico da Figura 6.13 mostra os resultados para as 10 arquiteturas de rede investigadas no sétimo experimento.

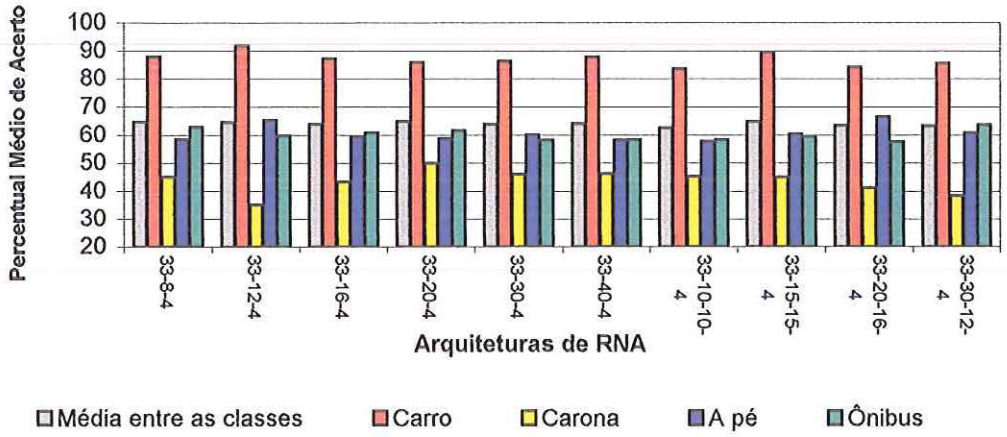


Figura 6.13: Gráfico dos resultados do experimento 7 para 4 modos de viagem.

Neste sétimo experimento o desempenho médio para as redes investigadas foi de 64,0% com desvio padrão entre classes de 16,5%, o que não demonstra uma melhoria de desempenho em relação aos experimentos anteriores.

O oitavo e último experimento realizado para os quatro modos utilizou um conjunto de atributos de entrada semelhante ao do experimento anterior, porém sem os atributos caronista, motivo de viagem e situação familiar (vide Tabela D.8). Os resultados para as 10 arquiteturas de rede submetidas ao experimento encontram-se na Figura 6.14.

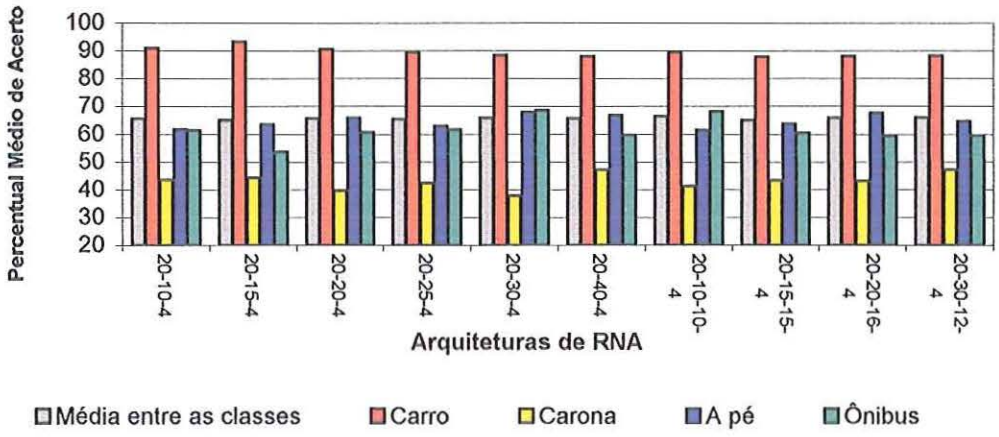


Figura 6.14: Gráfico dos resultados do experimento 8 para 4 modos de viagem.

As redes deste oitavo experimento tiveram desempenho médio de 65,7% e desvio padrão de 17,1% entre classes. Novamente os resultados se assemelham aos experimentos anteriores, não havendo alguma rede com desempenho destacável tanto em termos médios quanto para a classe carona.

Nos oito experimentos realizados, que envolveram 114 diferentes arquiteturas de rede, totalizando 2.052 simulações, o modo de viagem carona apresentou, sistematicamente em todas as redes investigadas, percentuais de acerto na classificação inferiores aos dos outros modos. Dentre todas as redes treinadas os melhores desempenhos observados para o modo carona classificam corretamente pouco mais da metade das instâncias relativas a este modo. Provavelmente a suposição de que a carona seria uma alternativa de modo de viagem disponível em todas as viagens coletadas pela pesquisa domiciliar é falha e não contorna a falta de informação nesta pesquisa sobre a disponibilidade de

modos à escolha do viajante. Ou ainda, pode haver falta de informação suficiente na pesquisa domiciliar para a caracterização do modo carona, pelo fato desta pesquisa, como afirma LOPEZ-REYES (1999), não ter tido seu desenvolvimento voltado a propiciar uma modelagem de escolha modal.

Como objetiva-se a obtenção de um modelo que seja representativo para todos os modos disponíveis que forneceria suporte à análise almejada do comportamento subjacente à escolha destes, decidiu-se pela exclusão do modo carona do conjunto de alternativas de modo disponíveis a serem modeladas, não estando mais, desta forma, a alternativa carona sob o foco deste estudo. Assim, os experimentos que sucedem esta decisão consideram a situação de escolha para três modos de viagem e seus resultados encontram-se apresentados na seção seguinte.

6.2.2 Resultados conclusivos

O primeiro experimento conduzido para três modos de viagem utilizou todas as variáveis, similarmente ao quinto experimento para quatro modos acima descrito, porém com exceção da variável caronista, não mais pertinente aos modos agora considerados: carro, a pé (ou caminhada) e ônibus (vide Tabela D.9). Os resultados deste experimento estão apresentados na Figura 6.15.

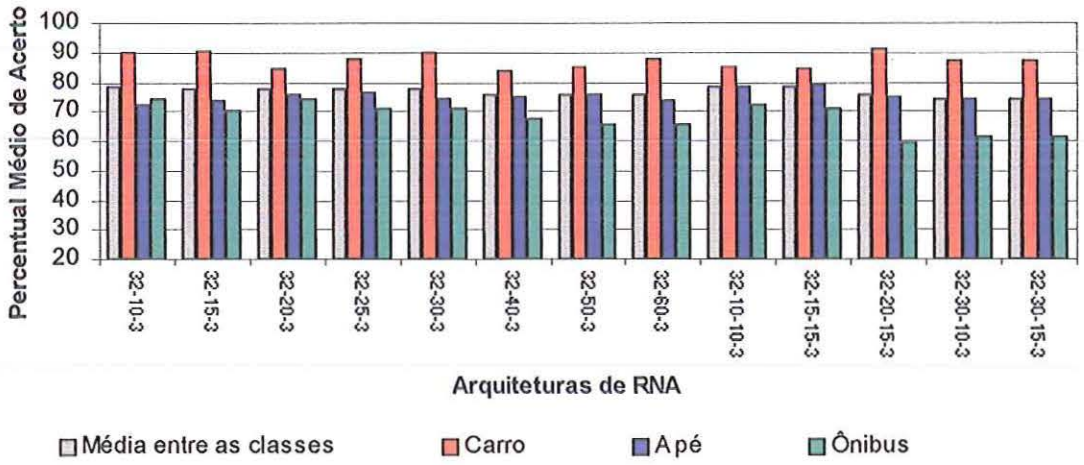


Figura 6.15: Gráfico dos resultados do experimento 1 para 3 modos de viagem.

O desempenho médio entre as 13 arquiteturas testadas neste experimento foi de 76,6% de acerto na classificação e desvio padrão entre classes de 9,2%. A arquitetura com maior percentual de acerto e menor desvio padrão, 78,4% e 6,0%, foi a 32-10-10-3 com resultados mostrados no gráfico da figura acima. Para esta arquitetura foi gerada uma árvore de decisão que atingiu a profundidade ótima ao nível 30, mostrada na Figura A.1 do anexo A, com percentuais de acerto na classificação médio entre classes de 65,4%, de proximidade, ou fidelidade, à rede de 94,2% para o conjunto de validação e com os percentuais para o conjunto de teste de 70,8% e 92,0%, respectivamente. Além do fato de não ser desejável uma árvore com nível de profundidade 30 por apresentar um grande número de regras, há uma incoerência nesta árvore: o nó raiz determina que viajantes que possuem carteira de motorista somente viajam de modo automóvel, carro, como mostra a figura 6.16.

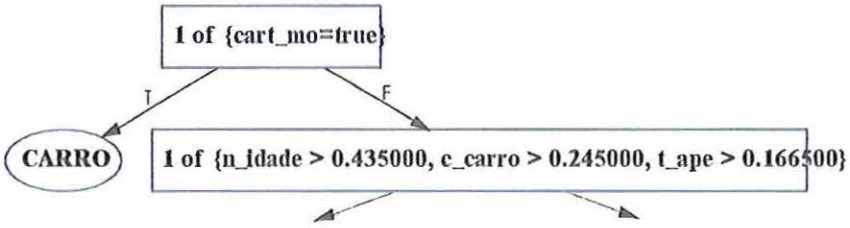


Figura 6.16: Detalhe da raiz da árvore extraída da rede 32-10-10.

Foi ainda observado nesta árvore que dentre as 22 folhas existentes no prolongamento do ramo, no qual não se possui carteira de motorista, há 5 folhas indicando a classe para o modo carro.

Extraiu-se então uma árvore da rede com segunda colocação em termos de desempenho como mostrado na Figura 6.15, a rede de arquitetura 32-15-15-3. Esta rede apresentou percentuais de classificação, médio e em cada classe, próximos da 32-10-10-3. O objetivo de extrair uma árvore desta rede segundo colocada foi observar a forma a qual se daria a estrutura de classificação de eficácia semelhante àquela extraída da 32-10-10-3, ou seja, verificar se haveriam os mesmos problemas encontrados na primeira árvore.

Mostrada por inteiro na Figura A.2, a árvore extraída ao nível ótimo de profundidade 31 obteve percentuais de acerto na classificação de 61,4% e 67,4% e

percentuais de fidelidade à rede de 90,2% e 91,3% para os conjuntos de validação e teste respectivamente. Repetiu-se nesta árvore, no nó raiz, o teste de divisão envolvendo o atributo carteira de motorista e com expansão somente para uma folha da classe carro quando satisfeita a restrição para a posse de carteira de motorista, como mostra a Figura 6.17. Desta vez dentre as 29 folhas de classes existentes no prolongamento do galho para aqueles que não possuem carteira de motorista, 12 são para o modo carro, o que mais uma vez sugere incoerência.

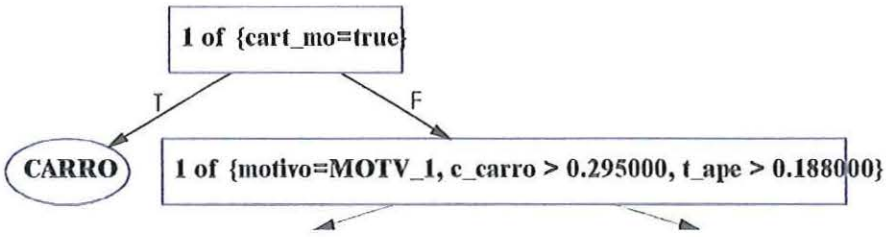


Figura 6.17: Detalhe da raiz da árvore extraída da rede 32-15-15.

Voltou-se aos dados e foram criados novos conjuntos para treinamento de redes sem o atributo carteira de motorista, mantendo-se todos os outros utilizados no experimento anterior (vide Tabela D.10). Um segundo experimento submeteu 17 arquiteturas de rede a processo de treinamento com resultados mostrados nos gráficos das Figuras 6.18 e 6.19.

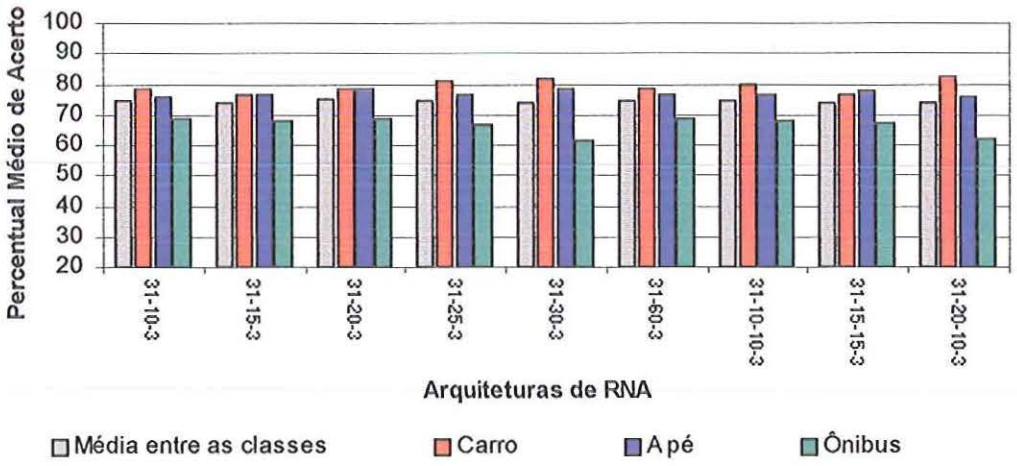


Figura 6.18: Gráfico dos resultados do experimento 2 para 3 modos de viagem.

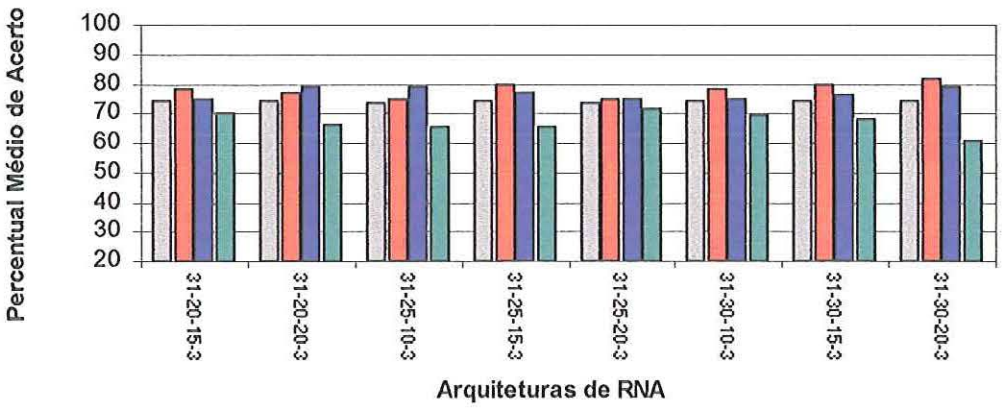


Figura 6.19: Continuação do gráfico da Figura 6.18.

No gráfico da figura acima, entre as 17 arquiteturas testadas o desempenho médio foi de 74,1% de acerto na classificação, com desvio padrão médio entre classes de 6,7%. Analisando-se os gráficos de resultados deste segundo experimento e aquele do experimento anterior, que continha o atributo carteira de motorista, observa-se que para todas as arquiteturas ocorreu uma diminuição do percentual de acerto neste experimento em relação ao anterior. Neste segundo experimento houve empate entre duas arquiteturas, a 31-10-10-3 e a 31-30-15-3, para o maior percentual de acerto que foi de 74,5% com coincidente desvio de padrão de 5,0% para ambas. As Figuras A.3 e A.4 mostram as árvores extraídas nos níveis ótimos de profundidade 48 e 29 para essas redes, com seus respectivos percentuais de acerto e de fidelidade para os conjuntos de validação e teste.

À procura de melhoria da eficácia preditiva da RNA, que registrava um percentual médio de 74,1% de acerto na previsão de escolha de modos de viagem até então, um terceiro experimento foi elaborado utilizando conjuntos de dados que incluíam os atributos do experimento anterior acrescidos dos atributos "trabalha" e "estuda" (vide Tabela D.11). Os gráficos das Figuras 6.20 6.21 apresentam os resultados das 18 arquiteturas investigadas neste experimento.

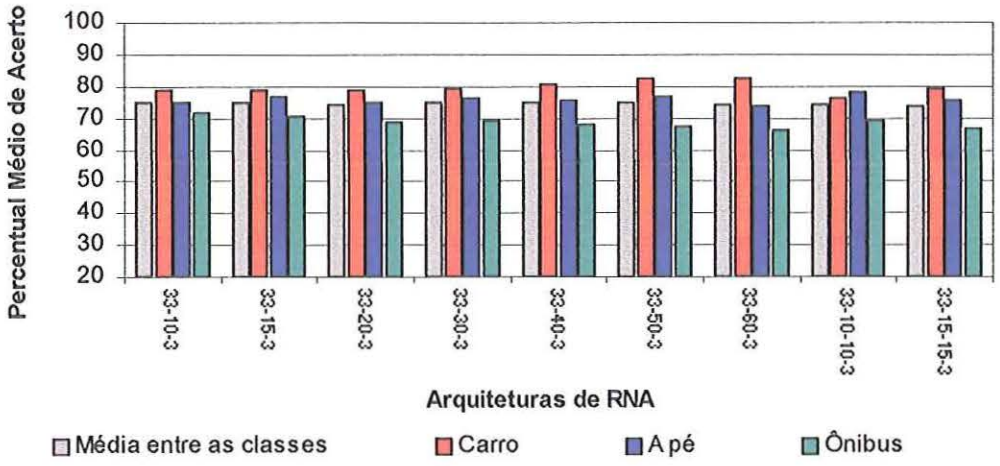


Figura 6.20: Gráfico dos resultados do experimento 3 para 3 modos de viagem.

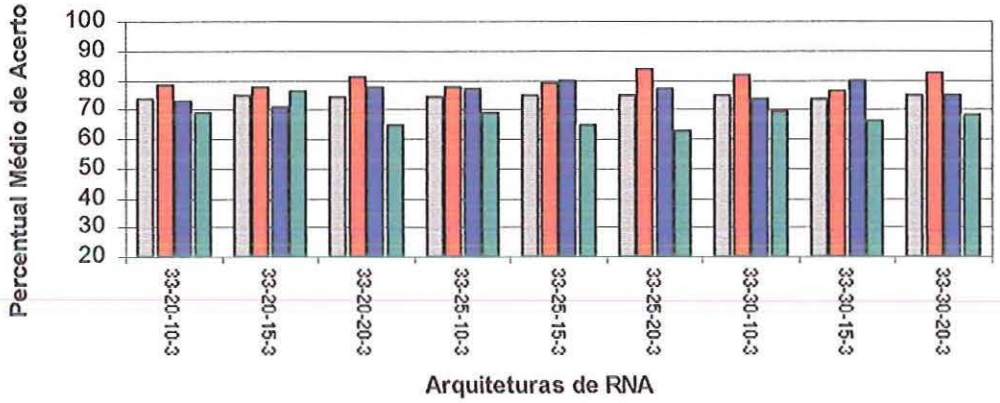


Figura 6.21: Continuação do gráfico da Figura 6.20.

Entre as 18 redes investigadas neste experimento o desempenho médio foi de 74,6% para acerto médio de classificação e 7,1% para desvio padrão entre classes. A arquitetura de melhor desempenho foi a 33-30-20-3 com percentual médio de acerto de 75,2% e desvio entre classes de 6,2% e a arquitetura de menor desvio padrão entre classes foi a 33-20-15-3 com 74,6% e 3,9% para percentuais de acerto médio e desvio. Foram extraídas árvores de decisão para essas duas arquiteturas. Tais árvores alcançaram os níveis de profundidade ótima de 35 e 21, correspondentes às redes 33-30-20-3 e 33-20-15-3 e estão apresentadas nas Figuras A.5 e A.6 com os respectivos índices de desempenho.

A inclusão dos dois atributos, "trabalha" e "estuda", neste terceiro experimento não resultou em um aumento significativo do desempenho preditivo das redes investigadas. Além disso, as árvores extraídas nos três experimentos realizados para três modos de viagem, Figuras A.1 até A.6, possuem muitos conceitos devido aos grandes níveis de profundidade ótima atingidos, performance média de classificação abaixo de 70% e um desequilíbrio na performance de classificação entre as classes. Constata-se esse desequilíbrio, presente na maioria dessas árvores, pela ocorrência de performance de classificação superior a 80% para um modo e inferior a 50% para outro.

A partir dos conjuntos de dados para treinamento utilizados no experimento anterior criou-se novos conjuntos sem os atributos de entrada: "situação familiar" (sit_fam), "grau de instrução" (gr_inst), "tamanho da família" (tam_fam) e "motivo de viagem" (motivo) (vide Tabela D.12). A decisão da remoção desses atributos baseou-se em informações do gráfico de importância relativa entre os atributos de entrada, Figura 6.11, e na observação das árvores extraídas até então que traziam repetições de categorias desses atributos nos testes de divisão de seus nós, trazendo dificuldades de compreensão das mesmas. Os resultados deste quarto experimento para três modos de viagem estão apresentados nas Figuras 6.22e 6.23.

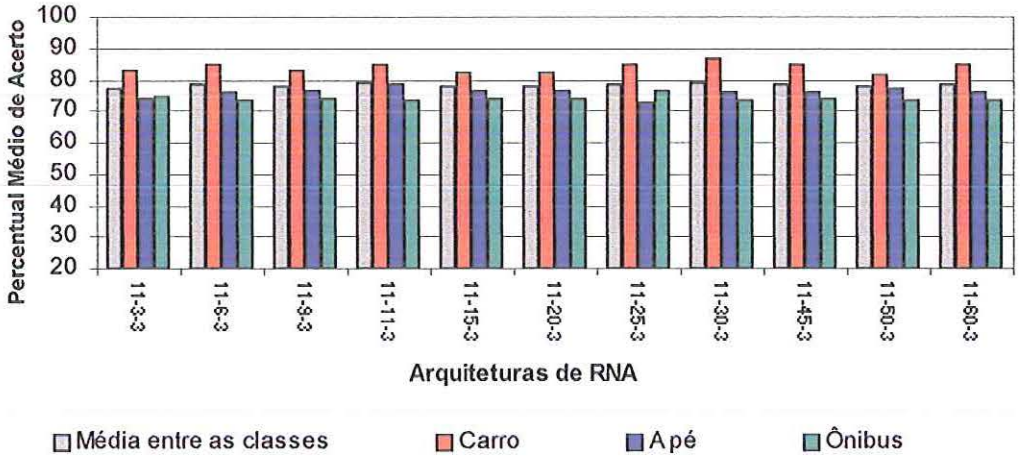


Figura 6.22: Gráfico dos resultados do experimento 4 para 3 modos de viagem.

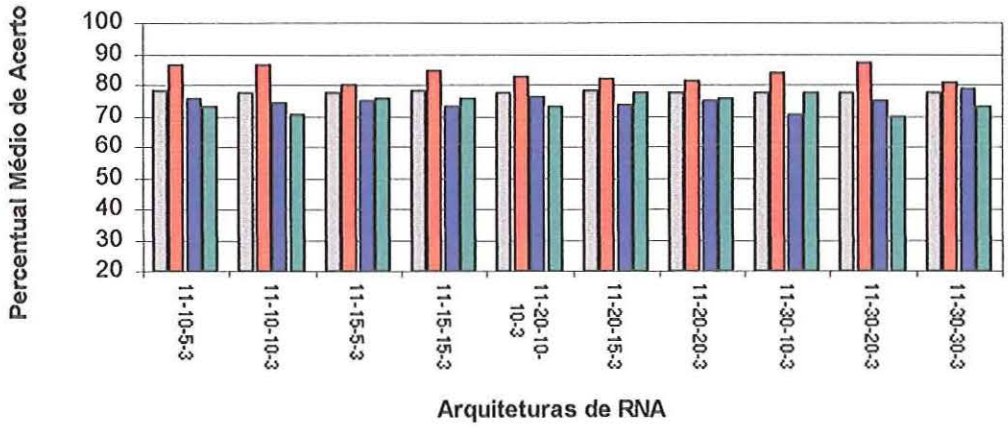


Figura 6.23: Continuação do gráfico da Figura 6.22.

A média de desempenho para as 21 arquiteturas com 11 unidades de entrada avaliadas nesse quarto experimento foi de 77,8% e a média de desvio padrão entre classes foi de 6,9%. Apesar da remoção de alguns atributos de entrada registrou-se um aumento de desempenho médio para todas as redes desse experimento em relação ao desempenho médio do experimento anterior. Foi extraída uma árvore para cada uma das três redes primeiro colocadas em termos de acerto à classificação, 11-11-3 (78,9% e 5,8%[2]), 11-30-3 (78,8% e 14,1%), 11-10-5-3 (78,3% e 6,9%), e uma árvore para a rede de menor desvio padrão, 11-15-15-3 (78,1% e 4,7%).

A árvore da rede vencedora, rede 11-11-3, se adequa ao tipo de resultado buscado, com percentuais de acerto próximos aos da rede originária e equilibrados considerando-se as classes (modos), e também com uma profundidade rasa, neste caso igual a 7. Essa árvore com seus respectivos índices de desempenho encontra-se apresentada na Figuras 7.1 do capítulo seguinte, as demais árvores obtidas neste experimento estão apresentadas nas Figuras B.1 a B.3 do anexo B.

Todavia, ainda que já houvesse um resultado para essa busca por paradigmas, um último experimento foi realizado. Neste experimento, que reuniu apenas 5 atributos de entrada (vide Tabela D.13), aqueles que aparecem nos nós da árvore da rede 11-11-3 com exceção do atributo "custo de viagem por modo

[2] Percentual médio de acerto entre classes e desvio padrão médio entre classes, respectivamente.

automóvel" (c_auto), foi feita uma procura extensa envolvendo 35 diferentes arquiteturas submetidas a processo de treinamento. A exclusão do atributo "c_auto" deve-se ao fato dele ser linearmente dependente do atributo "t_ape", tempo de viagem por modo caminhada, e haver a ocorrência simultânea desses dois atributos num mesmo nó, com valores limiares correspondentes, em várias das árvores extraídas, inclusive no nó raiz da árvore da rede 11-11-3. As Figuras 6.24, 6.25 e 6.26 apresentam os resultados para as redes investigadas neste quinto, e último, experimento para três modos de viagem.

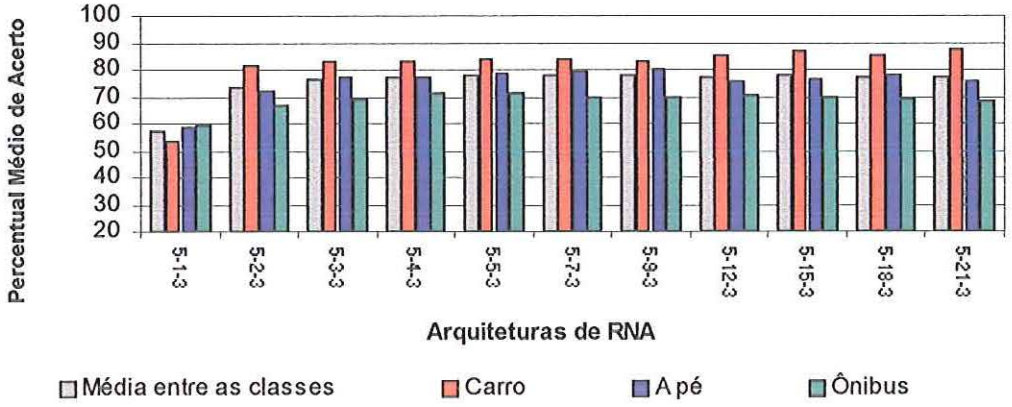


Figura 6.24: Gráfico dos resultados do experimento 5 para 3 modos de viagem.

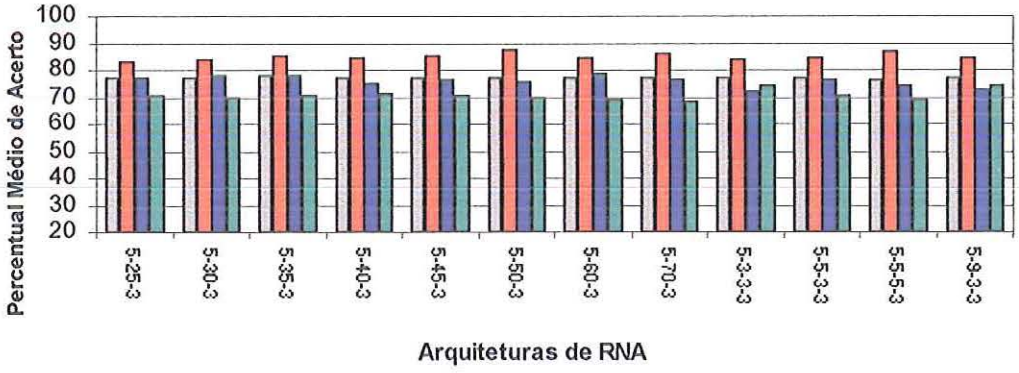


Figura 6.25: Continuação do gráfico da Figura 6.24.

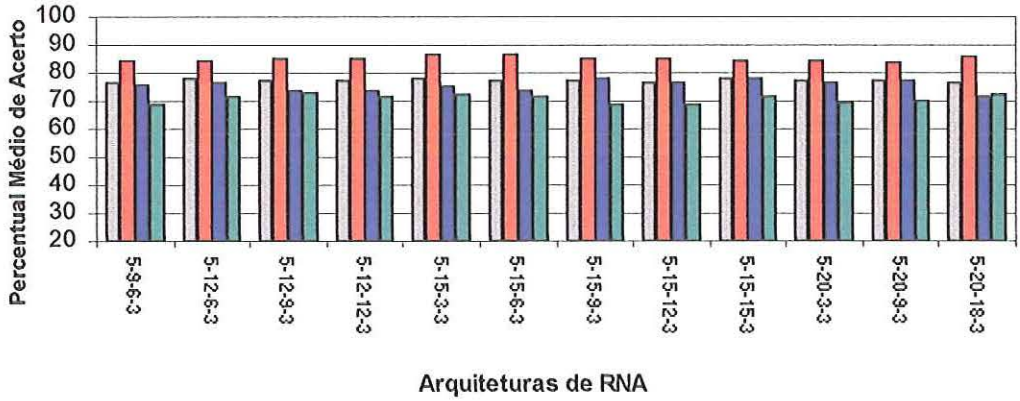


Figura 6.26: Continuação do gráfico da Figura 6.24.

Para as arquiteturas avaliadas neste experimento, a média de acerto de classificação foi de 76,46% com desvio padrão médio entre classes de 8,15%. Foi extraída uma árvore para a rede 5-9-3 a qual obteve o melhor percentual de acerto na classificação (77,8% e 6,23%), uma árvore para a rede 5-3-3-3 de menor desvio padrão (76,7% e 5,1%) e uma última árvore para a rede 5-12-9-3 com o segundo menor desvio padrão observado (77,3% e 6,6%). No capítulo seguinte apresentam-se as Figuras 7.2, 7.3 e 7.4 dessas árvores, que atingiram profundida ótima ao nível 7, conjuntamente com as correspondentes planilhas de condições e de regras.

Pode-se observar nas Figuras 7.2 e 7.3 que as árvores extraídas das redes 5-9-3 e 5-12-9-3 são praticamente iguais, com pequenas diferenças nos valores limiares do atribuo "t_ape" no nó raiz (nível 0) e no nó do nível 3 do ramo à esquerda, e no valor limiar do atributo "n_idade" no nó do nível 1 também do ramo à esquerda. Além da coincidência na profundidade ótima atingida e na forma da árvore, existe muita semelhança nos percentuais de acerto de classificação para os modos de viagem entre essas duas árvores, apresentados nos quadros de desempenho das Figura 7.2 e 7.3. A respeito dos percentuais de acerto de classificação dessas árvores, constata-se que não houve diminuição significativa em relação aos percentuais da árvore da rede 11-11-3 apesar da remoção de alguns atributos de entrada nas redes. A árvore da rede 5-3-3-3 (Figura 7.4) apresenta em seu quadro de desempenho uma distribuição de percentuais de acerto um pouco diferente das distribuições das outras duas árvores para 5 atributos de entrada. Nesta árvore, 5-3-3-3, a distribuição dos percentuais de acerto

assemelha-se mais àquela da árvore da rede 11-11-3, na qual o modo a pé apresenta-se com melhores percentuais na classificação que o modo ônibus

Para consolidação do conhecimento obtido nesta etapa de mineração preparou-se um conjunto de teste, para verificação da eficácia ou da capacidade preditiva das árvores, contendo todas as instâncias da base de dados para 3 modos de viagem, totalizando 6.434 instâncias. As árvores submetidas a esse teste com 6.434 instâncias foram a 11-11-3, a 5-9-3, a 5-12-9-3 e a 5-3-3-3, escolhidas pela simplicidade da estrutura, performance de classificação e por não requerem atributos de entrada os quais não se tem dados coletados além do conjunto amostral, como por exemplo os atributos referentes a viagens por modo ônibus. Os resultados desse teste estão apresentados no quadro de desempenho que acompanha cada árvore, na coluna chamada "Test. 6434", e de maneira geral assemelham-se aos resultados obtidos para os conjuntos de treinamento, validação e teste - partições da amostra. Para os resultados de fidelidade, apenas a árvore da rede 11-11-3 apresentou queda de percentual. Para os resultados de acerto de classificação, as redes 5-9-3 e 5-12-9-3 apresentaram queda de performance para o modo a pé nesse conjunto com 6.434 instâncias em relação às performances obtidas com o conjunto de partições da amostra.

6.3 CONSIDERAÇÕES FINAIS

A mineração para obtenção das quatro árvores de decisão , 11-11-3, 5-9-3, 5-12-9-3 e 5-3-3-3, como estruturas de conhecimento representativas e pequenas dos dados da amostra, envolveu 218 arquiteturas diferentes de redes tipo MLP que submetidas 18 vezes a processo de treinamento (taxas de aprendizagem X conjuntos de treinamento) geraram 3.924 simulações, além de outras mais realizadas através da interface gráfica do simulador, como por exemplo aquelas para definição das taxas de aprendizagem a serem usadas nas repetições do processo de treinamento. O resultado de cada uma dessas simulações foi analisado para a confecção da apresentação dos resultados contida neste capítulo.

7 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Este capítulo primeiramente apresenta as árvores de decisão resultantes e suas correspondentes planilhas contendo as condições originalmente encontradas nas árvores e outras contendo simplificações e regras. Em seguida apresenta uma discussão a respeito da análise e interpretação destes resultados a qual inclui os critérios definidos no sub-ítem 4.7.1 "Análise e Interpretação".

7.1 APRESENTAÇÃO DOS RESULTADOS

As árvores consideradas como resultados alcançados no processo de mineração, ou seja, as árvores das redes 11-11-3, 5-9-3, 5-3-3-3 e 5-12-9-3, encontram-se apresentadas nas Figuras 7.1 a 7.4. Essas figuras incluem os quadros de desempenho com os índices de fidelidade e acurácia preditiva, ou teste, de classificação para os determinados conjuntos utilizados.

Para essas árvores foi feita a desnormalização dos valores limiares dos atributos, presentes nos testes de divisão dos nós, e foram confeccionadas planilhas contendo exatamente as condições necessárias, ou caminho de restrições, correspondentes a cada ramo destas, as quais estão apresentadas nas páginas seguintes a cada árvore nas Planilhas 7.1, 7.4, 7.7 e 7.9. Para a leitura de uma planilha deve-se acompanhar a linha, iniciando pela esquerda (nó raiz) e interpretando cada pontuação de dois pontos (:) como a adição entre restrições, similarmente ao operador lógico aditivo ($\wedge$). Nas planilhas de condições contidas nas árvores sempre que houve a ocorrência do atributo "tempo de viagem por modo caminhada" (t_ape) ou "custo de viagem por modo automóvel" (c_carro) no teste de

QUADRO DE DESEMPENHO DA ÁRVORE					
Profundidade 7		Trein.	Valid.	Test.	Test. 6434
% Fidelidade à MLP-11-11-3		86,8	87,7	87,2	76,1
% Acerto à Classific.	Médio entre modos	72,5	75,8	74,7	73,5
	Carro	72,7	68,8	80,2	72,2
	A pé	82,4	85,1	72,9	80,5
	Ônibus	61,3	74,4	70,6	68,3

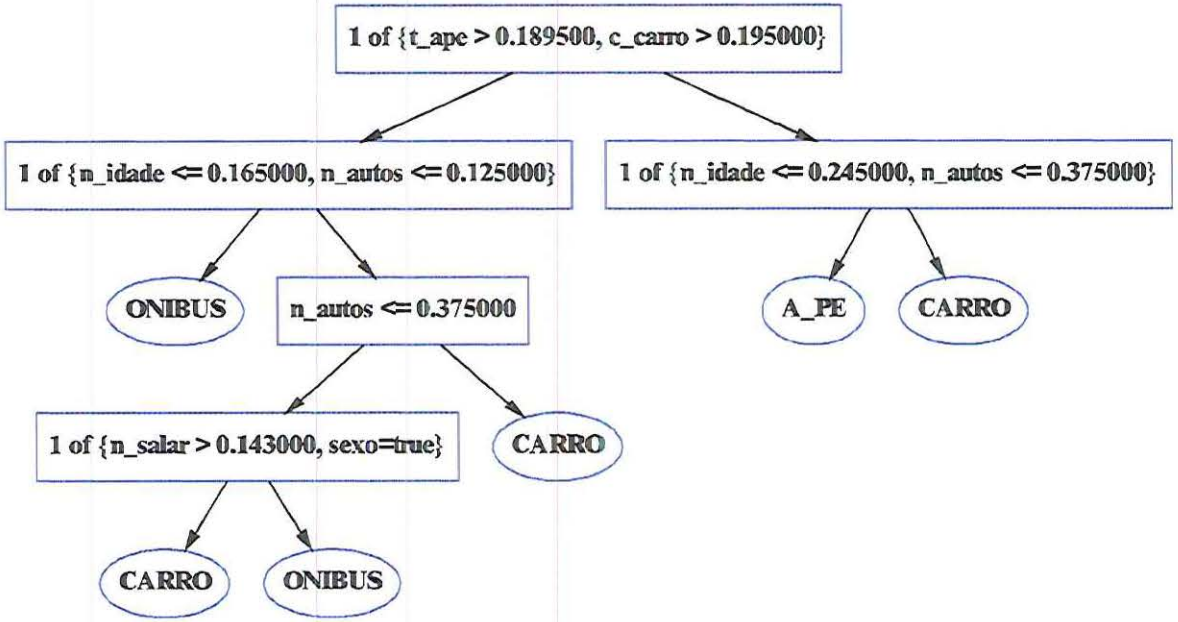


Figura 7.1: Árvore extraída da rede 11-11-3 com profundidade 7.

Planilha 7.1: Condições para os modos de viagem da árvore extraída da rede 11-11-3 com profundidade 7.

onibus	t_ape>33min ou dist>2,5km ou (c_carro>R$0,29 ou dis>=2,0km) :	idade<=17 ou n_autos=0 : ONIBUS
carro	t_ape>33min ou dist>2,5km ou (c_carro>R$0,29 ou dist>=2,0km) :	idade>17 ou n_autos>=1 : n_autos<=1: n_salar>14sal-mín ou homem : CARRO
onibus	t_ape>33min ou dist>2,5km ou (c_carro>R$0,29 ou dist>=2,0km):	idade>17 ou n_autos>=1 : n_autos<=1 : n_salar<=14sal-mín ou mulher : ONIBUS
carro	t_ape>33min ou dist>2,5km ou (c_carro>R$0,29 ou dist>=2,0km) :	idade>17 ou n_autos>=1 : n_autos>1 : CARRO
a_pe	t_ape<=33min ou dist<=2,5km ou (c_carro<R$0,29 ou dist<2,0km) :	idade<=24 ou n_autos<=1 : A_PE
carro	t_ape<=33min ou dist<=2,5km ou (c_carro<R$0,29 ou dist<2,0km) :	idade>=24 ou n_autos>=1 : CARRO

Planilha 7.2: Condições simplificadas para os modos de viagem da árvore extraída da rede 11-11-3 com profundidade 7.

onibus	dist>2,5km ou dist>=2,0km :	idade<=17 ou n_autos=0 : ONIBUS			
carro	dist>2,5km ou dist>=2,0km :	idade>17 ou n_autos>=1 :	n_autos<=1:	n_salar>14sal-mín ou homem : CARRO	
onibus	dist>2,5km ou dist>=2,0km :	idade>17 ou n_autos>=1 :	n_autos<=1 :	n_salar<=14sal-mín ou mulher : ONIBUS	
carro	dist>2,5km ou dist>=2,0km :	idade>17 ou n_autos>=1 :	n_autos>1 : CARRO		
a_pe	dist<=2,5km ou dist<2,0km:	idade<=24 ou n_autos<=1 : A_PE			
carro	dist<=2,5km ou dist<2,0km:	idade>24 ou n_autos>1 : CARRO			

Planilha 7.3: Regras para os modos de viagem da árvore extraída da rede 11-11-3 com profundidade 7. Com número de instâncias adicionais corretamente classificadas e correspondente percentual de acerto de classificação por regra, para um conjunto com 567 instânciais (conj. de treinamento).

regra1	dist>2,5km:	idade<=17: ONIBUS				20 --> 10,8%	
regra2	dist>2,5km:	n_autos=0: ONIBUS				68 --> 36,6%	
regra3	dist>=2,0km:	idade<=17: ONIBUS				2 --> 1,1%	
regra4	dist>=2,0km:	n_autos=0: ONIBUS				10 --> 5,4%	
regra5	dist>2,5km:	idade>17:	n_autos<=1:	n_salar>14sal-mín: CARRO		10 --> 6,1%	
regra6	dist>2,5km:	idade>17:	n_autos<=1:	homem: CARRO		31 --> 18,8%	
regra7	dist>2,5km:	n_autos=1:	n_salar>14sal-mín: CARRO			0 (coincidente à rgr. 5)	
regra8	dist>2,5km:	n_autos=1:	homem: CARRO			0 (coincidente à rgr. 6)	
regra9	dist>=2,0km:	idade>17:	n_autos<=1:	n_salar>14sal-mín: CARRO		2 --> 1,2%	
regra10	dist>=2,0km:	idade>17:	n_autos<=1:	homem: CARRO		0 (contida nas rgrs. 5, 6 e 9)	
regra11	dist>=2,0km:	n_autos=1:	n_salar>14sal-mín: CARRO			0 (coincidente à rgr. 9)	
regra12	dist>=2,0km:	n_autos=1:	homem: CARRO			0 (coincidente à rgr.10)	
regra13	dist>2,5km:	idade>17:	n_autos<=1:	n_salar<=14sal-mín: ONIBUS		12 --> 6,5%	
regra14	dist>2,5km:	idade>17:	n_autos<=1:	mulher : ONIBUS		0 (coincidente à rgr.13)	
regra15	dist>2,5km:	n_autos=1:	n_salar<=14sal-mín: ONIBUS			0 (coincidente à rgr.13)	
regra16	dist>2,5km:	n_autos=1:	mulher : ONIBUS			0 (coincidente à rgr.13)	
regra17	dist>=2,0km:	idade>17:	n_autos<=1:	n_salar<=14sal-mín: ONIBUS		2 --> 1,1%	
regra18	dist>=2,0km:	idade>17:	n_autos<=1:	mulher : ONIBUS		0 (coincidente à rgr.17)	
regra19	dist>=2,0km:	n_autos=1:	n_salar<=14sal-mín: ONIBUS			0 (coincidente à rgr.17)	
regra20	dist>=2,0km:	n_autos=1:	mulher : ONIBUS			0 (coincidente à rgr.17)	/ total ONIBUS 114 --> **61,3%**
regra21	dist>2,5km:	idade>17:	n_autos>1: CARRO			38 --> 23,0%	
regra22	dist>2,5km:	n_autos>1: CARRO				0 (coincidente à rgr.21)	
regra23	dist>=2,0km:	idade>17:	n_autos>1: CARRO			8 --> 4,8%	
regra24	dist>=2,0km:	n_autos>1: CARRO				0 (coincidente à rgr.23)	
regra25	dist<=2,5km:	idade<=24: A_PE				105 --> 50,0%	
regra26	dist<=2,5km:	n_autos<=1: A_PE				68 --> 32,4%	
regra27	dist<2,0km:	idade<=24: A_PE				0 (coincidente à rgr.25)	
regra28	dist<2,0km:	n_autos<=1: A_PE				0 (coincidente à rgr.26)	/ total A_PE 173 --> **82,4%**
regra29	dist<=2,5km:	idade>24: CARRO				31 --> 18,8%	
regra30	dist<=2,5km:	n_autos>1: CARRO				0 (coincidente à rgr.29)	
regra31	dist<2,0km:	idade>24: CARRO				0 (coincidente à rgr.29)	
regra32	dist<2,0km:	n_autos>1: CARRO				0 (coincidente à rgr.29)	/ total CARRO 120 --> **72,7%**

QUADRO DE DESEMPENHO DA ÁRVORE					
Profundidade 7		Trein.	Valid.	Test.	Test. 6434
% Fidelidade à MLP-5-9-3		92.1	89.8	91.5	90.1
% Acerto à Classific.	Médio entre modos	75.5	75.3	72.2	74.6
	Carro	78.2	82.4	80.6	80.9
	A pé	77.9	72.4	74.4	66.3
	Ônibus	70.7	71.1	61.3	75.8

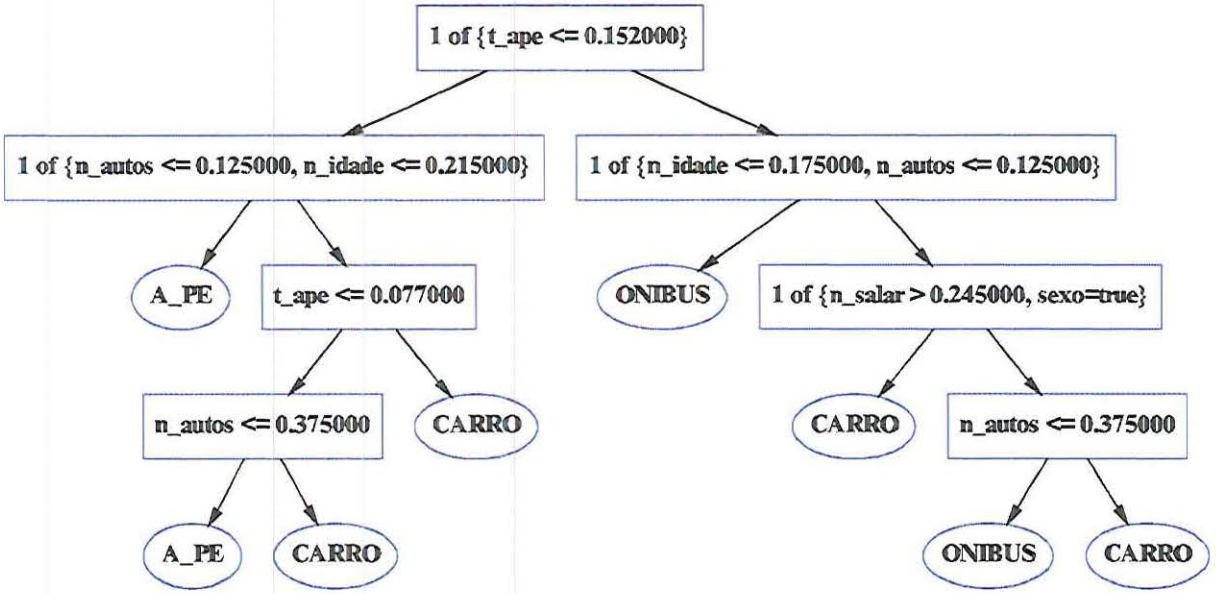


Figura 7.2: Árvore extraída da rede 5-9-3 com profundidade 7.

Planilha 7.4: Condições para os modos de viagem da árvore extraída da rede 5-9-3 com profundidade 7.

a_pe	t_ape<=26min ou dist<=2,0km:	n_autos=0 ou n_idade<=22: A_PE		
a_pe	t_ape<=26min ou dist<=2,0km:	n_autos>=1 ou n_idade>22:	t_ape<15min ou dist<1,1km:	n_autos<=1: A_PE
carro	t_ape<=26min ou dist<=2,0km:	n_autos>=1 ou n_idade>22:	t_ape<15min ou dist<1,1km:	n_autos>1: CARRO
carro	t_ape<=26min ou dist<=2,0km:	n_autos>=1 ou n_idade>22:	t_ape>=14min ou dist>=1,1km: CARRO	
carro	t_ape>26min ou dist>2,0km:	n_idade>18 ou n_autos>=1:	n_salar>22sal-mín ou homem: CARRO	
carro	t_ape>26min ou dist>2,0km:	n_idade>18 ou n_autos>=1:	n_salar<= 22sal-mín ou mulher:	n_autos > 1: CARRO
onibus	t_ape>26min ou dist>2,0km:	n_idade<=18 ou n_autos=0: ONIBUS		
onibus	t_ape>26min ou dist>2,0km:	n_idade>18 ou n_autos>=1:	n_salar<= 22sal-mín ou mulher:	n_autos<=1: ONIBUS

Planilha 7.5: Condições simplificadas para os modos de viagem da árvore extraída da rede 5-9-3 com profundidade 7.

a_pe	dist<=2,0km:	n_autos=0 ou n_idade<=22:	A_PE
a_pe	dist<1,1km:	n_autos>=1 ou n_idade>22:	n_autos<=1: A_PE
carro	dist<1,1km:	n_autos>=1 ou n_idade>22:	n_autos>1: CARRO
carro	1,1<=dist<=2,0km:	n_autos>=1 ou n_idade>22:	CARRO
onibus	dist>2,0km:	n_idade<=18 ou n_autos=0:	ONIBUS
carro	dist>2,0km:	n_idade>18 ou n_autos>=1:	n_salar>22sal-mín ou homem: CARRO
onibus	dist>2,0km:	n_idade>18 ou n_autos>=1:	n_salar<= 22sal-mín ou mulher: n_autos<=1: ONIBUS
carro	dist>2,0km:	n_idade>18 ou n_autos>=1:	n_salar<= 22sal-mín ou mulher: n_autos > 1: CARRO

Planilha 7.6: Regras para os modos de viagem da árvore extraída da rede 5-9-3 com profundidade 7. Com número de instâncias adicionais corretamente classificadas e correspondente percentual de acerto de classificação por regra, para um conjunto com 567 instânciais (conj. de treinamento).

regra 1	dist<=2,0km:	n_autos=0: A_PE			76 --> 40,0%	
regra 2	dist<=2,0km:	idade<=22: A_PE			46 --> 24,2%	
regra 3	dist<1,1km:	n_autos=1: A_PE			26 --> 13,7%	
regra 4	dist<1,1km:	n_idade>22:	n_autos<=1: A_PE		0 (contida nas regras 1 e 3)	/ total A_PE 148 --> **77,9%**
regra 5	dist<1,1km:	n_autos>1: CARRO			12 --> 6,7%	
regra 6	dist<1,1km:	n_idade>22:	n_autos>1: CARRO		0 (coincidente à rgr. 5)	
regra 7	1,1<=dist<=2,0km:	n_autos>=1: CARRO			35 --> 19,6%	
regra 8	1,1<=dist<=2,0km:	n_idade>22: CARRO			0 (coincidente à rgr. 7)	
regra 9	dist>2,0km:	n_idade<=18: ONIBUS			32 --> 16,2%	
regra 10	dist>2,0km:	n_autos=0: ONIBUS			81 --> 40,9%	
regra 11	dist>2,0km:	n_idade>18:	n_salar>22sal-mín: CARRO		36 --> 20,1%	
regra 12	dist>2,0km:	n_idade>18:	homem: CARRO		46 --> 25,7%	
regra 13	dist>2,0km:	n_autos>=1:	n_salar>22sal-mín: CARRO		0 (coincidente à rgr. 11)	
regra 14	dist>2,0km:	n_autos>=1:	homem: CARRO		0 (coincidente à rgr. 12)	
regra 15	dist>2,0km:	n_idade>18:	n_salar<=22sal-mín:	n_autos<=1: ONIBUS	27 --> 13,6%	
regra 16	dist>2,0km:	n_idade>18:	mulher:	n_autos<=1: ONIBUS	0 (coincidente à rgr. 15)	
regra 17	dist>2,0km:	n_autos=1:	n_salar<= 22sal-mín: ONIBUS		0 (coincidente à rgr. 15)	
regra 18	dist>2,0km:	n_autos=1:	mulher: ONIBUS		0 (coincidente à rgr. 15)	/ total ONIBUS 140 --> **70,7%**
regra 19	dist>2,0km:	n_idade>18:	n_salar<=22sal-mín:	n_autos > 1: CARRO	11 --> 6,1%	
regra 20	dist>2,0km:	n_idade>18:	mulher:	n_autos > 1: CARRO	0 (coincidente à rgr. 19)	
regra 21	dist>2,0km:	n_autos>1:	n_salar<=22sal-mín: CARRO		0 (coincidente à rgr. 19)	
regra 22	dist>2,0km:	n_autos>1:	mulher: CARRO		0 (coincidente à rgr. 19)	/ total CARRO 140 --> **78,2%**

QUADRO DE DESEMPENHO DA ÁRVORE					
Profundidade 7		Trein.	Valid.	Test.	Test. 6434
% Fidelidade à MLP-5-12-9-3		91,7	91,0	89,0	90,4
% Acerto à Classific.	Médio entre modos	75,8	74,5	73,3	73,1
	Carro	78,8	82,4	81,6	80,6
	A pé	76,3	67,8	74,4	60,2
	Ônibus	72,4	73,5	63,4	77,6

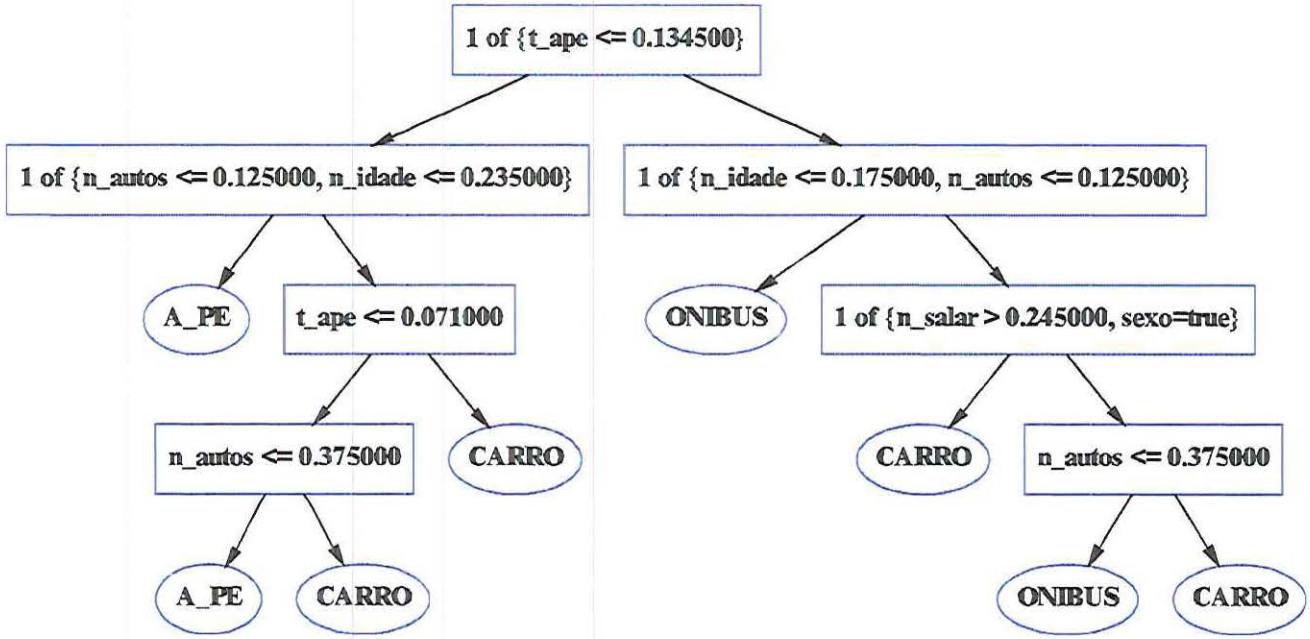


Figura 7.3: Árvore extraída da rede 5-12-9-3 com profundidade 7

Planilha 7.7: Condições para os modos de viagem da árvore extraída da rede 5-12-9-3 com profundidade 7.

a_pe	t_ape<=24min ou dist<1,8km:	n_autos=0 ou n_idade<=23: A_PE		
a_pe	t_ape<=24min ou dist<1,8km:	n_autos>=1 ou n_idade>23:	t_ape<=13min ou dist<=1,0km:	n_autos<=1: A_PE
carro	t_ape<=24min ou dist<1,8km:	n_autos>=1 ou n_idade>23:	t_ape<=13min ou dist<=1,0km:	n_autos>1: CARRO
carro	t_ape<=24min ou dist<1,8km:	n_autos>=1 ou n_idade>23:	t_ape>13min ou dist>1,0km: CARRO	
carro	t_ape>24min ou dist>=1,8km:	n_idade>18 ou n_autos>=1:	n_salar>22sal-mín ou homem: CARRO	
carro	t_ape>24min ou dist>=1,8km:	n_idade>18 ou n_autos>=1:	n_salar<=22sal-mín ou mulher:	n_autos>1: CARRO
onibus	t_ape>24min ou dist>=1,8km:	n_idade<=18 ou n_autos=0: ONIBUS		
onibus	t_ape>24min ou dist>=1,8km:	n_idade>18 ou n_autos>=1:	n_salar<=22sal-mín ou mulher:	n_autos<=1: ONIBUS

Planilha 7.8: Condições simplificadas para os modos de viagem da árvore extraída da rede 5-12-9-3 com profundidade 7.

a_pe	dist<1,8km:	n_autos=0 ou n_idade<=23: A_PE	
a_pe	dist<=1,0km:	n_autos>=1 ou n_idade>23:	n_autos<=1: A_PE
carro	dist<=1,0km:	n_autos>=1 ou n_idade>23:	n_autos>1: CARRO
carro	1,0<dist<1,8km:	n_autos>=1 ou n_idade>23: CARRO	
onibus	dist>=1,8km:	n_idade<=18 ou n_autos=0: ONIBUS	
carro	dist>=1,8km:	n_idade>18 ou n_autos>=1:	n_salar>22sal-mín ou homem: CARRO
onibus	dist>=1,8km:	n_idade>18 ou n_autos>=1:	n_salar<=22sal-mín ou mulher: n_autos<=1: ONIBUS
carro	dist>=1,8km:	n_idade>18 ou n_autos>=1:	n_salar<=22sal-mín ou mulher: n_autos>1: CARRO

QUADRO DE DESEMPENHO DA ÁRVORE					
Profundidade 7		Trein.	Valid.	Test.	Test. 6434
% Fidelidade à MLP-5-3-3-3		90,8	88,6	90,7	88,8
% Acerto à Classific.	Médio entre modos	76,4	76,9	75,4	74,5
	Carro	76,5	84,7	84,7	82,2
	A pé	86,3	77,0	81,1	73,9
	Ônibus	66,6	68,7	60,2	67,4

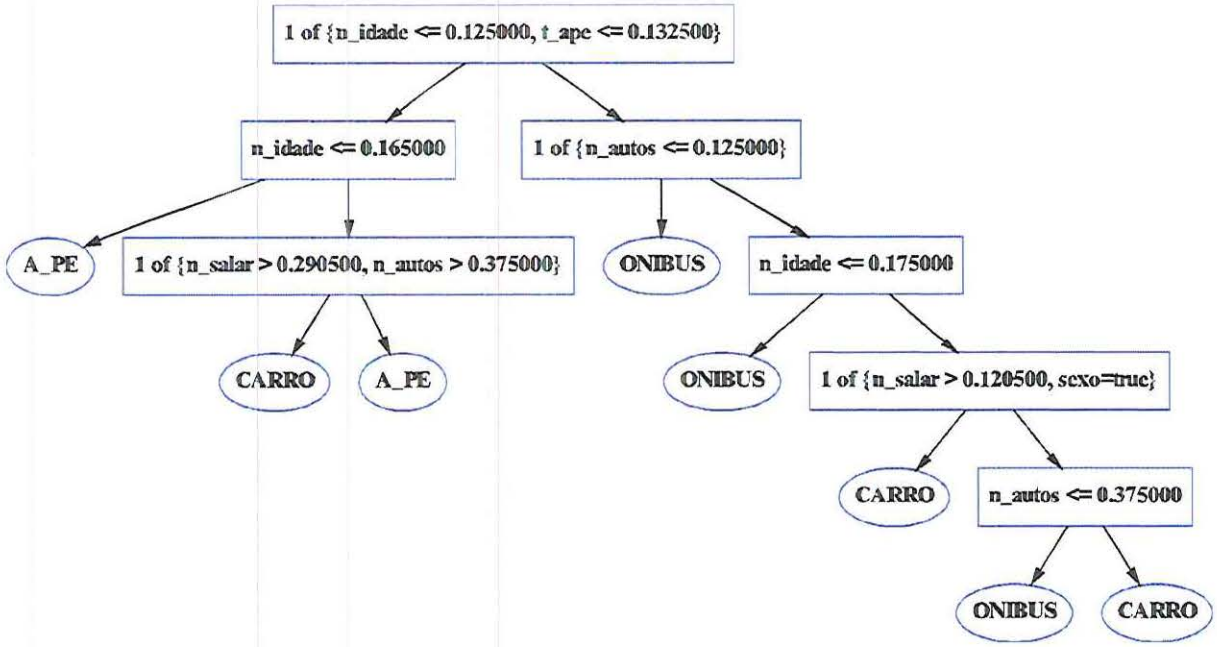


Figura 7.4: Árvore extraída da rede 5-3-3-3 com profundidade 7.

Planilha 7.9: Condições para os modos de viagem da árvore extraída da rede 5-3-3-3 com profundidade 7.

a_pe	n_idade<=14 ou t_ape<=23 ou dist<=1,7km:	n_idade<=17: A_PE	
a_pe	n_idade<=14 ou t_ape<=23 ou dist<=1,7km:	n_idade>17:	n_salar<=26sal-mín ou n_autos<=1: A_PE
carro	n_idade<=14 ou t_ape<=23 ou dist<=1,7km:	n_idade>17:	n_salar>26sal-mín ou n_autos>1: CARRO
carro	n_idade>14 ou t_ape>23 ou dist>1,7km:	n_autos>=1:	n_idade>18: n_salar>11sal-mín ou homen: CARRO
carro	n_idade>14 ou t_ape>23 ou dist>1,7km:	n_autos>=1:	n_idade>18: n_salar<=11sal-mín ou mulher: n_autos>1: CARRO
onibus	n_idade>14 ou t_ape>23 ou dist>1,7km:	n_autos=0: ONIBUS	
onibus	n_idade>14 ou t_ape>23 ou dist>1,7km:	n_autos>=1:	n_idade<=18: ONIBUS
onibus	n_idade>14 ou t_ape>23 ou dist>1,7km:	n_autos>=1:	n_idade>18: n_salar<=11sal-mín ou mulher: n_autos<=1: ONIBUS

Planilha 7.10: Condições simplificadas para os modos de viagem da árvore extraída da rede 5-3-3-3 com profundidade 7.

a_pe	n_idade<=14 ou dist<=1,7km:	n_idade<=17: A_PE	
carro	n_idade<=14 ou dist<=1,7km:	n_idade>17:	n_salar>26sal-mín ou n_autos>1: CARRO
a_pe	n_idade<=14 ou dist<=1,7km:	n_idade>17:	n_salar<=26sal-mín ou n_autos<=1: A_PE
onibus	n_idade>14 ou dist>1,7km:	n_autos=0: ONIBUS	
onibus	n_idade>14 ou dist>1,7km:	n_autos>=1:	n_idade<=18: ONIBUS
carro	n_idade>14 ou dist>1,7km:	n_autos>=1:	n_idade>18: n_salar>11sal-mín ou homen: CARRO
onibus	n_idade>14 ou dist>1,7km:	n_autos>=1:	n_idade>18: n_salar<=11sal-mín ou mulher: n_autos<=1: ONIBUS
carro	n_idade>14 ou dist>1,7km:	n_autos>=1:	n_idade>18: n_salar<=11sal-mín ou mulher: n_autos>1: CARRO

Planilha 7.11: Regras para os modos de viagem da árvore extraída da rede 5-3-3-3 com profundidade 7. Com número de instâncias adicionais corretamente classificadas e correspondente percentual de acerto de classificação por regra, para um conjunto com 567 instânciais (conj. de treinamento).

regra1	n_idade<=14: A_PE				66 --> 34,7%	
regra2	dist<=1,7km:	n_idade<=17: A_PE			23 --> 12,1%	
regra3	dist<=1,7km:	n_idade>17:	n_salar>26sal-mín ou n_autos>1: CARRO		16 --> 8,9%	
regra4	dist<=1,7km:	n_idade>17:	n_autos>1: CARRO		8 --> 4,5%	
regra5	dist<=1,7km:	n_idade>17:	n_salar<=26sal-mín: A_PE		75 --> 39,5%	
regra6	dist<=1,7km:	n_idade>17:	n_autos<=1: A_PE		0 (coincidente à rgr. 5)	/ total A_PE 164 --> **86,3%**
regra7	n_idade>14:	n_autos=0: ONIBUS			90 --> 45,5%	
regra8	dist>1,7km:	n_autos=0: ONIBUS			0 (coincidente à rgr. 7)	
regra9	n_idade>14:	n_autos>=1:	n_idade<=18: ONIBUS		20 --> 10,1%	
regra10	dist>1,7km:	n_autos>=1:	n_idade<=18: ONIBUS		0 (coincidente à rgr. 9)	
regra11	n_idade>18:	n_autos>=1:	n_salar>11sal-mín: CARRO		76 --> 42,5%	
regra12	n_idade>18:	n_autos>=1:	homen: CARRO		30 --> 16,8%	
regra13	dist>1,7km:	n_autos>=1:	n_idade>18:	n_salar>11sal-mín: CARRO	0 (coincidente à rgr.11)	
regra14	dist>1,7km:	n_autos>=1:	n_idade>18:	homen: CARRO	0 (coincidente à rgr.12)	
regra15	n_idade>18:	n_autos=1:	n_salar<=11sal-mín: ONIBUS		22 --> 11,1%	
regra16	n_idade>18:	n_autos=1:	mulher: ONIBUS		0 (coincidente à rgr. 15)	
regra17	dist>1,7km:	n_autos=1:	n_salar<=11sal-mín: ONIBUS		0 (coincidente à rgr. 15)	
regra18	dist>1,7km:	n_autos=1:	mulher: ONIBUS		0 (coincidente à rgr. 15)	/ total ONIBUS 132 --> **66,6%**
regra19	n_idade>18:	n_salar<=11sal-mín:	n_autos>1: CARRO		7 --> 3,9%	
regra20	n_idade>18:	mulher:	n_autos>1: CARRO		0 (coincidente à rgr. 19)	
regra21	dist>1,7km:	n_idade>18:	n_salar<=11sal-mín:	n_autos>1: CARRO	0 (coincidente à rgr. 19)	
regra22	dist>1,7km:	n_idade>18:	mulher:	n_autos>1: CARRO	0 (coincidente à rgr. 19)	/ total CARRO 137 --> **76,5%**

divisão de um nó, foi adicionado o atributo "distância de caminho mínimo" (dist), sob condição lógica alternativa (ou), porque uma vez normalizados esses atributos, não há distinção entre ambos por serem proporcionais, e também porque tal adição poderia vir a facilitar a compreensão de determinado teste de nó. Foram também confeccionadas para cada árvore resultante planilhas contendo simplificações das planilhas de condições. Essas simplificações referem-se à troca dos atributos proporcionais ao atributo "distância de caminho mínimo" acima referidos por este e estão mostradas nas Figuras 7.2, 7.5, 7.8 e 7.10.

A passagem pelo caminho de restrições a partir do nó raiz da árvore de decisão até uma folha de classe está sujeita às condições alternativas, ou disjuntivas, dos teste de divisão dos nós enfileirados ao longo desse caminho, as quais geram mais de uma maneira de se alcançar essa folha de classe destino. Essas diferentes formas de se chegar a uma folha de classe partindo-se do nó raiz numa árvore de decisão constituem as diferentes regras para determinada folha. As Planilhas 7.3, 7.6 e 7.11 contém as regras decorrentes das possíveis combinações entre as restrições por atributos considerando as linhas da planilha de condições de cada árvore, ou seja, considerando os ramos para as folhas de cada árvore. As regras correspondentes a cada árvore foram aplicadas com o auxílio de uma planilha eletrônica a um conjunto, ora já utilizado no treinamento das redes originárias às árvores, para a tarefa de classificação, com 567 instâncias distribuídas em 179, 190 e 198 instâncias para os modos carro, a pé e ônibus respectivamente. Os resultados, número de instâncias adicionais corretamente classificadas e correspondente percentual de acerto referente à quantidade de instâncias daquele modo de viagem existente no conjunto, estão apresentados nessas planilhas de regras à direita na linha da regra. Apenas para a árvore da rede 5-12-9-3 não há planilha de regras uma vez que esta árvore é idêntica àquela da rede 5-9-3, com uma pequena diferença no valor limiar do atributo "t_ape" no nó do nível de profundidade 3 à esquerda da raiz, e com acurácia preditiva similar à mesma, como pode ser visto nos correspondentes quadros de desempenho dessas árvores.

7.2 DISCUSSÃO DOS RESULTADOS

As árvores de decisão possuem uma hierarquia em sua estrutura para a aplicação. Tal hierarquia define a ordem com a qual as regras devem ser aplicadas para a discriminação das diferentes classes as quais pertencem os dados de um conjunto submetido à estrutura da árvore de decisão. Dessa forma, em uma árvore de decisão a estrutura, ou os ramos, ou ainda as regras são desenvolvidas considerando a interferência das regras antecessoras nas seguintes, ou seja, em determinada profundidade o teste de divisão de um nó cria uma regra que possui domínio de aplicação subjugado pelas regras antecessoras, composto pelos dados restantes ou não abrangidos por estas.

Para melhor ilustrar essa estrutura hierárquica para aplicação da árvore de decisão foi escolhida a árvore da rede 5-9-3 (Figura 7.2) como exemplo, por haver outra árvore de estrutura idêntica, a árvore da rede 5-12-9-3 (Figura 7.3) dentre as quatro árvores resultantes. Nessa árvore com simetria bilateral quanto à profundidade, mostrada na Figura 7.2, o primeiro ramo destina-se à folha do modo a pé, localizada na profundidade 2 do lado que satisfaz as restrições contidas no nó raiz, isto é, ramo à esquerda da raiz. Esse ramo corresponde à primeira linha da planilha de condições (Planilha 7.4) ou da planilha de condições simplificadas (Planilha 7.5) a qual origina as duas primeiras regras para o modo a pé dessa árvore, mostradas na Planilha 7.6. Quando aplicada ao conjunto de 567 instâncias, a primeira regra abrange 93 destas sendo que 76 pertencem de fato ao modo a pé e 17 instâncias pertencentes ao modo ônibus são erroneamente abrangidas, como mostra a Figura 7.5.

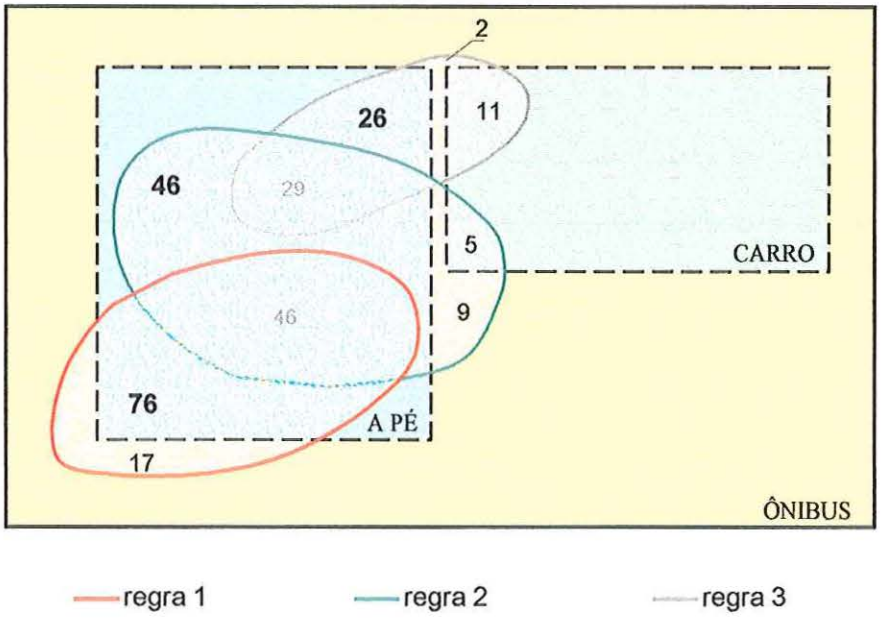
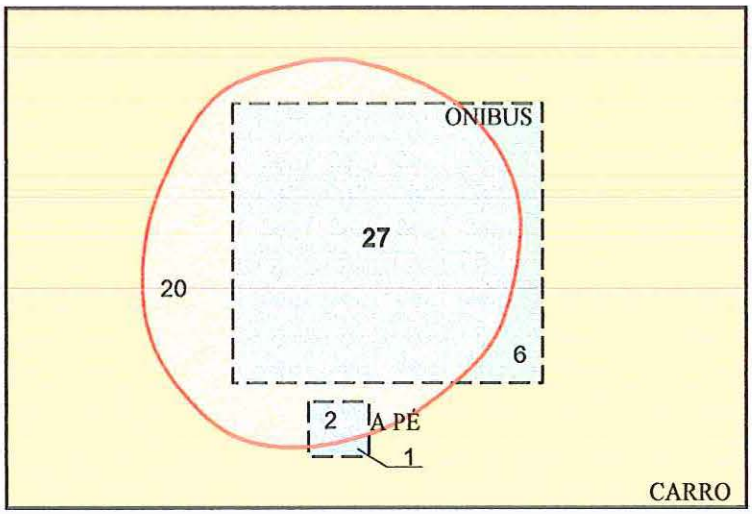


Figura 7.5.: Diagrama das regras da árvore da rede 5-9-3 para o modo a pé, com o número de instâncias compreendidas por local.

A segunda regra desse ramo abrange 106 instâncias do conjunto dentre as quais 46 instâncias são comuns à primeira regra, ou seja, já foram classificadas como pertencentes ao modo a pé pela regra anterior, a primeira. Dentre as 60 novas instâncias abrangidas pela segunda regra 46 realmente pertencem ao modo a pé, 9 instâncias pertencem ao modo ônibus e 5 ao modo carro. Apesar dessas duas regras estarem em um mesmo nível de profundidade, o que não determina uma ordem de aplicação entre elas, deve-se considerar apenas o número de instâncias adicionais corretamente classificadas em cada regra. Havendo a aplicação de uma delas ao conjunto de dados, a seguinte regra não encontraria mais aqueles que viriam a ser comuns à classificação de ambas. Por sua vez, o segundo ramo para modo a pé o qual alcança a folha desse modo de viagem localizada na profundidade 6 à esquerda da raiz da árvore gera outras duas regras, a terceira e a quarta regra da Planilha 7.6. A terceira regra engloba 68 instâncias adicionando 26 novas instâncias à discriminação do modo a pé, erroneamente classificando 13 que de fato pertencem aos modos carro e ônibus e repetindo a classificação para 29 instâncias já classificadas pela regra anterior, como mostra a Figura 7.5. A quarta e última regra para o modo a pé nesta árvore abrange 66 instâncias, todas já anteriormente classificadas pelas regras 1 e 3 de maneira que não vem a contribuir para a discriminação desse modo de viagem. Encerrada a aplicação das regras existentes para o modo a pé na Planilha 7.6 para a

determinada árvore, verifica-se então o mesmo índice de acerto de classificação fornecido como saída pelo algoritmo Trepan ao executar um teste ou previsão utilizando o mesmo conjunto de instâncias (vide quadro de desempenho da Figura 7.2 para o conjunto de treinamento). Foram igualmente aplicadas, com o auxílio de uma planilha eletrônica e segundo a hierarquia determinada pela árvore de decisão, as restantes regras para os outros modos e também foi verificado o mesmo percentual de acerto de saída do algoritmo para tal conjunto de instâncias.

A ocorrência de uma regra desnecessária, como por exemplo a quarta regra para o modo a pé da árvore em questão, provavelmente constitui uma forma encontrada pelo algoritmo de contornar a estrutura rígida da árvore de decisão binária para alcançar níveis de profundidade que tragam ganho à discriminação das classes envolvidas. Essa situação pode ser melhor visualizada quando são envolvidas outras regras da mesma árvore da Figura 7.2. A décima quinta regra, da árvore quando aplicada segundo a determinada hierarquia ao conjunto de 567 instâncias, abrange 27 novas instâncias a classe ônibus porém as seguintes regras são coincidentes, ou seja, abrangem exatamente as mesmas instâncias dessa regra 15, conforme mostrado na Figura 7.6.



regras 15, 16, 17 e 18

Figura 7.6: Diagrama das regras coincidentes 15, 16, 17 e 18 da árvore da rede 5-9-3 para o modo ônibus, com o número de instâncias compreendidas por local.

Comparando-se as regras 15 e 16 vê-se que estas somente diferenciam-se na passagem pelo nó da profundidade 3 à direita da raiz que impõe restrição alternativa entre os atributos a respeito da renda e do sexo, e que neste caso realmente não faz diferença: ambas as regras possuem imagem coincidente. A montante, as regras antecessoras de números 11 e 12 (Planilha 7.6), ao passarem pelo mesmo nó, formam a folha para o modo carro na profundidade 4 à direita da raiz e tais regras não são coincidentes, ou seja, "tiram proveito" de ambas as restrições por atributos relacionadas naquele nó para a discriminação da classe carro, como mostrado na coluna à direita da Planilha 7.6, nas linhas dessas regras. Situação semelhante pode ser observada um nível acima no nó da profundidade 1 também à direita da raiz. Se para todas as regras sucessoras às desse nível não faz diferença alguma a passagem pela restrição por atributo de idade do viajante ou número de automóveis do domicílio, para as regras 9 e 10 (Planilha 7.6) que formam a folha da classe ônibus nesse nível de profundidade 1 as restrições por aqueles diferentes atributos contribuem muito à classe ônibus. A jusante, a regra sucessora de número 19 para a folha da classe carro, localizada na última profundidade (nível 7) à direita da raiz, forma par complementar à regra 15 para a classe ônibus, tendo como única diferença em relação à regra 11 para a mesma classe apenas a inclusão da restrição para o número de automóveis no domicílio, o que traz um ganho para a discriminação da classe carro, não importando a passagem pelas restrições alternativas que compõem as demais regras naquela profundidade, regras 20, 21 e 22.

A correspondência entre os índices de acerto de classificação obtidos na Planilha 7.6 e aqueles referentes ao mesmo conjunto no quadro de desempenho da árvore da Figura 7.2 (coluna Trein.) indicam que o ordenamento das regras conforme dispostas na Planilha 7.6 corresponde à estrutura hierárquica de aplicação dessa árvore. O caso de aplicação das regras independentemente, ou isoladamente, à tal estrutura hierárquica foi estudado para a mesma árvore da Figura 7.2 e os resultados encontrados para cada regra aplicada àquele conjunto de 567 instâncias estão apresentados na Planilha 7.12. Nesse caso não há a remoção das instâncias abrangidas pela regra anterior e cada regra possui o domínio de aplicação de todo o conjunto de 567 instâncias.

Planilha 7.12: Número de instâncias corretamente classificadas e correspondente percentual de acerto de classificação das regras para os modos de viagem da árvore extraída da rede 5-9-3 com profundidade 7 aplicadas individualmente, fora da estrutura hierarquica da árvore, a um conjunto de 567 instâncias.

regra 1	dist<=2,0km:	n_autos=0: A_PE			76 --> 40,0%
regra 2	dist<=2,0km:	idade<=22: A_PE			46 --> 24,2%
regra 3	dist<1,1km:	n_autos=1: A_PE			26 --> 13,7%
regra 4	dist<1,1km:	n_idade>22:	n_autos<=1: A_PE		60 --> 32,0%
regra 5	dist<1,1km:	n_autos>1: CARRO			12 --> 6,7%
regra 6	dist<1,1km:	n_idade>22:	n_autos>1: CARRO		12 --> 6,7%
regra 7	1,1<=dist<=2,0km:	n_autos>=1: CARRO			40 --> 22,3%
regra 8	1,1<=dist<=2,0km:	n_idade>22: CARRO			35 --> 19,6%
regra 9	dist>2,0km:	n_idade<=18: ONIBUS			32 --> 16,2%
regra 10	dist>2,0km:	n_autos=0: ONIBUS			89 --> 44,9%
regra 11	dist>2,0km:	n_idade>18:	n_salar>22sal-mín: CARRO		36 --> 20,1%
regra 12	dist>2,0km:	n_idade>18:	homem: CARRO		69 --> 38,5%
regra 13	dist>2,0km:	n_autos>=1:	n_salar>22sal-mín: CARRO		37 --> 20,7%
regra 14	dist>2,0km:	n_autos>=1:	homem: CARRO		68 --> 30,8%
regra 15	dist>2,0km:	n_idade>18:	n_salar<=22sal-mín:	n_autos<=1: ONIBUS	121 --> 61,1%
regra 16	dist>2,0km:	n_idade>18:	mulher:	n_autos<=1: ONIBUS	77 --> 38,9%
regra 17	dist>2,0km:	n_autos=1:	n_salar<= 22sal-mín: ONIBUS		55 --> 27,8%
regra 18	dist>2,0km:	n_autos=1:	mulher: ONIBUS		38 --> 19,2%
regra 19	dist>2,0km:	n_idade>18:	n_salar<=22sal-mín:	n_autos > 1: CARRO	32 --> 17,9%
regra 20	dist>2,0km:	n_idade>18:	mulher:	n_autos > 1: CARRO	24 --> 13,4%
regra 21	dist>2,0km:	n_autos>1:	n_salar<=22sal-mín: CARRO		25 --> 14,0%
regra 22	dist>2,0km:	n_autos>1:	mulher: CARRO		24 --> 13,4%

Planilha 7.13: Número de instâncias adicionais corretamente classificadas e correspondente percentual de acerto de classificação das regras para os modos de viagem da árvore extraída da rede 5-9-3 com profundidade 7, aplicadas de acordo com a hierarquia da árvore apenas em relação ao modo de viagem, a um conjunto de 567 instâncias.

regra 1	dist<=2,0km:	n_autos=0: A_PE			76 --> 40,0%	
regra 2	dist<=2,0km:	idade<=22: A_PE			46 --> 24,2%	
regra 3	dist<1,1km:	n_autos=1: A_PE			26 --> 13,7%	
regra 4	dist<1,1km:	n_idade>22:	n_autos<=1: A_PE		0 (contida nas regras 1 e 3)	/ total A_PE 148 --> **77,9%**
regra 5	dist<1,1km:	n_autos>1: CARRO			12 --> 6,7%	
regra 6	dist<1,1km:	n_idade>22:	n_autos>1: CARRO		0 (coincidente à rgr. 5)	
regra 7	1,1<=dist<=2,0km:	n_autos>=1: CARRO			40 --> 22,3%	
regra 8	1,1<=dist<=2,0km:	n_idade>22: CARRO			0	
regra 11	dist>2,0km:	n_idade>18:	n_salar>22sal-mín: CARRO		36 --> 20,1%	
regra 12	dist>2,0km:	n_idade>18:	homem: CARRO		48 --> 26,8%	
regra 13	dist>2,0km:	n_autos>=1:	n_salar>22sal-mín: CARRO		1 --> 0,6%	
regra 14	dist>2,0km:	n_autos>=1:	homem: CARRO		0	
regra 19	dist>2,0km:	n_idade>18:	n_salar<=22sal-mín:	n_autos > 1: CARRO	11 --> 6,1%	
regra 20	dist>2,0km:	n_idade>18:	mulher:	n_autos > 1: CARRO	0	
regra 21	dist>2,0km:	n_autos>1:	n_salar<=22sal-mín: CARRO		0	
regra 22	dist>2,0km:	n_autos>1:	mulher: CARRO		0	/ total CARRO 148 --> **82,7%**
regra 9	dist>2,0km:	n_idade<=18: ONIBUS			32 --> 16,2%	
regra 10	dist>2,0km:	n_autos=0: ONIBUS			81 --> 40,9%	
regra 15	dist>2,0km:	n_idade>18:	n_salar<=22sal-mín:	n_autos<=1: ONIBUS	41 --> 20,7%	
regra 16	dist>2,0km:	n_idade>18:	mulher:	n_autos<=1: ONIBUS	1 --> 0,5%	
regra 17	dist>2,0km:	n_autos=1:	n_salar<= 22sal-mín: ONIBUS		0	
regra 18	dist>2,0km:	n_autos=1:	mulher: ONIBUS		0	/ total ONIBUS 155 --> **78,3%**

Na análise da Planilha 7.12 verifica-se que embora algumas regras individualmente aplicadas tenham atingido um percentual de acerto de classificação notável, como é o caso das regras 15, 10, 1, 16 e12,em ordem decrescente de percentual, nenhuma regra sozinha alcança os percentuais de acerto de classificação, correspondente a cada classe, atingidos pelas regras aplicadas conjuntamente segundo o arranjo ordenado da estrutura da árvore. Tal fato sugere que as regras não devam ser tomadas separadamente da árvore como um conhecimento isolado e completo por si. Neste contexto, cada regra perfaz uma tarefa dentro da estrutura de decisão para a qual foi criada.

Também foi estudado o caso da aplicação das regras apenas com hierarquia dentro de uma classe, ou modo de viagem. Dessa forma, a primeira regra para um modo tem todas as 567 instâncias do conjunto como domínio de aplicação, a regra seguinte possui somente as instâncias restantes, ou não abrangidas pela primeira, como seu domínio de aplicação e assim por diante. O ciclo de aplicação das regras é reiniciado para cada modo, ou seja, a primeira regra de cada modo tem todo o conjunto de instâncias como domínio novamente. Os resultados desse caso estudado são mostrados na Planilha 7.13. Na análise dessa planilha verifica-se que os percentuais de acerto de classificação para as classes se assemelham aos alcançados pelas regras aplicadas segundo a hierarquia da árvore de decisão (Planilha 7.6), porém com um pequeno aumento nos percentuais dos modos carro e ônibus devido ao fato de o domínio de aplicação ser maior neste caso pela não remoção das instâncias causada por regras antecessoras.

Portanto, a discussão que se segue a respeito da análise e interpretação das árvores resultantes baseia-se na informação ou conhecimento contido na árvore como um corpo único e não apenas nas partes constituintes, as regras.

7.2.1 Análise e Interpretação

As árvores resultantes possuem um percentual de acerto de classificação médio entre modos para todos os conjuntos testados, incluindo o conjunto com toda a base de dados, de 74,6%. Esse número percentual é próximo dos desempenhos médios atingidos pelas correspondentes redes originárias, o que associado ao percentual médio para as quatro árvores de fidelidade à rede, também para todos os conjuntos testados, de 88,9%, evidencia a representatividade dessas estruturas

de conhecimento em relação a suas redes originárias, bem como sua eficácia para os dados do domínio estudado.

De certa forma, de acordo com as árvores resultantes, apesar das redes originárias das árvores possuírem diferentes arquiteturas e terem sido treinadas com diferentes atributos de entrada, sob condições iniciais de pesos de conexões diferentes e com diferentes taxas de aprendizagem, concordam sobre a maneira de discriminar cada um dos modos de viagem no espaço de instâncias estudado. A redundância ou quantidade de similaridades observadas entre as quatro árvores resultantes pode conduzir a um delineamento das características gerais de uso de cada modo ou a uma caracterização, ainda que grosseira, do comportamento, ou situação, de escolha para cada modo de viagem, restrita à confiabilidade do percentual de acerto de classificação dessas árvores e mapeada pelas correspondentes redes neurais.

Para o modo a pé, ou caminhada, pode-se dizer que a escolha seria feita quando a distância de viagem é menor que aproximadamente 2 quilômetros, há um ou nenhum automóvel no domicílio e principalmente por aqueles indivíduos jovens adultos com idade inferior a aproximadamente 22 anos. O principal fator determinante de escolha por esse modo mostra-se, em todas as árvores, ser a distância de viagem.

Para o modo carro, ou automóvel, a hipótese da escolha para distâncias menores que aproximadamente 2 quilômetros é que viajariam de carro aqueles com mais de um carro no domicílio e majoritariamente os maiores de 23 anos. Para distâncias maiores que essa, viajariam de automóvel os maiores de idade do sexo feminino apenas quando houvesse mais de um automóvel no domicílio, ou seja, havendo restrição de existência de um único automóvel no domicílio viajariam os maiores de idade do sexo masculino. Além da distância, o outro atributo que se demonstra relevante para a escolha desse modo é o número de automóveis do domicílio além da maioridade.

Todas as árvores escolhidas concordam que, quando a distância é superior a 2 quilômetros e os viajantes são menores de idade ou não há automóvel no domicílio, o ônibus seria o modo de viagem preferencialmente escolhido. Também concordam que para a distância superior a 2 quilômetros se só há um automóvel no domicílio do viajante maior de idade e ou a renda familiar é inferior a

aproximadamente 15 salários mínimos ou seu sexo é feminino, o modo escolhido para realizar a viagem seria o ônibus. Essa última hipótese em contraponto com outra complementar para o modo carro sugere, em todas as árvores, que o viajante do sexo feminino com um único automóvel no domicílio somente trocaria do modo de viagem ônibus para o modo carro quando a renda domiciliar ultrapassasse aproximadamente 15 salários mínimos. Ou seja, o atributo sexo deixaria de ser importante para os maiores de idade na decisão de escolha entre os modos carro e ônibus para realizar viagens intraurbanas em distâncias maiores que aproximadamente 2 quilômetros, havendo a restrição de um único automóvel pertencente ao domicílio, quando a renda fosse superior a tal limite. Havendo 2 ou mais automóveis no domicílio a escolha do modo carro para realizar viagens a quaisquer distâncias parece independer de outros atributos para os viajantes maiores de idade.

Em termos da utilidade dos resultados para os objetivos buscados por este estudo, não foi possível a partir das estruturas de conhecimento, as árvores de decisão extraídas das redes neurais artificiais ajustadas aos dados da pesquisa de Bauru, apontar qual das estruturas do processo de escolha modal, implícitas nos modelos compensatórios, não-compensatórios e semi-compensatórios, revistas no capítulo 2 deste texto, mais se aproxima delas. Porém para o objetivo de compreender melhor o processo de escolha do modo de viagem este estudo traz um auxílio para o estabelecimento de hipóteses para a elaboração de modelos desse processo na medida em que indica não somente quais atributos são relevantes, como por exemplo fez o estudo conduzido por SHMUELI et al. (1996), mas algo sobre a relação entre eles, incluindo seus valores limiares às restrições, e de acordo com cada modo de viagem. Por exemplo, vê-se, nas árvores de decisão escolhidas, que há uma relação de compensação entre o atributo sexo do viajante e os atributos relacionados com a capacidade econômica de seu domicílio, representada pelos atributos renda domiciliar e número de automóveis do domicílio, na decisão de escolha do modo carro para a realização de uma viagem.

A novidade, ou desvio sobre o conhecimento prévio sobre o domínio, foi a necessidade de apenas um atributo que representasse a função custo para o deslocamento realizado em uma viagem, o atributo tempo de viagem por modo caminhada (t_ape), apresentado pela árvore de menor profundidade extraída da rede 11-11-3. Os outros atributos de entrada relacionados ao deslocamento que

esta rede dispunha, como por exemplo o tempo, e em menor grau o custo de viagem por modo ônibus e o custo ou o tempo de viagem por modo automóvel, sequer foram contemplados à necessidade para as condições de escolha dos modos aos quais estes atributos se referem. Caso se dispusesse deste conhecimento prévio certamente não se teria confeccionado uma amostra, para utilização neste estudo, para se contornar a falta de informações na pesquisa domiciliar sobre viagens por modo ônibus, uma vez que a obtenção desta informação, feita de forma indireta, posterior à coleta em campo, mesmo que auxiliada por ferramentas de SIG, mostrou-se muito custosa, mas sim se utilizado todos os registros de viagem contidos na base de dados dessa pesquisa domiciliar já na etapa de mineração de dados. O fato de a rede ter dispensado os demais atributos relacionados ao deslocamento de viagem pode dever-se à forma indireta com que estes foram obtidos, calculados a partir da distância de caminho mínino (dist), a qual pode ter sido mapeada, ou reconhecida, quanto à proporcionalidade ou dependência linear.

Como esses atributos que foram dispensados tiveram seu cálculo baseado em hipóteses que se aproximam da realidade do ambiente donde provém os dados (vide sub-ítem 5.3.3.2 deste texto), talvez o fato de a rede os ter dispensado reflita a não necessidade de outros atributos de viagem além da distância de caminho mínimo para a modelagem da escolha do modo de viagem por redes neurais artificiais em cidades de pequeno e médio porte, como é o caso de Bauru, onde por não haver o problema de congestionamento de trânsito, encontrado na maioria das cidades de grande porte do mundo, há uma proporcionalidade ou relação de dependência linear entre os atributos relacionados ao deslocamento que os torna, exceto um, a distância de caminho mínimo, dispensáveis a essa modelagem.

Expediente matemático necessário aos modelos convencionais de escolha (MNL), o emprego dos atributos relativos ao deslocamento para cada modo de viagem, conforme observado nos estudos anteriores para a modelagem de escolha por redes neurais artificiais, listados no ítem 3.3 "Redes Neurais no Domínio da Escolha Modal em Transportes" deste texto, provavelmente reflita uma forma de compor o conjunto dos atributos de entrada *ad instar* àquela necessária aos modelos convencionais, uma vez que tais estudos estipulam uma comparação de performance entre os paradigmas neural artificial e o matemático convencional.

Quanto à simplicidade dos achados, as quatro árvores de decisão escolhidas como resultados da etapa de mineração demonstram baixa complexidade sintática porque possuem profundidade rasa, nível 7, o que auxilia a interpretação da informação contida nessas estruturas.

8 CONCLUSÕES E RECOMENDAÇÕES

8.1 CONCLUSÕES

Embora as condições, ou restrições a sujeitar-se, apresentadas nas árvores resultantes, por conterem informações necessárias a uma discriminação entre os modos de viagem, possam aproximar a situação de descrição (mostrada na Figura 6.1) para a divisão modal do conjunto de dados empregado (fato evidenciado pelo desempenho médio aproximado de 75% de acerto na classificação), certamente estas não são suficientes ao fechamento da caracterização, ou completa descrição, sobre o comportamento de escolha para os modos de transporte envolvidos, e tampouco apontam de qual das estruturas subjacentes aos modelos apresentados no capítulo 2 mais se aproximam. Entretanto as árvores resultantes sugerem uma unicidade de estrutura de conhecimento para o espaço de instâncias estudado, pela quantidade de similaridades observadas entre elas tanto na forma da estrutura quanto na informação contida. Informação ou conhecimento o qual pôde proporcionar as considerações a respeito de cada um dos modos de viagem enfocados, apresentadas no capítulo anterior. Dentre essas considerações destaca-se a relação observada em todas as árvores resultantes de compensação entre o atributo sexo do viajante e os atributos relacionados com a capacidade econômica de seu domicílio, representados pelos atributos renda domiciliar e número de automóveis do domicílio, na decisão de escolha do modo carro para a realização de uma viagem.

Como um achado, as árvores resultantes sugerem a não necessidade de coleta de informações, ou atributos de viagem, referentes a cada modo de transporte que sejam função do deslocamento realizado em uma viagem, para a modelagem da divisão modal por redes neurais artificiais do tipo MLP. O dispensar

de mais de um atributo relacionado ao deslocamento de viagem pode estar baseado na forma indireta pela qual estes atributos foram obtidos, ou seja, calculados a partir da distância de caminho mínimo. Os atributos assim calculados possuem uma dependência linear ou proporcionalidade ao atributo original, distância de caminho mínimo (dist), a qual pode ter sido mapeada, ou reconhecida, pela rede neural artificial que os dispensou quanto a auxiliares na sua tarefa de discriminação das classes, ou modos de viagem envolvidos. Todavia, esses atributos dispensados pela RNA foram calculados a partir de hipóteses próximas da realidade do ambiente donde provém os dados. Isso torna possível a idéia de que para a modelagem por RNA da escolha do modo de viagem em cidades de porte pequeno e médio (como é o caso da cidade de Bauru, onde não há o problema de congestionamento de trânsito, encontrado na maioria das cidades de grande porte do mundo) seja dispensável mais do que um atributo relacionado ao deslocamento de viagem por haver proporcionalidade ou relação de dependência linear entre tais atributos. O haver informações em cada instância sobre atributos de viagem para os modos além daquele pelo qual a viagem se realizou surgiu por uma necessidade matemática de modelos probabilísticos como os modelos de utilidade aleatória, como por exemplo o modelo Logit. O dispensar da coleta de informações referentes aos modos alternativos não escolhidos à realização de uma viagem colaboraria para uma economia de recursos na fase de aquisição de dados nos estudos para fins de previsão da divisão modal em transportes que venham a utilizar RNA para a modelagem.

As condições apresentadas nas árvores para a escolha do modo carro, sempre envolvendo a maioridade, inclusive em algumas delas com o valor limiar do atributo no teste de divisão do nó próximo da idade (18 anos) permitida para obtenção de habilitação para condução de veículos automotores, sugere que as redes tenham aprendido a informação contida no atributo carteira de motorista mesmo que este não tenha sido apresentado ao treinamento delas.

A seleção de instâncias exclusivas para amostragem feita através de segmentação de mercado, justificada pela característica de repetição de apresentação do processo de treinamento de RNA escolhido, pelo alto custo de coleta de um registro para a amostra, e pelo perfil desagregado da modelagem almejado, que despreza eventuais modas (tendências) de comportamento

agregado, mostrou-se eficaz ao propósito de transmitir informações a respeito da "população", base de dados da pesquisa domiciliar, para a aprendizagem das redes, pela semelhante performance de predição entre os conjuntos de teste da amostra e de toda a base de dados. Além disso, o número de instâncias do conjunto de treinamento, aproximadamente 550, mostrou-se capaz de fornecer quantidade suficiente de informação ao treinamento de uma rede para que esta obtenha capacidade de generalização, no domínio da escolha modal enfocado.

O valor limiar de restrição para o atributo distância de viagem[1] em torno de 2 quilômetros para separação entre modos de transporte, descoberto no conhecimento adquirido pelas redes treinadas no espaço de instâncias empregado e apresentado pelas árvores de decisão resultantes, é concordante com os valores observados para esse atributo em relação ao modo a pé por ARRUDA (2000) (Figuras C.1 e C.2 do anexo C) e próximo dos observados pelo FHWA (1992) e SANCHES (1988).

Ao contrário do algoritmo de indução de árvores a partir de dados , CART, utilizado por SHMUELI et al. (1996), a árvore extraída pelo algoritmo Trepan mostrou-se um auxílio útil na determinação de quais atributos são relevantes, ou seja, na seleção dos atributos de entrada para a modelagem por RNA para o domínio da escolha do modo de viagem estudado. É provável que isso se deva ao fato de o algoritmo Trepan utilizar a própria rede neural como motor classificador, possuindo assim um *bias* semelhante ao da rede.

8.2 RECOMENDAÇÕES

Recomenda-se aplicar as quatro árvores de decisão e redes neurais resultantes deste trabalho a dados provenientes de outras localidades para estimação da acurácia preditiva dessas estruturas de conhecimento fora do ambiente no qual foram ajustadas. Além dessa verificação da capacidade de translação espacial, recomenda-se também a geração de novas árvores ajustadas para outras localidades a fim de compará-las às obtidas por este estudo.

[1] Essa distância refere-se a distância de caminho mínimo (dist), atributo de viagem definido no capítulo 5, sub-ítem 5.3.3.2, deste texto.

As árvores de decisão, por possuírem uma raiz única de onde derivam os ramos para as diferentes classes ou folhas, eventualmente criam nós, profundidade abaixo, com repetições em seus testes de divisão dos atributos do nó raiz, porém com outros valores de limiar de comparação, que por estarem mais próximos às folhas são mais específicos a determinada classe do que o determinado pelo nó raiz. A saída da estrutura rígida arbórea por repetições de atributos, para um melhor discriminar das classes utilizada pelo Trepan, pode aumentar a profundidade da árvore e dificultar a interpretação do conhecimento contido. Recomenda-se duas alternativas a trabalhos futuros: treinar redes neurais tipo MLP para discriminar os modos de viagem de forma seqüencial e aninhada, similarmente à forma empregada pelos modelos Logit Aninhados[2] (NMNL); e ao ponto das árvores obtidas no trabalho, transformar as árvores nos "gráficos acíclicos de exceção dirigida" (EDAG[3]) apresentados por GAINES (1996). O EDAG pode ser visto como uma generalização da árvore de decisão, o que supera o problema dos conceitos disjuntivos que envolvem replicações de sub-árvores, dentro das árvores de decisão. Problemas observados em algumas árvores geradas pelo algoritmo Trepan, como por exemplo na árvore da Figura A.3, em que há a inclusão de várias categorias de um mesmo atributo nos testes de divisão de nós, o que dificulta a interpretação sobre qual é a restrição satisfeita, e seu correspondente caso contrário, nos ramos à jusante do nó, são tratados quando se transforma a estrutura de uma árvore de decisão em estrutura de um EDAG.

[2] Ver SOBEL (1980), BEN-AKIVA & LERMAN (1985) e ORTÚZAR & WILLUMSEN (1994) para maiores detalhes sobre os modelos Logit Aninhados (Nested Multinomial Logit).
[3] EDAG abrevia o original em inglês "Exception Directed Acyclic Graph".

ANEXO A:

Árvores de decisão

QUADRO DE DESEMPENHO DA ÁRVORE				
Profundidade 30		Trein.	Valid.	Teste
% Fidelidade à MLP-32-10-10-3		94,5	94,2	92,4
% Acerto à Classific.	Médio entre modos	68,4	65,2	70,5
	Carro	97,7	99,0	97,6
	A pé	52,7	42,7	49,4
	Ônibus	55,7	53,5	65,2

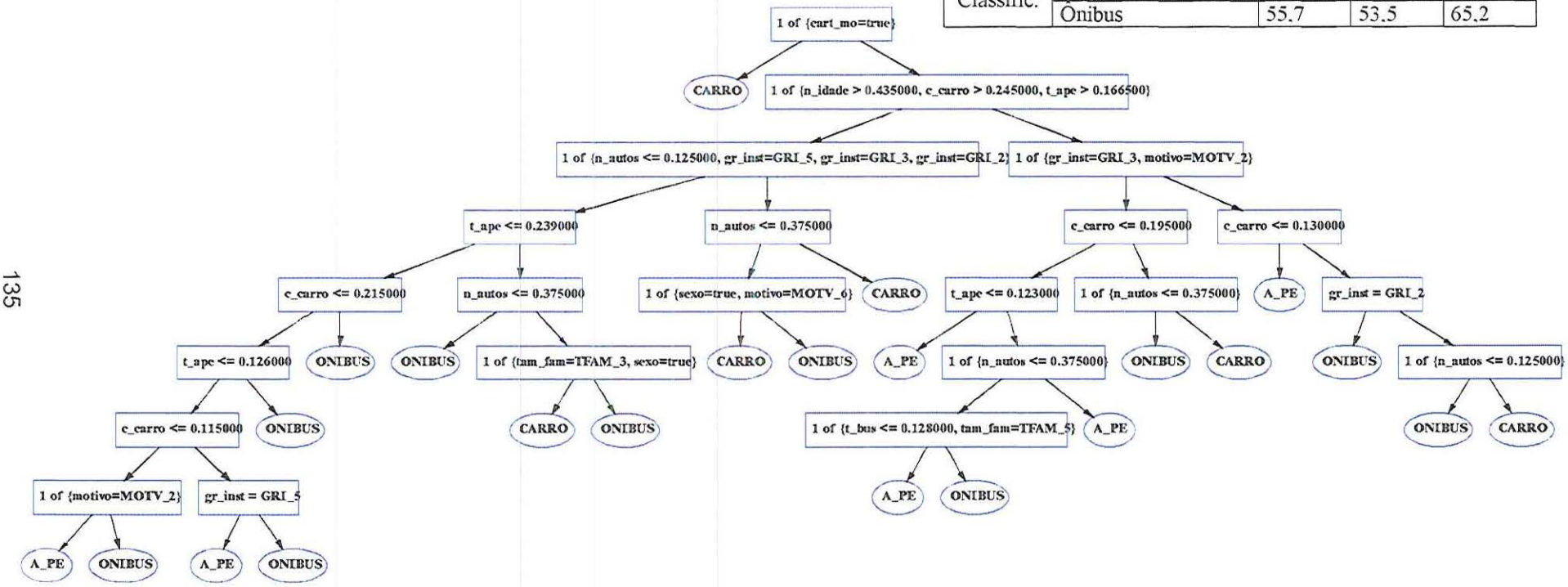


Figura A.1: Árvore de profundidade 30 extraída da rede 32-10-10-3

QUADRO DE DESEMPENHO DA ÁRVORE				
Profundidade 31		Trein.	Valid.	Teste
% Fidelidade à MLP-32-15-15-3		91,0	90,2	91,3
% Acerto à Classific.	Médio entre modos	63,6	61,4	67,4
	Carro	98,9	99,0	100,0
	A pé	41,3	31,3	36,8
	Ônibus	51,9	52,5	66,3

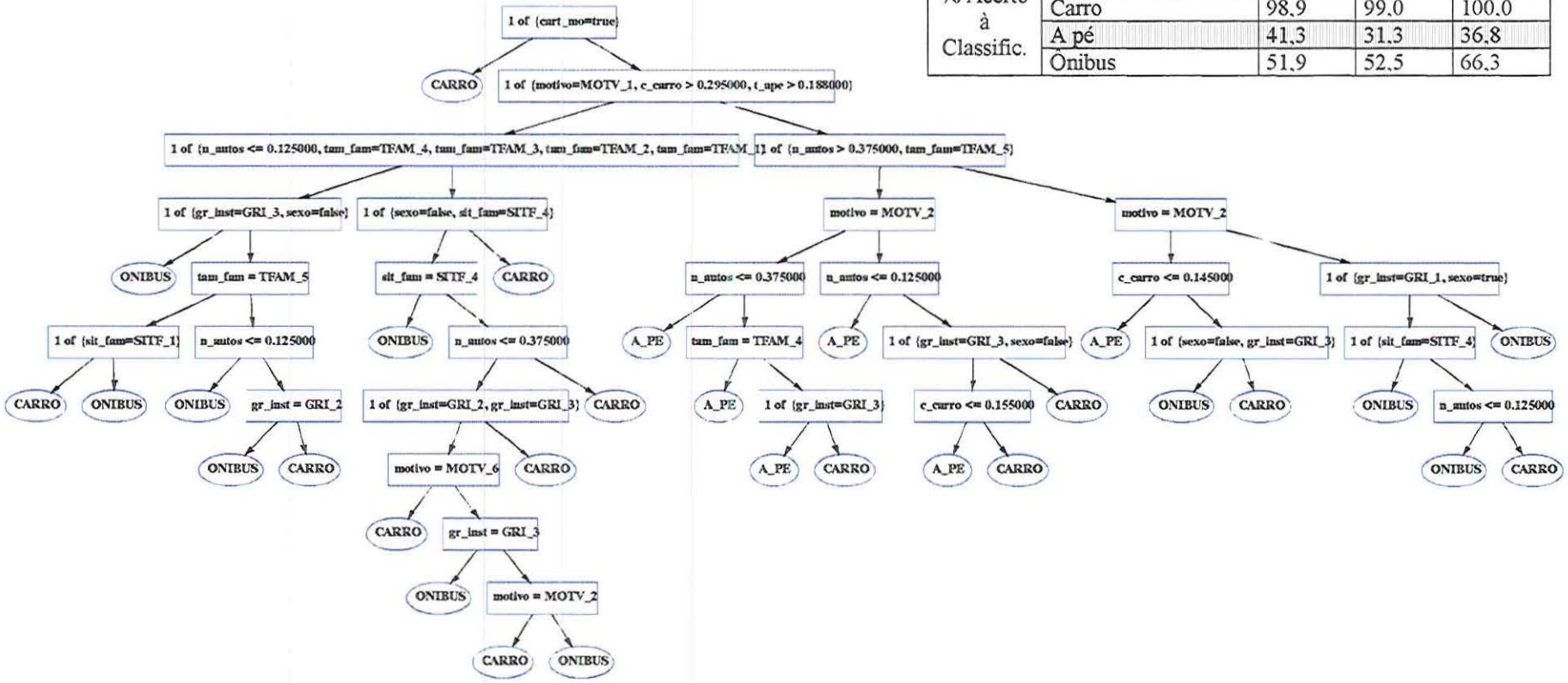


Figura A.2:Árvore de profundidade 31 extraída da rede 32-15-15-3.

QUADRO DE DESEMPENHO DA ÁRVORE				
Profundidade 48		Trein.	Valid.	Teste
% Fidelidade à MLP-31-10-10-3		82,4	78,6	83,3
% Acerto à Classific.	Médio entre modos	62,3	62,4	69,7
	Carro	70,1	66,0	70,6
	A pé	44,0	38,5	56,3
	Ônibus	73,2	81,8	81,5

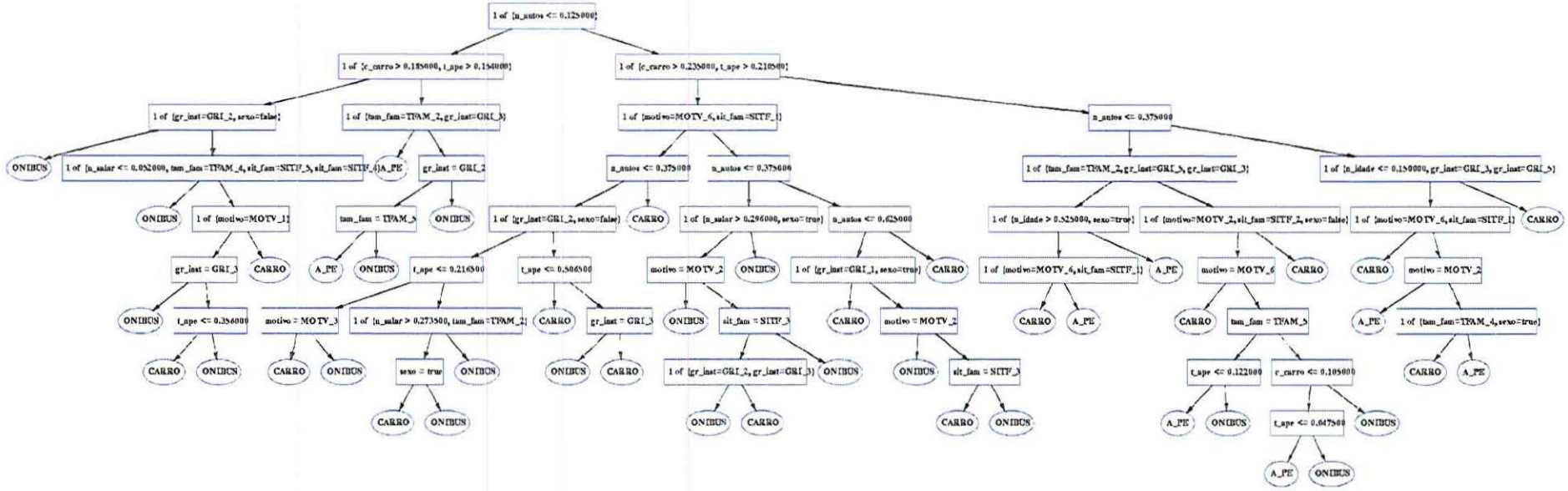


Figura A.3: Árvore de profundidade 48 extraída da rede 31-10-10-3.

QUADRO DE DESEMPENHO DA ÁRVORE				
Profundidade 29		Trein.	Valid.	Teste
% Fidelidade à MLP-31-30-15-3		80,7	76,9	80,3
% Acerto à Classific.	Médio entre modos	60,8	56,9	61,4
	Carro	74,0	66,0	67,1
	A pé	67,9	60,4	73,6
	Ônibus	50,0	44,4	44,6

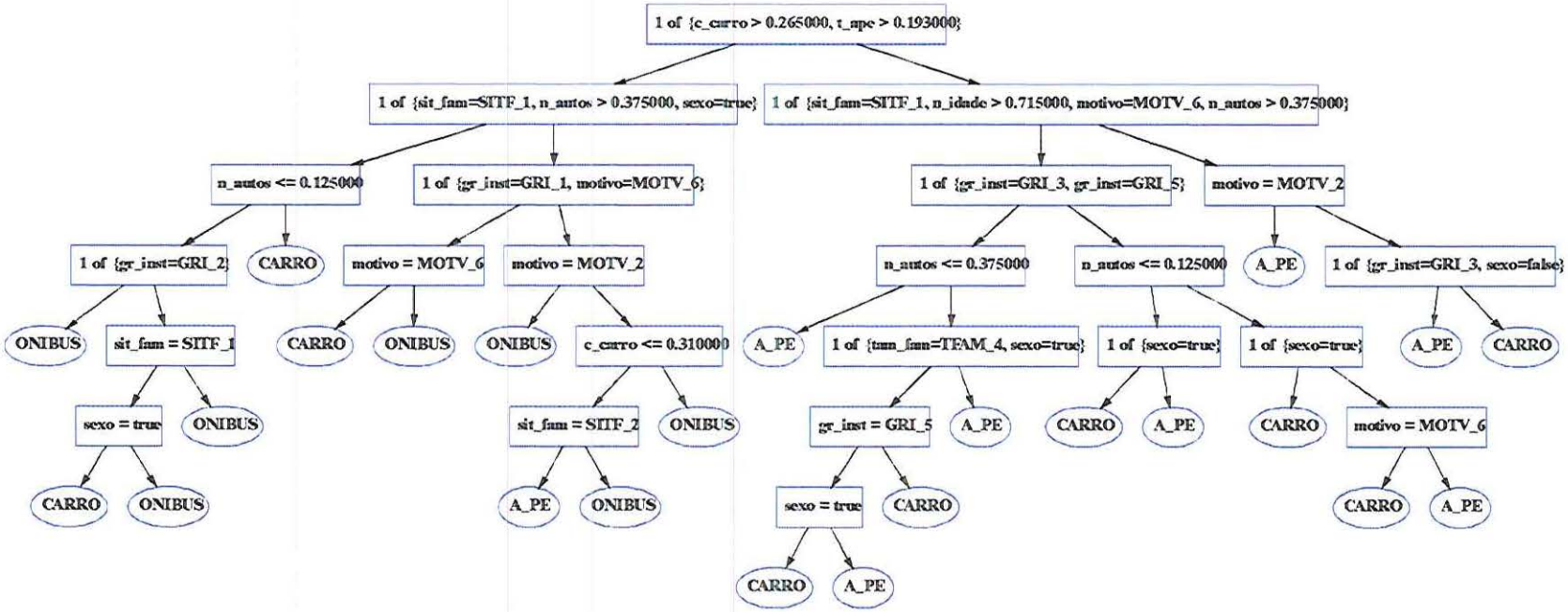


Figura A.4: Árvore de profundidade 29 extraída da rede 31-30-15-3.

QUADRO DE DESEMPENHO DA ÁRVORE				
Profundidade 35		Trein.	Valid.	Teste
% Fidelidade à MLP-33-30-20-3		81,6	81,4	86,6
% Acerto à Classific.	Médio entre modos	57,3	54,3	58,5
	Carro	88,3	85,1	87,6
	A pé	33,5	36,9	40,6
	Ônibus	52,3	38,5	46,5

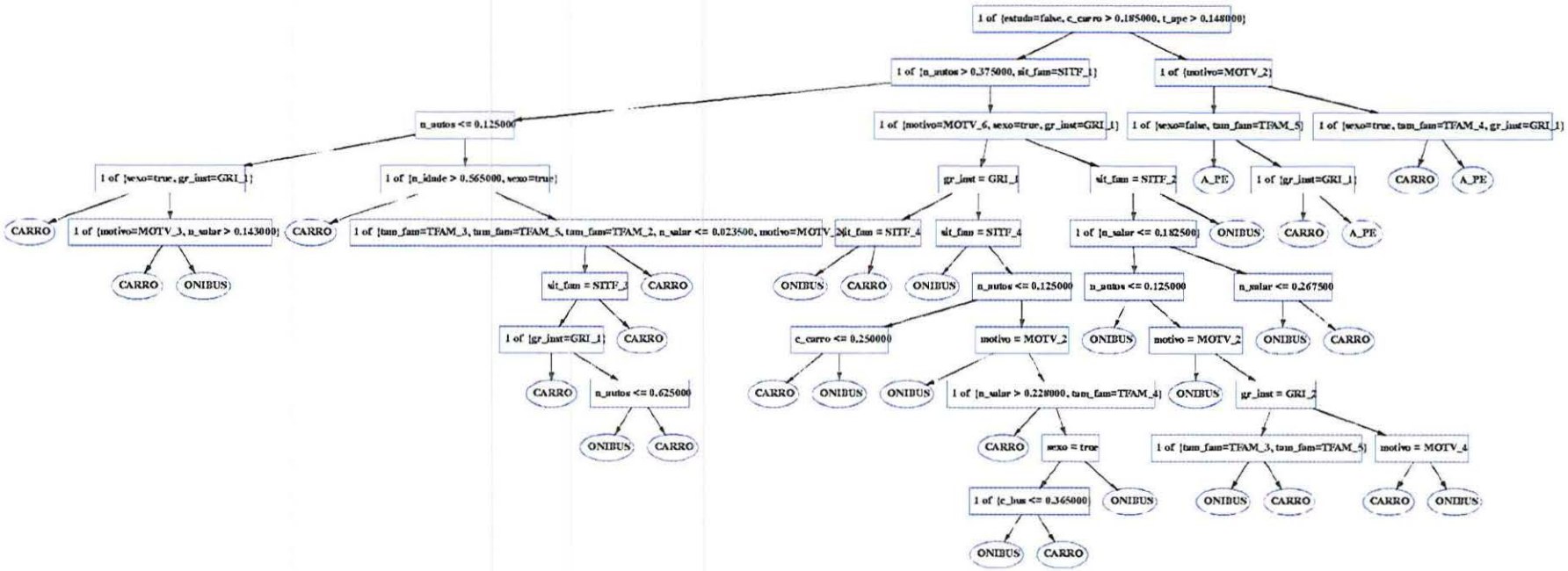


Figura A.5: Árvore extraída da rede 33-30-20-3 com profundidade 35.

QUADRO DE DESEMPENHO DA ÁRVORE				
Profundidade 21		Trein.	Valid.	Teste
% Fidelidade à MLP-33-20-15-3		82,2	79,2	79,6
% Acerto à Classific.	Médio entre modos	60,9	59,5	63,4
	Carro	88,3	80,9	87,6
	A pé	47,8	47,6	52,5
	Ônibus	49,2	45,1	48,8

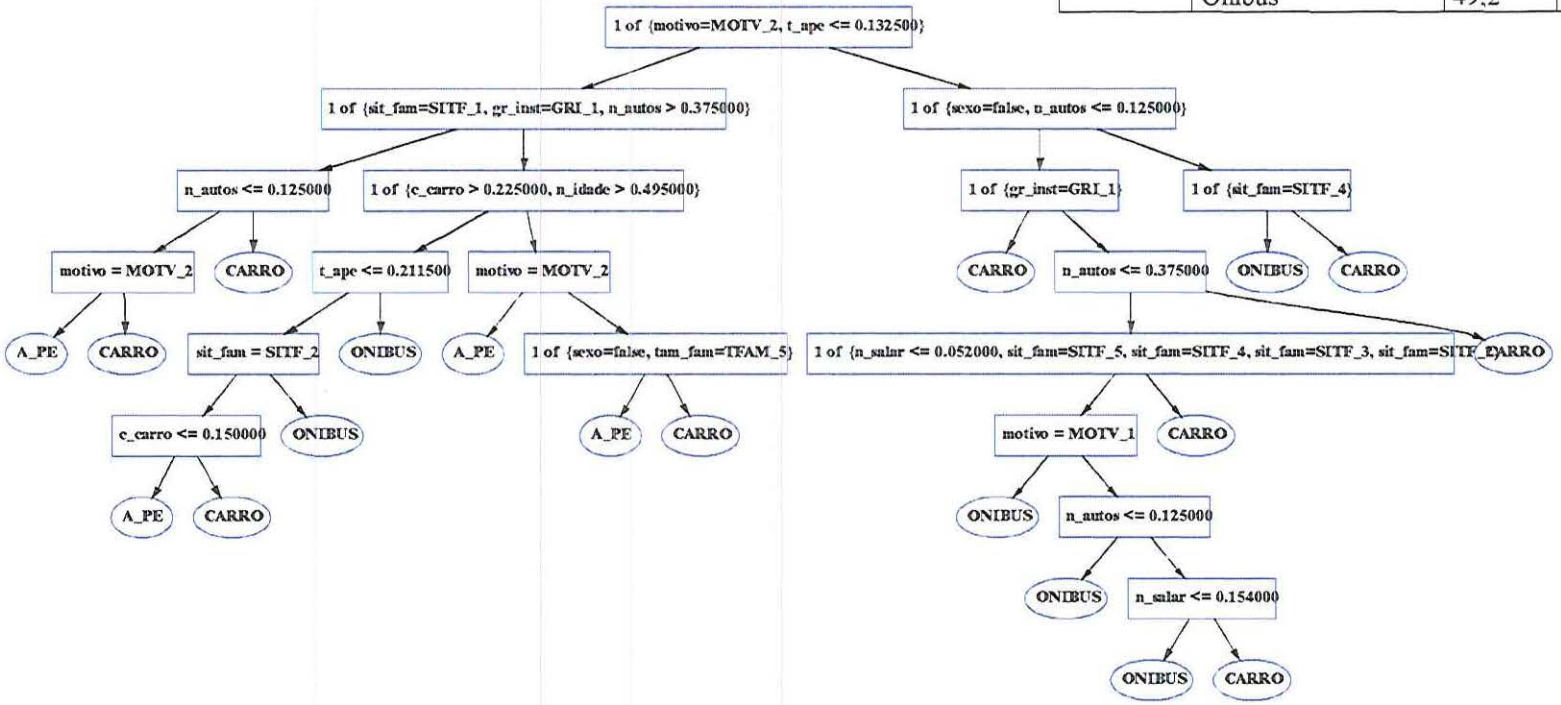


Figura A.6: Árvore extraída da rede 33-20-15-3 com profundidade 21.

ANEXO B:

Árvores de decisão e planilhas de condições

QUADRO DE DESEMPENHO DA ÁRVORE				
Profundidade 46		Trein.	Valid.	Teste
% Fidelidade à MLP-11-30-3		88,8	86,5	84,4
% Acerto à Classific.	Médio entre modos	79,2	76,4	70,5
	Carro	87,2	78,7	77,2
	A pé	79,1	74,0	70,1
	Ônibus	71,1	76,5	65,0

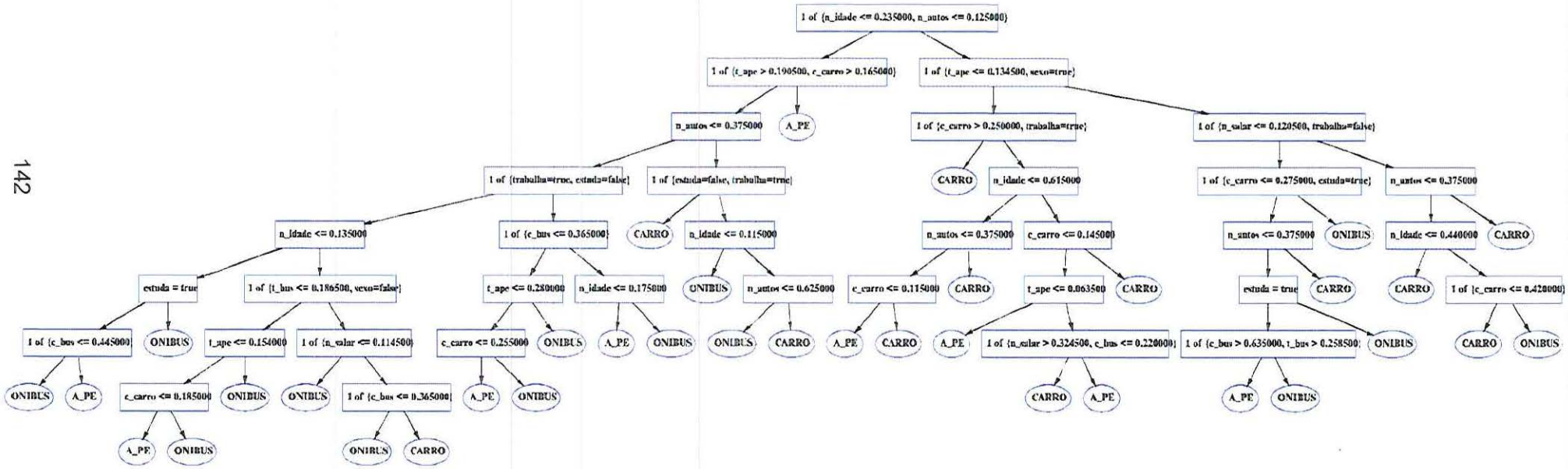


Figura B.1: Árvore extraída da rede 11-30-3 com profundidade 46.

Planilha B.1: Condições para os modos de viagem da árvore extraída da rede 11-30-3 com profundidade 46 apresentadas conforme ordenamento por modo.

a_pe	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	trabalha ou ñ_estuda :	idade<16 :
a_pe	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	trabalha ou ñ_estuda :	idade>15 :
a_pe	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	ñ_trabalha ou estuda :	c_bus<=R$0,45(meia tarifa c/ desconto) :
a_pe	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	ñ_trabalha ou estuda :	c_bus>=R$0,65(uma tarifa) :
a_pe	idade<24 ou n_autos=0 :	t_ape<33min ou dist<1,6km ou c_carro<R$0,24 : A_PE			
a_pe	idade>24 ou n_autos>=1 :	t_ape<24min ou homem :	dist<2,7km ou c_carro<R$0,38 ou ñ_trabalha :	idade<=54 :	n_autos<=1 :
a_pe	idade>24 ou n_autos>=1 :	t_ape<24min ou homem :	dist<2,7km ou c_carro<R$0,38 ou ñ_trabalha :	idade>54 :	dist<1,35km ou c_carro<R$0,21 :
a_pe	idade>24 ou n_autos>=1 :	t_ape<24min ou homem :	dist<2,7km ou c_carro<R$0,38 ou ñ_trabalha :	idade>54 :	dist<1,35km ou c_carro<R$0,21 :
a_pe	idade>24 ou n_autos>=1 :	t_ape>24min ou mulher :	n_salar<12sal-mín ou ñ_trabalha :	dist<3,0km ou c_carro<R$0,42 ou estuda :	n_autos<=1 :
carro	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	trabalha ou ñ_estuda :	idade>15 :
carro	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos>1 :	ñ_estuda ou trabalha : CARRO	
carro	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos>1 :	estuda ou ñ_trabalha :	idade>13 :
carro	idade>24 ou n_autos>=1 :	t_ape<24min ou homem :	dist>2,7km ou c_carro>R$0,38 ou trabalha : CARRO		
carro	idade>24 ou n_autos>=1 :	t_ape<24min ou homem :	dist<2,7km ou c_carro<R$0,38 ou ñ_trabalha :	idade<=54 :	n_autos<=1 :
carro	idade>24 ou n_autos>=1 :	t_ape<24min ou homem :	dist<2,7km ou c_carro<R$0,38 ou ñ_trabalha :	idade<=54 :	n_autos>1 : CARRO
carro	idade>24 ou n_autos>=1 :	t_ape<24min ou homem :	dist<2,7km ou c_carro<R$0,38 ou ñ_trabalha :	idade>54 :	dist<1,35km ou c_carro<R$0,21 :
carro	idade>24 ou n_autos>=1 :	t_ape<24min ou homem :	dist<2,7km ou c_carro<R$0,38 ou ñ_trabalha :	idade>54 :	dist>1,35km ou c_carro>R$0,21 : CARRO
carro	idade>24 ou n_autos>=1 :	t_ape<24min ou homem :	dist<2,7km ou c_carro<R$0,38 ou ñ_trabalha :	idade>54 :	dist<1,35km ou c_carro<R$0,21 :
carro	idade>24 ou n_autos>=1 :	t_ape<24min ou homem :	dist<2,7km ou c_carro<R$0,38 ou ñ_trabalha :	idade>54 :	dist>1,35km ou c_carro>R$0,21 : CARRO
carro	idade>24 ou n_autos>=1 :	t_ape>24min ou mulher :	n_salar<12sal-mín ou ñ_trabalha :	dist<3,0km ou c_carro<R$0,42 ou estuda :	n_autos>1 : CARRO
carro	idade>24 ou n_autos>=1 :	t_ape>24min ou mulher :	n_salar>11sal-mín ou trabalha :	n_autos<=1 :	idade<=40 : CARRO
carro	idade>24 ou n_autos>=1 :	t_ape>24min ou mulher :	n_salar>11sal-mín ou trabalha :	n_autos<=1 :	idade>40 :
carro	idade>24 ou n_autos>=1 :	t_ape>24min ou mulher :	n_salar>11sal-mín ou trabalha :	n_autos>1 : CARRO	
onibus	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	trabalha ou ñ_estuda :	idade<16 :
onibus	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	trabalha ou ñ_estuda :	idade<16 :
onibus	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	trabalha ou ñ_estuda :	idade>15 :
onibus	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	trabalha ou ñ_estuda :	idade>15 :
onibus	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	trabalha ou ñ_estuda :	idade>15 :
onibus	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	trabalha ou ñ_estuda :	idade>15 :
onibus	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	ñ_trabalha ou estuda :	c_bus<=R$0,45(meia tarifa c/ desconto) :
onibus	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	ñ_trabalha ou estuda :	c_bus<=R$0,45(meia tarifa c/ desconto) :
onibus	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos<=1 :	ñ_trabalha ou estuda :	c_bus>=R$0,65(uma tarifa) :
onibus	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos>1 :	estuda ou ñ_trabalha :	idade<=13 : ONIBUS
onibus	idade<24 ou n_autos=0 :	t_ape>33min ou dist>1,6km ou c_carro>R$0,24 :	n_autos>1 :	estuda ou ñ_trabalha :	idade>13 :
onibus	idade>24 ou n_autos>=1 :	t_ape>24min ou mulher :	n_salar<12sal-mín ou ñ_trabalha :	dist<3,0km ou c_carro<R$0,42 ou estuda :	n_autos<=1 :
onibus	idade>24 ou n_autos>=1 :	t_ape>24min ou mulher :	n_salar<12sal-mín ou ñ_trabalha :	dist<3,0km ou c_carro<R$0,42 ou estuda :	n_autos<=1 :
onibus	idade>24 ou n_autos>=1 :	t_ape>24min ou mulher :	n_salar<12sal-mín ou ñ_trabalha :	dist>3,0km ou c_carro>R$0,42 ou ñ_estuda : ONIBUS	
onibus	idade>24 ou n_autos>=1 :	t_ape>24min ou mulher :	n_salar>11sal-mín ou trabalha :	n_autos<=1 :	idade>40 :

Obs: A planilha B1 CONTINUA na próxima página.

idade 46 apresentadas conforme ordenamento por modo.

	trabalha ou ñ_estuda :	idade<16 :	estuda :	c_bus>R$0,65(uma tarifa) : A_PE
	trabalha ou ñ_estuda :	idade>15 :	t_bus<32min ou mulher :	t_ape<27min : dist<1,85km ou c_carro<R$0,27 : A_PE
	ñ_trabalha ou estuda :	c_bus<=R$0,45(meia tarifa c/ desconto) :	t_ape<48min :	dist<2,8km ou c_carro<R$0,38 : A_PE
	ñ_trabalha ou estuda :	c_bus>=R$0,65(uma tarifa) :	n_idade <= 0.175 : A_PE	
<R$0,38 ou ñ_trabalha :	idade<=54 :	n_autos<=1 :	dist<1,0km ou c_carro<R$0,17 : A_PE	
<R$0,38 ou ñ_trabalha :	idade>54 :	dist<1,35km ou c_carro<R$0,21 :	t_ape<12min : A_PE	
<R$0,38 ou ñ_trabalha :	idade>54 :	dist<1,35km ou c_carro<R$0,21 :	t_ape>12min :	n_salar<30sal-mín ou c_bus=ñgratuito : A_PE
ñ_trabalha :	dist<3,0km ou c_carro<R$0,42 ou estuda :	n_autos<=1 :	estuda :	c_bus>R$1,00(duas tarifas) ou t_bus>44min : A_PE
	trabalha ou ñ_estuda :	idade>15 :	t_bus>32min ou homem :	n_salar>11sal-mín : c_bus>=R$0,65(uma tarifa) : CARRO
	ñ_estuda ou trabalha : CARRO			
	estuda ou ñ_trabalha :	idade>13 :	n_autos>2 : CARRO	
>R$0,38 ou trabalha : CARRO				
<R$0,38 ou ñ_trabalha :	idade<=54 :	n_autos<=1 :	dist>1,0km ou c_carro>R$0,16 : CARRO	
<R$0,38 ou ñ_trabalha :	idade<=54 :	n_autos>1 : CARRO		
<R$0,38 ou ñ_trabalha :	idade>54 :	dist<1,35km ou c_carro<R$0,21 :	t_ape>12min :	n_salar>29sal-mín ou c_bus=gratuito : CARRO
<R$0,38 ou ñ_trabalha :	idade>54 :	dist>1,35km ou c_carro>R$0,21 : CARRO		
ñ_trabalha :	dist<3,0km ou c_carro<R$0,42 ou estuda :	n_autos>1 : CARRO		
trabalha :	n_autos<=1 :	idade<=40 : CARRO		
trabalha :	n_autos<=1 :	idade>40 :	dist<4,8km ou c_carro<R$0,64 : CARRO	
trabalha :	n_autos>1 : CARRO			
	trabalha ou ñ_estuda :	idade<16 :	estuda :	c_bus<R$0,65(uma tarifa) : ONIBUS
	trabalha ou ñ_estuda :	idade<16 :	ñ_estuda : ONIBUS	
	trabalha ou ñ_estuda :	idade>15 :	t_bus<32min ou mulher :	t_ape<27min : dist>1,85km ou c_carro>R$0,27 : ONIBUS
	trabalha ou ñ_estuda :	idade>15 :	t_bus<32min ou mulher :	t_ape>27min : ONIBUS
	trabalha ou ñ_estuda :	idade>15 :	t_bus>32min ou homem :	n_salar<=11sal-mín : ONIBUS
	trabalha ou ñ_estuda :	idade>15 :	t_bus>32min ou homem :	n_salar>11sal-mín : c_bus<=R$0,45(meia tarifa c/ desconto) : ONIBUS
	ñ_trabalha ou estuda :	c_bus<=R$0,45(meia tarifa c/ desconto) :	t_ape<48min :	dist>2,7km ou c_carro>R$0,385 : ONIBUS
	ñ_trabalha ou estuda :	c_bus<=R$0,45(meia tarifa c/ desconto) :	t_ape>48min : ONIBUS	
	ñ_trabalha ou estuda :	c_bus>=R$0,65(uma tarifa) :	idade>18 : ONIBUS	
	estuda ou ñ_trabalha :	idade<=13 : ONIBUS		
	estuda ou ñ_trabalha :	idade>13 :	n_autos<=2 : ONIBUS	
ñ_trabalha :	dist<3,0km ou c_carro<R$0,42 ou estuda :	n_autos<=1 :	estuda :	c_bus<R$1,00(duas tarifas), t_bus < 0.2585 : ONIBUS
ñ_trabalha :	dist<3,0km ou c_carro<R$0,42 ou estuda :	n_autos<=1 :	ñ_estuda : ONIBUS	
ñ_trabalha :	dist>3,0km ou c_carro>R$0,42 ou ñ_estuda : ONIBUS			
trabalha :	n_autos<=1 :	idade>40 :	dist>4,8km ou c_carro>R$0,64 : ONIBUS	

Obs: a planilha é CONTINUAÇÃO da página anterior.

QUADRO DE DESEMPENHO DA ARVORE				
Profundidade 15		Trein.	Valid.	Teste
% Fidelidade à MLP-11-10-5-3		83,2	84,7	82,1
% Acerto à Classific.	Médio entre modos	77,1	74,7	72,5
	Carro	89,2	82,7	80,4
	A pé	78,6	72,6	76,6
	Ônibus	63,2	69,1	61,1

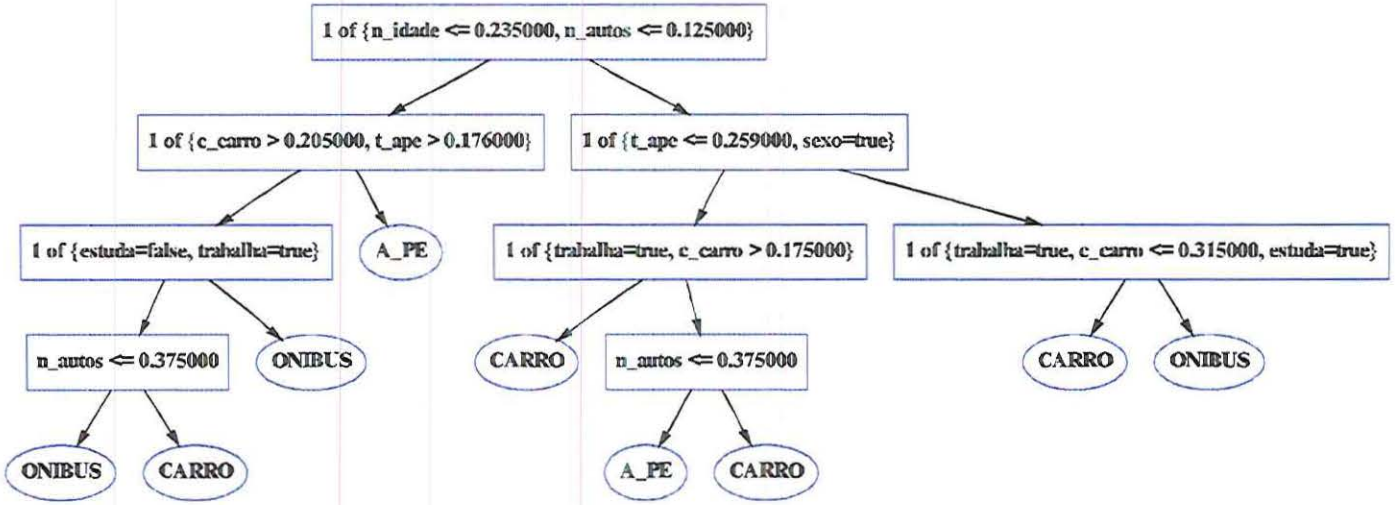


Figura B.2: Árvore extraída da rede 11-10-5-3 com profundidade 15.

Planilha B.2: Condições para os modos de viagem da árvore extraída da rede 11-10-5-3 com profundidade 15 apresentadas conforme ordenamento por modo.

a_pe	idade<24 ou n_autos=0 :	dist<2,1km ou c_carro<R$0,30 ou t_ape<31min :	A_PE	
a_pe	idade>24 ou n_autos>=1 :	t_ape<45min ou homem :	ñ_trabalha ou dist<1,75km ou c_carro<R$0,26 :	n_autos<=1 : A_PE
carro	idade<24 ou n_autos=0 :	dist>2,1km ou c_carro>R$0,30 ou t_ape>31min :	ñ_estuda ou trabalha :	n_autos>1 : CARRO
carro	idade>24 ou n_autos>=1 :	t_ape<45min ou homem :	trabalha ou dist>1,75km ou c_carro>R$0,26 : CARRO	
carro	idade>24 ou n_autos>=1 :	t_ape<45min ou homem :	ñ_trabalha ou dist<1,75km ou c_carro<R$0,26 :	n_autos>1 : CARRO
carro	idade>24 ou n_autos>=1 :	t_ape>44min ou homem :	trabalha ou dist<3,55km ou c_carro<R$0,48 ou estuda : CARRO	
onibus	idade<24 ou n_autos=0 :	dist>2,1km ou c_carro>R$0,30 ou t_ape>31min :	ñ_estuda ou trabalha :	n_autos<=1 : ONIBUS
onibus	idade<24 ou n_autos=0 :	dist>2,1km ou c_carro>R$0,30 ou t_ape>31min :	estuda ou ñ_trabalha : ONIBUS	
onibus	idade>24 ou n_autos>=1 :	t_ape>44min ou homem :	ñ_trabalha ou dist>3,55km ou c_carro>R$0,48 ou ñ_estuda : ONIBUS	

QUADRO DE DESEMPENHO DA ÁRVORE				
Profundidade 22		Trein.	Valid.	Teste
% Fidelidade à MLP-11-15-15-3		89,9	87,0	89,3
% Acerto à Classific.	Médio entre modos	73,8	77,0	73,0
	Carro	74,6	78,5	69,3
	A pé	70,6	72,3	76,8
	Ônibus	76,2	79,3	72,5

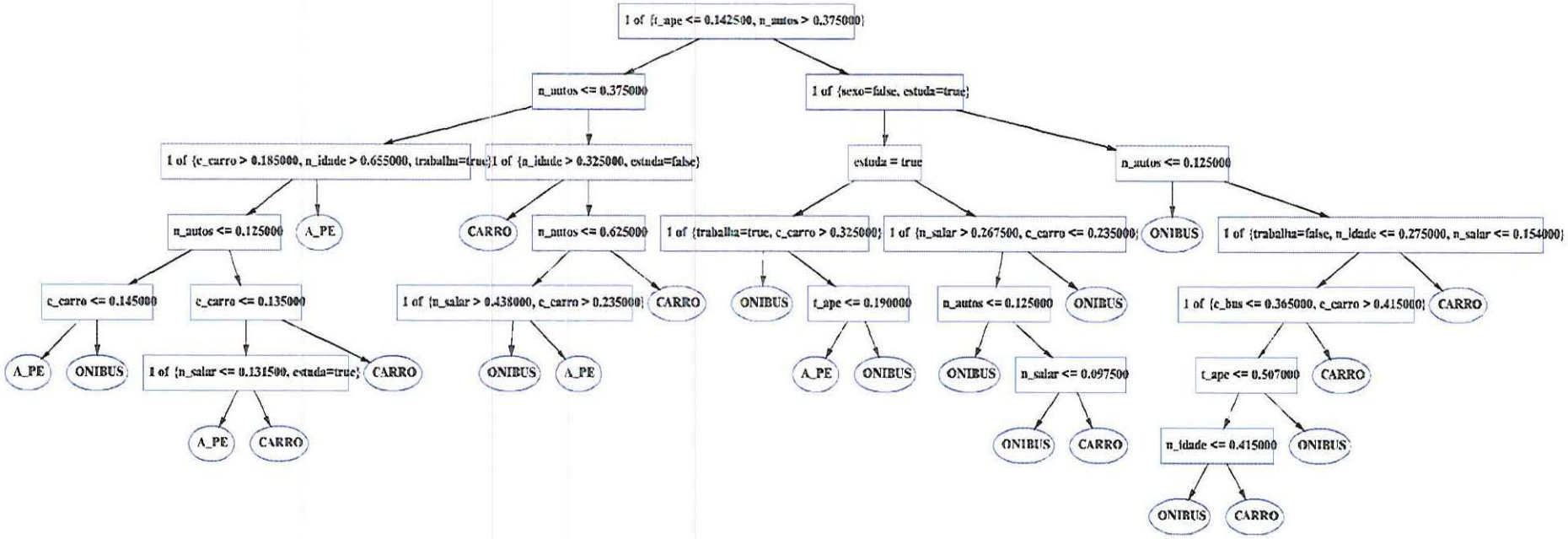


Figura B.3: Árvore extraída da rede 11-15-15-3 com profundidade 22.

Planilha B.3: Condições para os modos de viagem da árvore extraída da rede 11-15-15-3 com profundidade 22 apresentadas conforme ordenamento por modo.

a_pe	t_ape<25min ou n_autos>1 :	n_autos<=1 :	dist>1,85km ou c_carro>R$0,27 ou idade>58 ou trabalha :	n_autos=0 :	dist<1,35
a_pe	t_ape<25min ou n_autos>1 :	n_autos<=1 :	dist>1,85km ou c_carro>R$0,27 ou idade>58 ou trabalha :	n_autos>0 :	dist<1,2k
a_pe	t_ape<25min ou n_autos>1 :	n_autos<=1 :	dist<1,85km ou c_carro<R$0,27 ou idade<58 ou ñ_trabalha : A_PE		
a_pe	t_ape<25min ou n_autos>1 :	n_autos>1 :	idade<31 ou estuda :	n_autos<=2 :	n_salar<
a_pe	t_ape>25min ou n_autos<=1:	mulher ou estuda :	estuda :	ñ_trabalha ou dist<3,6km ou c_carro<R$0,49 :	t_ape<33
carro	t_ape<25min ou n_autos>1 :	n_autos<=1 :	dist>1,85km ou c_carro>R$0,27 ou idade>58 ou trabalha :	n_autos>0 :	dist<1,2k
carro	t_ape<25min ou n_autos>1 :	n_autos<=1 :	dist>1,85km ou c_carro>R$0,27 ou idade>58 ou trabalha :	n_autos>0 :	dist>1,2k
carro	t_ape<25min ou n_autos>1 :	n_autos>1 :	idade>31min ou ñ_estuda : CARRO		
carro	t_ape<25min ou n_autos>1 :	n_autos>1 :	idade<31 ou estuda :	n_autos>2 : CARRO	
carro	t_ape>25min ou n_autos<=1:	mulher ou estuda :	ñ_estuda :	n_salar>25sal-mín ou dist<2,5km ou c_carro<R$0,35 :	n_autos>
carro	t_ape>25min ou n_autos<=1:	homem, ñ_estuda :	n_autos>0 :	ñ_trabalha ou idade<=27 ou n_salar<15sal-mín :	c_bus<=I
carro	t_ape>25min ou n_autos<=1:	homem, ñ_estuda :	n_autos>0 :	ñ_trabalha ou idade<=27 ou n_salar<15sal-mín :	c_bus>=I
carro	t_ape>25min ou n_autos<=1:	homem, ñ_estuda :	n_autos>0 :	trabalha ou idade>27 ou n_salar>15sal-mín : CARRO	
onibus	t_ape<25min ou n_autos>1 :	n_autos<=1 :	dist>1,85km ou c_carro>R$0,27 ou idade>58 ou trabalha :	n_autos=0 :	dist>1,35
onibus	t_ape<25min ou n_autos>1 :	n_autos>1 :	idade<31 ou estuda :	n_autos<=2 :	n_salar>
onibus	t_ape>25min ou n_autos<=1:	mulher ou estuda :	estuda :	trabalha ou dist>3,6km ou c_carro>R$0,49 : ONIBUS	
onibus	t_ape>25min ou n_autos<=1:	mulher ou estuda :	estuda :	ñ_trabalha ou dist<3,6km ou c_carro<R$0,49 :	t_ape>33
onibus	t_ape>25min ou n_autos<=1:	mulher ou estuda :	ñ_estuda :	n_salar>25sal-mín ou dist<2,5km ou c_carro<R$0,35 :	n_autos=
onibus	t_ape>25min ou n_autos<=1:	mulher ou estuda :	ñ_estuda :	n_salar>25sal-mín ou dist<2,5km ou c_carro<R$0,35 :	n_autos>
onibus	t_ape>25min ou n_autos<=1:	mulher ou estuda :	ñ_estuda :	n_salar<25sal-mín ou dist>2,5km ou c_carro>R$0,35 : ONIBUS	
onibus	t_ape>25min ou n_autos<=1:	homem, ñ_estuda :	n_autos=0 : ONIBUS		
onibus	t_ape>25min ou n_autos<=1:	homem, ñ_estuda :	n_autos>0 :	ñ_trabalha ou idade<=27 ou n_salar<15sal-mín :	c_bus<=I
onibus	t_ape>25min ou n_autos<=1:	homem, ñ_estuda :	n_autos>0 :	ñ_trabalha ou idade<=27 ou n_salar<15sal-mín :	c_bus<=I

Obs: a planilha B3 CONTINUA na próxima página.

15-15-3 com profundidade 22 apresentadas conforme ordenamento por modo.

27 ou idade>58 ou trabalha :	n_autos=0 :	dist<1,35km ou c_carro<R$0,21 : A_PE	
27 ou idade>58 ou trabalha :	n_autos>0 :	dist<1,2km ou c_carro<R$0,20 :	n_salar<13sal-mín ou estuda : A_PE
27 ou idade<58 ou ñ_trabalha : A_PE			
	n_autos<=2 :	n_salar<40sal-mín ou dist<2,5km ou c_carro<R$0,35 : A_PE	
	ñ_trabalha ou dist<3,6km ou c_carro<R$0,49 :	t_ape<33min : A_PE	
27 ou idade>58 ou trabalha :	n_autos>0 :	dist<1,2km ou c_carro<R$0,20 :	n_salar>13sal-mín ou ñ_estuda : CARRO
27 ou idade>58 ou trabalha :	n_autos>0 :	dist>1,2km ou c_carro>R$0,20 : CARRO	
ARRO			
	n_autos>2 : CARRO		
	n_salar>25sal-mín ou dist<2,5km ou c_carro<R$0,35 :	n_autos>0 :	n_salar>9sal-mín : CARRO
	ñ_trabalha ou idade<=27 ou n_salar<15sal-mín :	c_bus<=R$0,45(meia tarifa c/ desconto) ou dist>4,8km ou c_carro>R$0,6:	t_ape<85min : idade>38 : CARRO
	ñ_trabalha ou idade<=27 ou n_salar<15sal-mín :	c_bus>=R$0,65(uma tarifa) ou dist<4,8km ou c_carro<R$0,63 : CARRO	
	trabalha ou idade>27 ou n_salar>15sal-mín : CARRO		
27 ou idade>58 ou trabalha :	n_autos=0 :	dist>1,35km ou c_carro>R$0,21 : ONIBUS	
	n_autos<=2 :	n_salar>40sal-mín ou dist>2,5km ou c_carro>R$0,35 : ONIBUS	
	trabalha ou dist>3,6km ou c_carro>R$0,49 : ONIBUS		
	ñ_trabalha ou dist<3,6km ou c_carro<R$0,49 :	t_ape>33min : ONIBUS	
	n_salar>25sal-mín ou dist<2,5km ou c_carro<R$0,35 :	n_autos=0 : ONIBUS	
	n_salar>25sal-mín ou dist<2,5km ou c_carro<R$0,35 :	n_autos>0 :	n_salar<=9sal-mín : ONIBUS
	n_salar<25sal-mín ou dist>2,5km ou c_carro>R$0,35 :	ONIBUS	
	ñ_trabalha ou idade<=27 ou n_salar<15sal-mín :	c_bus<=R$0,45(meia tarifa c/ desconto) ou dist>4,8km ou c_carro>R$0,6:	t_ape<85min : idade<=38 : ONIBUS
	ñ_trabalha ou idade<=27 ou n_salar<15sal-mín :	c_bus<=R$0,45(meia tarifa c/ desconto) ou dist>4,8km ou c_carro>R$0,6:	t_ape>85min : ONIBUS

Obs: esta planilha é CONTINUAÇÃO da página anterior.

ANEXO C

Método aplicado para a separação entre os modos bicicleta e caminhada incluídos no conjunto de dados referentes ao modo 3 da pesquisa domiciliar

Um critério para se obter a distinção do modo bicicleta do modo caminhada seria através da velocidade desenvolvida na viagem. É bem sabido que o indivíduo ao caminhar possui em média velocidade aproximada de 5 km/h, o que levaria a classificar viagens por modo 3 acima desta velocidade como sendo realizadas por bicicleta. Sendo assim, as hipóteses para direcionamento deste método poderiam ser: o indivíduo viajando por modo 3 até 5 km/h está caminhando e o indivíduo que viaja, acima de 5 km/h também por modo 3, o faz por modo bicicleta. Contudo, a velocidade não é uma informação colhida pela pesquisa domiciliar, ou seja , não foi medida de forma direta. Assim, é necessário obtê-la de forma indireta através da distância percorrida e do tempo gasto revelado pelo entrevistado. Porém, como discutido a seguir, as imprecisões nas medidas de distância e tempo dificultam a adoção de uma única hipótese, a respeito dos 5km/h de velocidade, para a separação dos modos.

A distância percorrida, para o caso do modo 3, foi obtida utilizando-se uma ferramenta do software de SIG que a partir dos endereços de origem e destino encontra o menor caminho possível, considerando a rede viária existente, e calcula o valor da distância. Essa ferramenta demonstra ser apropriada para o cálculo de distâncias para este modo, uma vez que, usualmente, o indivíduo ao locomover-se de um ponto ao outro na cidade tanto por caminhada quanto por bicicleta, otimiza o trajeto, minimizando a distância percorrida, pelo fato de não estar sujeito a restrições de trânsito como o sentido obrigatório de vias, por exemplo. Esse valor de distância está sujeito a um erro de 100 metros pelo fato da localização dos endereços, no SIG, ser feita pelo cruzamento de ruas mais próximo e como as quadras da cidade possuem aproximadamente 100 metros de lado, atribui-se 50 metros de erro tanto na localização do endereço da origem quanto do destino.

O tempo de viagem foi coletado na pesquisa domiciliar através da hora de saída na origem e da hora de chegada no destino, ambas declaradas pelo entrevistado, referindo-se a viagens realizadas no dia anterior. Este tempo de

viagem, revelado pelo entrevistado, é uma medida imprecisa pois está sujeita à percepção de cada indivíduo. Embora o tempo de viagem possa ser medido em um relógio, um indivíduo quando inquerido sobre quanto tempo gastou ao fazer uma viagem freqüentemente julga erroneamente o real tempo gasto, como verificou WATSON (1974). Curiosamente, tal erro individual de julgamento também foi percebido para o julgamento da distância real percorrida num estudo conduzido por CADWALLADER (1975). Assim sendo, a velocidade calculada a partir desse tempo de viagem declarado, está, como observado, sujeita à imprecisão que impossibilita a adoção desta como critério único na distinção dos viajantes por modo bicicleta dentre os pertencentes ao modo 3.

Na tentativa de se contornar esse problema de imprecisão buscou-se adicionar, neste método, outros fatores além da velocidade, que pudessem auxiliar a divisão entre os modos não motorizados. ARRUDA (2000) e FHWA (1992) fazem um apanhado de estudos sobre fatores dados como influenciadores na opção pelos modos de transporte não motorizados, alistando e discutindo cada um destes fatores. Mediante os vários fatores, ou atributos, alistados por estes autores e os disponíveis da pesquisa domiciliar, decidiu-se por incluir o motivo da viagem e a distância percorrida como fatores, tornando o método agora baseado na análise de três dimensões: a velocidade, a distância e o motivo.

Com relação à distância de viagem, estudos do FHWA (1992) identificaram a distância aproximada de 3,0 km, que pode ser percorrida a pé em cerca de 30 a 40 minutos, como o limite aceitável para caminhadas diárias. Para THAMIZ *et al.* (1996) a distância máxima de caminhada admitida, para indivíduos que não possuem automóvel, varia entre 1,3 e 2,5 km, dependendo de suas características sócio-econômicas e da localização de suas residências em relação ao centro da cidade. No estudo de ARRUDA (2000), conduzido na cidade de São Carlos no interior do estado de São Paulo e com características semelhantes à cidade de Bauru, foi coletada uma amostra de 768 viagens, todas de caráter utilitário (não foram consideradas as viagens realizadas por modo a pé pelo motivo exercício físico), com a distribuição de viagens por modo a pé em relação ao comprimento da viagem mostrada na figura C.1.

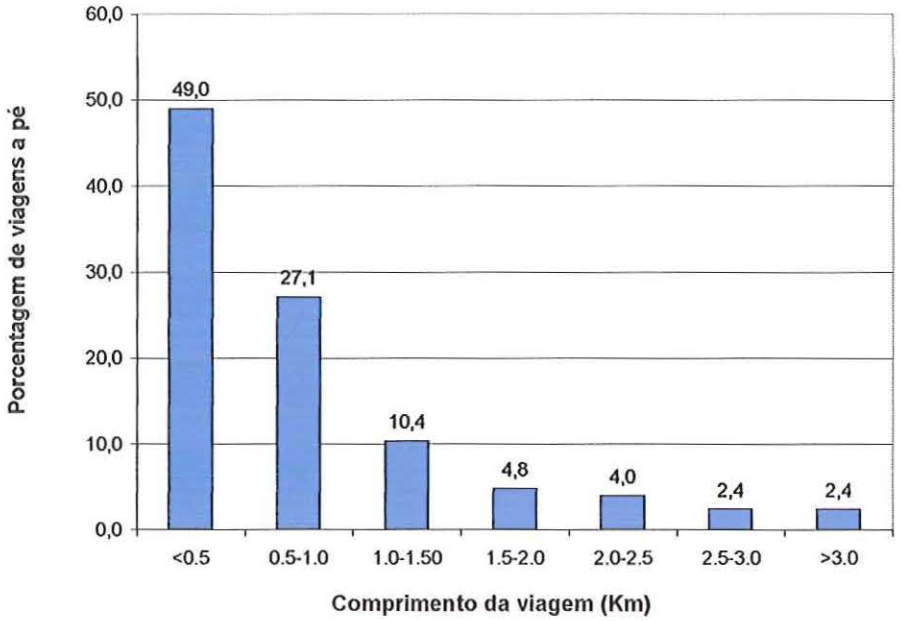


Figura C.1: Viagens a pé X Comprimento da viagem. Fonte: ARRUDA (2000).

Após a calibração de um modelo Logit multinomial ARRUDA (2000) obteve a variação da probablilidade de escolha do modo a pé em função do comprimento da viagem mostrada na figura C.2.

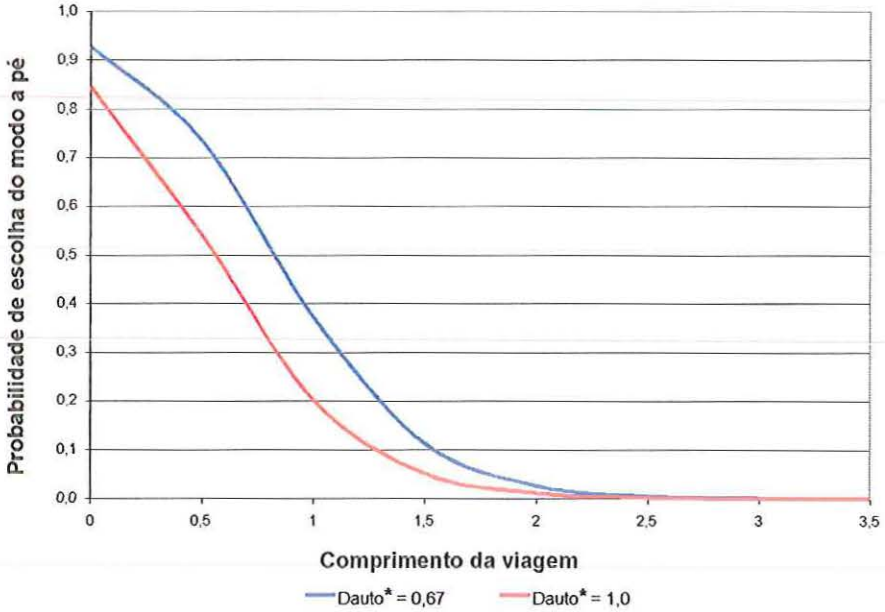


Figura C.2: Probabilidade de escolha modo a pé X Comprimento da viagem. Fonte: ARRUDA (2000).

* Disponibilidade do automóvel no domicílio.

A figura C.3 mostra como varia a probabilidade de escolha pelo modo a pé em função da distância, em cidades brasileiras de porte médio, Bauru inclusa entre as cidades desse porte

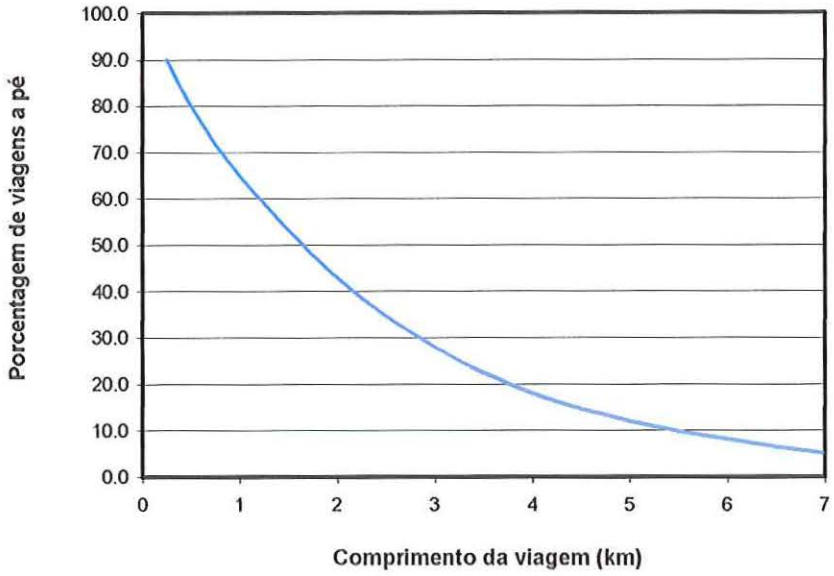


Figura C.3: Probabilidade de escolha do modo a pé em função da distância.
Fonte: SANCHES (1988).

De maneira geral, os estudos mencionados indicam que poucas pessoas caminham mais que 2 km, o que indica a relevância da consideração da distância neste método.

O motivo de viagem é um atributo que foi diretamente colhido na entrevista e consta nos dados disponíveis da pesquisa domiciliar. A consideração do motivo de viagem pode ser esclarecedora em alguns casos. Dadas as características inerentes do modo, provavelmente poucas pessoas viajariam de bicicleta por motivo de saúde ou de compras. Estudos do FHWA (1992) suportam o fato de que apenas uma minoria usa o modo bicicleta para viagens por motivos regulares do dia a dia, sendo o menor percentual observado para motivo ESCOLA, 2,8%[1] O motivo de viagem majoritário para viagens por bicicleta, apontado pelo FHWA (1992) é o motivo LAZER[2] , seguido pelo motivo OUTROS[3].

[1] Referente ao estudo de *Portland* contido no FHWA (1992).
[2] Traduzido do original "*Recreation*".
[3] Traduzido do original "*Utility*" lá definido como qualquer viagem por bicicleta como meio de transporte exceto por prazer e pelos demais motivos alternativos a este.

Definidos então os atributos, a velocidade de viagem, a distância percorrida e o motivo da viagem, foi feito um cruzamento destas informações, com o auxílio de um software de estatística. Os atributos distância e velocidade foram divididos em categorias e cruzados com cada um dos motivos de viagem existentes, gerando assim as seguintes tabelas de freqüências:

Tabela C.1. Tabela distância X velocidade para o motivo de viagem "trabalho".

Motivo TRABALHO								
Distância / Velocidade	<0,5km	0,5-1km	1-1,5km	1,5-3km	3-5km	>5km	Subtotal	Subtotal p/ Bicicleta
<5km/h	134	108	51	58	7	2	360	
5-7km/h	12	21	8	29	18	1	89	
7-8km/h	1	9	6	12	9	3	40	
8-9km/h	1	12	6	5	1	0	25	
9-10km/h	0	3	0	8	3	1	15	12
>10km/h	3	9	9	41	23	42	127	106
							Total	118

Tabela C.2. Tabela distância X velocidade para o motivo de viagem "escola".

Motivo ESCOLA								
Distância / Velocidade	<0,5km	0,5-1km	1-1,5km	1,5-3km	3-5km	>5km	Subtotal	Subtotal p/ Bicicleta
<5km/h	277	219	66	39	0	0	601	
5-7km/h	21	48	44	28	5	0	146	
7-8km/h	1	9	13	12	8	0	43	
8-9km/h	0	7	19	9	5	0	40	
9-10km/h	0	7	0	12	2	0	21	2
>10km/h	0	27	34	47	78	56	242	134
							Total	136

Tabela C.3. Tabela distância X velocidade para o motivo de viagem "compras".

Motivo COMPRAS								
Distância / Velocidade	<0,5km	0,5-1km	1-1,5km	1,5-3km	3-5km	>5km	Subtotal	Subtotal p/ Bicicleta
<5km/h	73	54	24	12	0	0	163	
5-7km/h	4	18	6	5	4	0	37	
7-8km/h	0	7	1	4	0	0	12	
8-9km/h	0	6	2	2	0	1	11	
9-10km/h	0	1	0	3	1	0	5	4
>10km/h	2	6	20	22	8	13	71	43
							Total	47

Tabela C.4. Tabela distância X velocidade para o motivo de viagem "saúde".

Motivo SAÚDE								
Distância / Velocidade	<0,5km	0,5-1km	1-1,5km	1,5-3km	3-5km	>5km	Subtotal	Subtotal p/ Bicicleta
<5km/h	10	16	1	2	1	0	30	
5-7km/h	0	2	1	3	0	0	6	
7-8km/h	0	3	0	1	0	0	4	
8-9km/h	0	0	3	1	0	0	4	
9-10km/h	0	4	0	0	0	0	4	0
>10km/h	0	2	0	4	1	1	8	6
							Total	6

Tabela C.5. Tabela distância X velocidade para o motivo de viagem "lazer".

Motivo LAZER								
Distância / Velocidade	<0,5km	0,5-1km	1-1,5km	1,5-3km	3-5km	>5km	Subtotal	Subtotal p/ Bicicleta
<5km/h	47	48	32	19	1	0	147	
5-7km/h	5	8	4	6	0	0	23	
7-8km/h	1	6	8	5	0	0	20	
8-9km/h	1	4	3	3	0	1	12	
9-10km/h	1	1	0	1	1	0	4	2
>10km/h	4	6	8	15	3	13	49	39
							Total	41

Tabela C.6. Tabela distância X velocidade para o motivo de viagem "outros".

Motivo OUTROS								
Distância / Velocidade	<0,5km	0,5-1km	1-1,5km	1,5-3km	3-5km	>5km	Subtotal	Subtotal p/ Bicicleta
<5km/h	55	40	26	16	6	0	143	
5-7km/h	5	13	6	17	2	1	44	
7-8km/h	0	5	4	3	1	0	13	
8-9km/h	0	1	3	4	0	0	8	
9-10km/h	1	0	0	7	2	2	12	11
>10km/h	1	7	17	19	20	27	91	66
							Total	77

Da análise dessas tabelas de freqüências pôde-se constatar a existência de uma linha diagonal de elementos nulos ou pequenos, em algumas tabelas pouco e em outras bem definida, que separa duas concentrações de freqüência: uma no canto superior esquerdo (provável modo caminhada) e outra no canto inferior direito (provável modo bicicleta). Assumiu-se, então, essa linha como um possível "divisor de águas" entre os viajantes por bicicleta e por caminhada e passou-se a analisar cada um dos motivos de viagem à procura de critérios para a distinção dos prováveis viajantes por bicicleta.

Um critério adotado foi o de não considerar a tabela de cruzamento do motivo ESCOLA para fins de distinção do modo bicicleta, pelo fato de não haver no Brasil o costume por parte dos estudantes de locomoverem-se até as escolas, para fins de estudo, por modo bicicleta, sendo poucas as escolas brasileiras providas de local destinado ao estacionamento de bicicletas durante o horário letivo. Ainda observa-se para o motivo ESCOLA que a concentração no canto inferior direito, da respectiva tabela é a mais alta dentre todas as tabelas, ou motivos. Atribui-se isso ao fato do entrevistado ser criança ou adolescente e provavelmente mais impreciso em seu relato de tempo gasto na viagem, bem como à interpretação errônea do real motivo da viagem realizada, uma vez que, por exemplo, uma viagem para atividade de lazer realizada nas dependências da escola fora do horário de aula não caracteriza o motivo ESCOLA.

Para os motivos de viagem TRABALHO, COMPRAS, SAÚDE e OUTROS, optou-se por um critério comum, com limites mínimos para velocidade e distância, que definissem o provável conjunto das viagens realizadas por bicicleta: velocidade acima de 9 km/h e distância acima e 1,5 km. No caso das viagens realizadas por motivo LAZER, acredita-se que o modo bicicleta possa estar sendo escolhido para distâncias um pouco mais curtas, e assim adotou-se um menor limite mínimo, ou seja, distância a partir de 1km, e manteve-se o limite de velocidade de 9 km/h.

Esses critérios formaram um filtro que foi aplicado no conjunto de viagens por modo 3 (2.670 viagens), identificando 289 prováveis viagens por modo bicicleta as quais foram excluídas deste conjunto. Essas 289 viagens por bicicleta representam 3,2% de todas as viagens coletadas em Bauru e disponíveis para este trabalho, 8984 viagens. Um estudo realizado por GEIPOT (1985) na cidade de Bauru estimou, naquela época, que cerca de 2% da população estaria utilizando o modo "ciclo" que incluía bicicletas e ciclomotores. Embora o método aqui aplicado tenha identificado um percentual maior de prováveis viajantes por modo bicicleta que o estudo mais antigo realizado pelo GEIPOT, e considerando que a inclusão de outros fatores como a idade e o sexo poderiam talvez refinar o método, admite-se que este método tenha atingido seu objetivo de apenas eliminar distorções grosseiras no conjunto de dados do modo 3, e não de vir a ser um método de aplicação genérica para distinção entre o modo caminhada e o modo bicicleta.

ANEXO D:

Atributos de entrada utilizados nos experimentos

Tabela D.1: Atributos de entrada do primeiro experimento da etapa preliminar.

▪Idade do viajante	▪Motivo de viagem
▪Sexo do viajante	▪Distância de caminho mínimo
▪Situação familiar	▪Tempo de viagem por modo caminhada
▪Grau de instrução	▪Tempo de viagem por modo automóvel
▪Carteira de motorista	▪Tempo de viagem por modo ônibus
▪Tamanho da família	▪Custo de viagem por modo automóvel
▪Renda domiciliar	▪Custo de viagem por modo ônibus
▪Número de automóveis no domicílio	▪ Tarifa dupla

Tabela D.2: Atributos de entrada do segundo experimento da etapa preliminar.

▪Idade do viajante	▪Motivo de viagem
▪Sexo do viajante	▪Distância de caminho mínimo
▪Situação familiar	▪Tempo de viagem por modo caminhada
▪Grau de instrução	▪Tempo de viagem por modo automóvel
▪Carteira de motorista	▪Tempo de viagem por modo ônibus
▪Tamanho da família	▪Custo de viagem por modo automóvel
▪Renda domiciliar	▪Custo de viagem por modo ônibus
▪Número de automóveis no domicílio	▪ Caronista

Tabela D.3: Atributos de entrada do terceiro experimento da etapa preliminar.

▪Idade do viajante	▪Tempo de viagem por modo caminhada
▪Sexo do viajante	▪Tempo de viagem por modo automóvel
▪Situação familiar	▪Tempo de viagem por modo ônibus
▪Grau de instrução	▪Custo de viagem por modo automóvel
▪Carteira de motorista	▪Custo de viagem por modo ônibus
▪Tamanho da família	▪Caronista
▪Renda domiciliar	
▪Número de automóveis no domicílio	

Tabela D.4: Atributos de entrada do quarto experimento da etapa preliminar.

▪Idade do viajante	▪Motivo de viagem
▪Sexo do viajante	▪Distância de caminho mínimo
▪Situação familiar	▪Tempo de viagem por modo caminhada
▪Grau de instrução	▪Tempo de viagem por modo automóvel
▪Carteira de motorista	▪Tempo de viagem por modo ônibus
▪Tamanho da família	▪Custo de viagem por modo automóvel
▪Renda domiciliar	▪Custo de viagem por modo ônibus
▪Número de automóveis no domicílio	▪Caronista

Tabela D.5: Atributos de entrada do quinto experimento da etapa preliminar.

▪Idade do viajante	▪Motivo de viagem
▪Sexo do viajante	▪Tempo de viagem por modo caminhada
▪Situação familiar	▪Tempo de viagem por modo automóvel
▪Grau de instrução	▪Tempo de viagem por modo ônibus
▪Carteira de motorista	▪Custo de viagem por modo automóvel
▪Tamanho da família	▪Custo de viagem por modo ônibus
▪Renda domiciliar	▪Caronista
▪Número de automóveis no domicílio	

Tabela D.6: Atributos de entrada do sexto experimento da etapa preliminar.

▪Idade do viajante	▪Motivo de viagem
▪Sexo do viajante	▪Tempo de viagem por modo caminhada
▪Grau de instrução	▪Tempo de viagem por modo automóvel
▪Carteira de motorista	▪Tempo de viagem por modo ônibus
▪Tamanho da família	▪Custo de viagem por modo automóvel
▪Renda domiciliar	▪Custo de viagem por modo ônibus
▪Número de automóveis no domicílio	▪Caronista

Tabela D.7: Atributos de entrada do sétimo experimento da etapa preliminar.

▪Idade do viajante	▪Tempo de viagem por modo caminhada
▪Sexo do viajante	▪Tempo de viagem por modo automóvel
▪Situação familiar	▪Tempo de viagem por modo ônibus
▪Grau de instrução	▪Custo de viagem por modo automóvel
▪Carteira de motorista	▪Custo de viagem por modo ônibus
▪Tamanho da família	▪Caronista
▪Renda domiciliar	▪Trabalha
▪Número de automóveis no domicílio	▪Estuda
▪Motivo de viagem	

Tabela D.8: Atributos de entrada do oitavo experimento da etapa preliminar.

▪Idade do viajante	▪Tempo de viagem por modo caminhada
▪Sexo do viajante	▪Tempo de viagem por modo automóvel
▪Grau de instrução	▪Tempo de viagem por modo ônibus
▪Carteira de motorista	▪Custo de viagem por modo automóvel
▪Tamanho da família	▪Custo de viagem por modo ônibus
▪Renda domiciliar	▪Trabalha
▪Número de automóveis no domicílio	▪Estuda

Tabela D.9: Atributos de entrada do primeiro experimento da etapa conclusiva.

▪Idade do viajante	▪Motivo de viagem
▪Sexo do viajante	▪Tempo de viagem por modo caminhada
▪Situação familiar	▪Tempo de viagem por modo automóvel
▪Grau de instrução	▪Tempo de viagem por modo ônibus
▪Carteira de motorista	▪Custo de viagem por modo automóvel
▪Tamanho da família	▪Custo de viagem por modo ônibus
▪Renda domiciliar	
▪Número de automóveis no domicílio	

Tabela D.10: Atributos de entrada do segundo experimento da etapa conclusiva.

▪Idade do viajante	▪Motivo de viagem
▪Sexo do viajante	▪Tempo de viagem por modo caminhada
▪Situação familiar	▪Tempo de viagem por modo automóvel
▪Grau de instrução	▪Tempo de viagem por modo ônibus
▪Tamanho da família	▪Custo de viagem por modo automóvel
▪Renda domiciliar	▪Custo de viagem por modo ônibus
▪Número de automóveis no domicílio	

Tabela D.11: Atributos de entrada do terceiro experimento da etapa conclusiva.

▪Idade do viajante	▪Tempo de viagem por modo caminhada
▪Sexo do viajante	▪Tempo de viagem por modo automóvel
▪Situação familiar	▪Tempo de viagem por modo ônibus
▪Grau de instrução	▪Custo de viagem por modo automóvel
▪Tamanho da família	▪Custo de viagem por modo ônibus
▪Renda domiciliar	▪Trabalha
▪Número de automóveis no domicílio	▪Estuda
▪Motivo de viagem	

Tabela D.12: Atributos de entrada do quarto experimento da etapa conclusiva.

▪Idade do viajante	▪Tempo de viagem por modo ônibus
▪Sexo do viajante	▪Custo de viagem por modo automóvel
▪Renda domiciliar	▪Custo de viagem por modo ônibus
▪Número de automóveis no domicílio	▪Trabalha
▪Tempo de viagem por modo caminhada	▪Estuda
▪Tempo de viagem por modo automóvel	

Tabela D.13: Atributos de entrada do quinto experimento da etapa conclusiva.

▪Idade do viajante
▪Sexo do viajante
▪Renda domiciliar
▪Número de automóveis no domicílio
▪Tempo de viagem por modo caminhada

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, F.C.; SIQUEIRA, J.O. (1997). Comparação entre regressão logística e redes neurais na falência de bancos brasileiros. In: CONGRESSO BRASILEIRO DE REDES NEURAIS, 3./ ESCOLA DE REDES NEURAIS, 4., Florianópolis. Anais. p.27-32.

ANDREWS, R.; DIEDERICH, J.; TICKLE, A.B. (1995). A Survey and critique of technique for extracting rules from trained artificial neural networks. Knowledge-Based Systems Journal, v.8, n.6, p.373-389.

ANDREWS, R.; GEVA, S. (1995). RULEX & CEBP networks as the basis for a rule refinement system. In: HALLAM, J., ed. Hybrid problems hybrid solutions. Amsterdam, IOS Press. p.1-12.

ANFAVEA - ASSOCIAÇÃO NACIONAL DOS FABRICANTES DE VEÍCULOS AUTOMOTORES (1999). Anuário estatístico da indústria automobilística brasileira. São Paulo, ANFAVEA.

APTÉ, C.; WEISS, S. (1997). Data mining with decision trees and decision rules. Future Generation Computer Systems, v.13, p.197-210.

ARRUDA, F.S. (2000). *Integração dos modos não motorizados nos modelos de planejamento dos transportes.* São Carlos. 94 p. Tese (Mestrado) - Universidade Federal de São Carlos.

BABCOCK, C. (1994). Parallel processing mines retail data: Wal-Mart use parallel processing to mine its retail data. Computer World, n.26, p 6.

BARNES, T.J.; SHEPPARD, E. (1992). Is there a place for the rational actor? A geographical critique of the rational choice paradigm. Economic Geograhy, v.68, p.1-21.

BEALE, R.; JACKSON, T. (1992). Neural computing: an introduction. Philadelphia, IOP.

BEN-AKIVA, M.E.; LERMAN, S.R. (1985). Discrete choice analysis: theory and application to travel demand. Cambridge, MIT Press.

BEN-AKIVA, M.E. et al. (1999). Extended framework for modeling choice behavior. Marketing Letters, v.10, p.11-44.

BEN-AKIVA, M.E.; MCFADDEN, D.; TRAIN, K. (1987). The demand for local telephone service: a fully discrete model of residential calling patterns and service choices. The Rand Journal of Economics, v.18, p.109-123.

BERKOVIC, J. (1985). New car sales and used car stocks: a model of the automobile market. Rand Journal of Economics, v.16, p.195-21.

BISHOP, C. (1995). Neural networks for pattern recognition. New York, Oxford Press.

BITTENCOURT, G. (1996). Inteligência artificial: ferramentas e teorias. Campinas, UNICAMP/ Escola de Computação.

BRACHMAN, R.J.; ANAND, T. (1996). The process of knowledge discovery in databases. In: FAYYAD, U.M. et al, eds. Advances in knowledge discovery and data mining. Menlo Park, AAAI Press. p.37-57.

BRAGA, J.L. et al. (1998). Extração de conhecimento: uma comparação entre os métodos clássico e conexionista. In: SIMPÓSIO BRASILEIRO DE REDES NEURAIS, 5., Belo Horizonte. Anais. p.126-131.

BREIMAN, L.; OLSHEN, R.; STONE, C. (1984). Classification and regression trees. Monterey, Wadsworth and Brooks.

BROMLEY, R.D.F.; THOMAS, C.J. (1993). The retail revolution, the carless shopper and disadvantage. Transactions of the Institute of British Geographers NS, v.18, p.222-236.

CABENA, P. et al. (1997). Discovering data mining: from concept to implementation. Upper Saddle River, NJ, Prentice Hall.

CADWALLADER, M. (1975). A behavioral model of consumer spatial decision making. Economic Geography, v.51, p.339-349.

CARVALHO, A.P.L.F.; BRAGA, A.P.; LUDERMIR, T.B. (1999). Redes neurais artificiais: teoria e aplicações. São Carlos, ICMC/USP. /Material didático do curso de Redes Neurais/

CLARK, P.; NIBLET, T. (1989). The CN2 induction algorithm. Machine Learning Journal, v.3, n.4 p.261-283.

CRAVEN, M.W. (1996). Extracting comprehensible models from trained neural networks. Madison. 199p. Tese (Doutorado) - University of Wisconsin.

CRAVEN, M.W.; SHAVLIK, J.W. (1996). Extracting tree-structured representations of trained networks. In: TOURETZKY, D.S.; MOZER, M.C.; HASSELMO, M.E., eds. Advances in neural information processing systems. Denver, MIT Press. v.8, p.37-45.

CRAVEN, M.W.; SHAVLIK, J.W. (1994). Using sampling and queries to extract rules from trained neural networks. INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 11., New Brunswick. Proceedings. p.37-45.

CRAVEN, M.W.; SHAVLIK, J.W. (1997). Using neural networks for data mining. Future Generation Computer Systems, v.13, special issue on data mining, p.211-229.

CYBENKO, G. (1988). Continuous valued neural networks with two hidden layers are sufficient. Medford, Tufts University/Department of Computer Science. /Technical Report/ .

CYBENKO, G. (1989). Approximation by superpositions of sigmoidal function. Mathematics of Control, Signals, and Systems, v.2, p.303-314.

DAGANZO, C.F. (1979). Multinomial probit: the theory and its applications to demand forecasting. New York, Academic Press.

DALVI, Q. (1978). Economic theories of travel choice. In: HENSHER, D.A.; DALVI, Q., eds. Determinants of Travel Choice. England, Saxon House. p.66-99.

DAPENA, A.R.; MEDINA, J.R. (2000). Empleo de redes neuronales para modelos de elección de opciones de transporte. In: CONGRESO DE INGENIERIA DEL TRANSPORTE, 4., Valencia. Anais. p.1-10.

DE CARVALHO, M.C.M. (1999). Previsão de demanda de transportes: uma abordagem não-linear usando redes neuronais. In: CONGRESSO DE PESQUISA E ENSINO EM TRANSPORTE, 13., São Carlos. Anais. v.1, p.187-200.

DOMENCICH, T.; MCFADDEN, D. (1975). Urban travel demand: a behavioral analysis. New York, North-Holland Publishing.

DOMINGOS, B.; PAZZANI, M. (1996). Beyond independence: conditions for the optimality of the simple bayesian classifier. In: SAITTA, L., ed. Machine Learning: proceedings of thirteenth International Conference. p.105-112.

DOUGHERTY, M.S. (1997). *Guest editorial: applications of neural networks in* transportation. Transportation Research C, v.5, n.5, p.255-257.

DOUGHERTY, M.S. (1995). A review of neural networks applied to transport. Transportation Research C, v.3, n.4, p.247-260.

EMDURB - EMPRESA MUNICIPAL DE DESENVOLVIMENTO URBANO E RURAL DE BAURU (2001). http://www.emdurb.com.br (10 Jul.).

FAGHRI, A.; ANEJA, S. (1996). Artificial neural network-based approach To modeling trip production. Transportation Research Record, n.1556, p.131-136.

FAMILI, A. et al. (1997). Data preprocessing and intelligent data analysis. Intelligent Data Analysis, v.1, n.1.

FAYYAD, U.M.; PIATESTKY-SHAPIRO, G.; SMYTH, P. (1996). From data mining to knowledge discovery: an overview. In: FAYYAD, U.M. et al., eds. Advances in knowledge discovery and data mining. Menlo Park, CA, AAAI Press. p.1-34.

FEDERAL HIGHWAY ADMINISTRATION (1992). Reasons why bicycling and walking are and are not being used more extensively as travel modes. /FHWA-PD-92-041/.

FISCHER, M.M.; NIJKAMP, P. (1987). From static toward dynamic discrete choice modelling. Regional Science and Urban Economics, v.7, p.3-27.

FOERSTER, J.F. (1979). Mode choice decision process models: a comparison of compensatory and non-compensatory structures. Transportation Research A, v.13, p.17-28.

FU, L. (1994). Rule generation from neural networks. IEEE Transactions on Systems, Man, and Cybernetics, v.24, n.8, p.1114-1123.

FURTADO, A.N.D. (1998). Uma nova abordagem na avaliação de projetos de transportes: o uso de redes neurais artificiais como técnica para avaliar e ordenar alternativas. São Carlos. 249 p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo.

GAINES, B.R. (1996). Transforming rules and trees. In: FAYYAD, U.M. et al, eds. Advances in knowledge discovery and data mining. Menlo Park, AAAI Press. p.205-226.

GARSON, D.G. (1991). Interpreting neural-network connection weights. AI Expert, n.6, v.4, p.47-51.

GAUDRY, M.J.I. (1981). The Inverse power transformation logit and dogit mode choice models. Transportation Research, v.15B, p.97-103.

GEIPOT (1985). Estudo da demanda de transportes urbanos: diagnóstico/1981, prognóstico/1985-1990. Brasília.

GOLOB, T.G.; RICHARDSON, A.J. (1981). Noncompensatory and discontinuous constructs in travel-behavior models. In: STOPHER, P.R.; MEYBURG, A.H.; BROG, W., eds. New horizons in travel-behavior research. Massachusetts, Lexington Books. p.369-384.

GOLOB, F.T.; MEURS, H. (1986). Biases in response over time in a seven-day travel diary. Transportation, v.13, p.163-181.

GONZÁLES-ESCRIVÁ, J.A.; SÁNCHEZ, T.R. (2000). Utilización de redes neuronales para ajustar modelos de elección modal en transporte interurbano de personas. In: CONGRESO DE INGENIERIA DEL TRANSPORTE, 4., Valencia. Anais. p.273-280.

GORDON, P.; KUMAR, A.; RICHARDSON, H.W. (1988). Beyond the journey to work. Transportion Research A, v.22, p.419-426.

HAGERSTRAND, T. (1970). What about people in regional science? Local, Papers of the Regional Science Association, v.24, p. 7-21.

HAND, D.J. (1998). Data mining: statistics and more? The American Statician, v.2, n.52, p.112-118.

HAYKIN, S. (1999). Neural networks – a comprehensive foundation. 2.ed. Upper Saddle River, NJ, Prentice Hall.

HERTZ, J.; KROGH, A.; PALMER, R.G. (1991). Introduction to the theory of neural computation. Addison-Wesley, Santa Fe Institute Studies in The Science of Complexity. v.1 /Lecture Notes/ .

HIMANEN, V. et al. (1998). Neural networks in transport: applications. Aldershot - Gower House, Ashgate.

HINTON, G. (1986). Learning distributed representations of concepts. In: ANNUAL CONFERENCE OF THE COGNITIVE SCIENCE SOCIETY, 8., Amherst. Proceedings. p.1-12.

HIRSCH, J. (2000). 2 U.S. economists win nobel. Los Angeles Times Newspaper, section C. http://emlab.berkeley.edu/users/mcfadden/index.html. (07 Dez.).

HOPFIELD, J.J. (1982). Neural networks and physical systems with emergent collective properties. Proceedings of Natonal Academy of Sciences, v.79, p.2554-2558.

HOROWITZ, J.L. (1985). *Travel and location behavior: state of the art and* research opportunities. Transportation Research A, v.19, n.5/6, p.41-53.

JAPKOWICZ, N. (2000). The class imbalance problem: significance and strategies. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, Las Vegas, 2000. Proceedings. p.111-117.

KAWAMOTO, E. (1987). Novo enfoque do processo de escolha em transporte com tratamento baseado na psicofísica multidimensional. São Carlos. 126 p. Tese (Doutorado) – Escola de Engenharia de São Carlos, Universidade de São Paulo.

KAWAMOTO, E. (1989). Processo semi-compensatório de escolha: um modelo probabilístico. Revista de Transporte e Tecnologia, v.1, v.2, p.9-31.

KAWAMOTO, E. (1990). Comparação de desempenho dos modelos de divisão modal: modelo logit multinomial X modelo semi-compensatório. In: CONGRESSO PANAMERICANO DE INGENIERIA DE TRANSITO Y TRANSPORTE, 6., Popayán. Anais. p.203-209.

KAWAMOTO, E. (1994). Modelo semi-compensatório de escolha modal em transporte de passageiros. São Carlos. 96p. Tese (Livre Docência) - Escola de Engenharia de São Carlos, Universidade de São Paulo.

KAWAMOTO, E.; SETTI, J.R. (1992). Procedure for the calibration of a semicompensatory mode choice model. Transportation Research Record, n.1357, p.66-71.

KITAMURA, R. (1997). Activity-based travel forecasting: what are some issues? Applications of models of activity behavior for activity based demand

forecasting. In: ACTIVITY-BASED TRAVEL FORESCASTING CONFERENCE, New Orleans. Proceedings.

KITAMURA, R.; BOVY, P. (1987). Analysis of attrition biases and trip reporting errors for panel data. Transportation Research A, v.21, p.287-302.

KLÖSGEN, W. (1996). Explora: a multipattern and multistrategy discovery assistant. In: FAYYAD, U.M. et al, eds. Advances in knowledge discovery and data mining. Menlo Park, AAAI Press. p.249-271.

KOVÁCS, Z.L. (1996). Redes neurais artificiais: fundamentos e aplicações. São Paulo, Edição Acadêmica.

KURANI, K.S.; LEE-GOSSELIN, M.E.H. (1997). Synthesis of past activity analysis applications. In: ENGELKE, L.J. ed. ACTIVITY-BASED TRAVEL FORECASTING CONFERENCE: summary, recommendations and compendium of papers. Texas Transportation Institute. Report of the travel model improvement program. Arlington, TX. p.51-78.

LANCASTER, K.J. (1966). A new approach to consumer theory. Journal of Political Economy, v.14, n.2, p.132-57.

LIMA, E. (1999). Bauru ganha dez mil carros novos por ano. Jornal da Cidade, Bauru, 18 abr. Caderno JC nos Bairros, p.2.

LIU, H.; MOTODA, H. (2000). Feature selection for knowledge discovery and data mining. Norwel, MA, Kluwer Academic. (Kluwer International Series in Engineering and Computer Science, 454).

LÓPEZ-REYES, D.E. (1999). Um Procedimento para peterminação de matriz origem-destino para diferentes modos: método indireto baseado em modelo de escolha discreta. São Carlos. 98 p. Tese (Doutorado) – Escola de Engenharia de São Carlos, Universidade de São Paulo.

LUCE, R.D. (1959). Individual choice behavior. New York, John Wiley.

MARBLE, D.; HANSON, P.; HANSON, S. (1972). Household travel behavior study: field operaions and questionnaires. Evanston, The Transportation Center at Northwestern University.

MARQUES, H.N. (1998). Um sistema de informações para usuários de transporte coletivo em cidades de médio porte. São Carlos. 89 p. Dissertação (Mestrado) - Escola de Engenharia de São Carlos, Universidade de São Paulo.

MARTINELI, E. (1999). Extração de conhecimento de redes neurais artificiais. São Carlos. 104 p. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação.

MCFADDEN, D. (2000). Disaggregate behavioral travel demand's RUM side: a 30-year retrospective. In: INTERNATIONAL ASSOCIATION FOR TRAVEL BEHAVIOR CONFERENCE, Queensland, 2000. http://emlab.berkeley.edu/users/mcfadden/index.html. (07 Dez.).

MCFADDEN, D. (1975). Aggregate travel demand forecasting from disaggrgated behavioral models. Transportation Research Record, n.534, p.24-37, mês.

MCFADDEN, D. (1997). Rationality for economists. In: NSF SYMPOSIUM ON ELICITING PREFERENCES. U. C. Berkeley, Jun 1997. http:/elsa.berkeley.ed/symposia/nsf97wp.html

MCFADDEN, D. (1986). The choice theory approach to market research. Marketing Science, v.5, p.275-297.

MENDES FILHO, E.F. (1997). Projeto evolucionário de redes neurais para avaliação de crédito financeiro. São Carlos. 85 p. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

MILLER, G.; TODD, P.; HEDGE, S.U. (1989). Designing neural networks using genetic algorithms. In: INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS, 3. Proceedings. Morgan Kaufman. p.379-384.

MILLER, I.; FREUND, J.; JOHNSON, R. (1990). Probability and Statistics for Engineers. Upper Saddle River, NJ, Prentice-Hall.

MOLDAU, J.H. (1985). A existência de objetivos irredutíveis e implicações para a análise das escolhas. São Paulo. 238p. Tese (Livre Docência) - Faculdade de Economia e Administração, Universidade de São Paulo.

MORIKAWA, T.; MCFADDEN, D. (1986). Discrete response to unobserved variables for which there are multiple indicators. Massachusetts, MIT / Department of Economics.

MOXON, B. (1996). Defining data mining. DBMS Online: DBMS data warehouse Supplement.

NIJKAMP, P. et al. (1996). Modelling inter-urban transport flows in Italy: a comparison between neural network analysis and logit analysis. Transportation Research C, v.4, n.6, p.323-338.

NÚMERO de coletivos é maior que a demanda de usuários. (1999). Jornal da Cidade, Bauru, 6 jun. Cardeno JC nos Bairros. p.1.

OPACIC, S.; POTTER, R.B. (1986). Grocery store cognitions of disadvantaged consumer groups: a reading case study. Tijdschrift voor Economische en Sociale Geografe, v.77, p.288-298.

OPENSHAW, S. (1977). Optimal zoning system for spatial interaction models. Enviroment and Planning A, v.9, p.169-184, mês.

ORTÚZAR, J.D.; WILLUMSEN, L. (1994). Modelling Transport. 2.ed. London, John Wiley.

PAU, L. F.; GÖTZCHE, T. (1992). Explanation facility for neural networks. Journal of Intelligent and Robotic Systems, v.5, p.193-206.

PIPKIN, J. S. (1995). Disaggregate models of travel behavior. In: HANSON, S., ed. The geography of urban transportation. 2.ed. New York, The Guilford Press. Cap.8, p.188-218.

PRECHELT. L. (1994). Proben 1: a set of neural network benchmark problems and benchmark rules. Germany, Fakultät für Informatik / Universität Karlsruhe. /Technical Report, 21/

QUINLAN, J.R. (1993). C4.5: programs for machine learning. San Mateo, Morgan Kaufman.

QUINLAN, J.R. (1986). Induction of decision trees. Machine Learning, v.1, n. , p.81-106.

RAIA JUNIOR, A.A. (2000). Acessibilidade e mobilidade na estimativa de um índice de potencial de viagens utilizando redes neurais artificiais e sistemas de informações geográficas. São Carlos. 217 p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo.

RECKER, W.W.; GOLOB, T.F. (1979). A non-compensatory model of transportation behavior based on sequential consideration of atributes. Transportation Research B, v.13, n.4, p.269-280.

RECKER, W.W.; MCNALLY, M.G.; ROOT, G.S. (1986a). A model of complex travel behavior: part I – theoretical development. Transportation Research A, v.20, p.307-318.

RECKER, W.W.; MCNALLY, M.G.; ROOT, G.S. (1986b). A model of complex travel behavior: part II – an operational model. Transportation Research A, v. 20, p.319-330.

RICHARDS, M.G.; BEN-AKIVA, M.E. (1975). A disaggregate travel demand model. Lexington, Lexington Books. (Saxon House Studies)

RIPLEY, B.D. (1996). Pattern recognition and neural networks. Cambridge, Cambridge University Press.

RIPLEY, B.D. (1993). Statistical aspects of neural networks. In: BARNDORFF-NIELSEN, O.E.; JENSEN, J.L.; KENDALL, W.S., eds. Networks and chaos-statistical and probabilistic aspects. London, Chapman and Hall. p.40-123.

RUMELHART, D.E.; HINTON, G.E.; WILLIAMS, R.J. (1986). Learning internal representation by error propagation. Parallel distributed processing: exploration in the microstructure of cognition, v.1, p.318-362.

SANCHES, S.P. (1988). Contribuição à análise operacional de redes de transporte coletivo em cidades de porte médio. São Carlos. 117 p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo.

SEADE – FUNDAÇÃO SISTEMA ESTADUAL ANÁLISE DE DADOS (2001). Pesquisa condições de vida e habitação. http://www.seade.gov.br (10 Jul.).

SETIONO, R.; LEOW, W.K. (2000). FERNN: an algorithm for fast extraction of rules from neural networks. Applied Intelligence, v.12, p.15-25.

SHAVLIK, J.W.; MOONEY, R.; TOWELL, G. (1991). Symbolic and neural net learning algorithms: an empirical comparison. Machine Learning, v.6, p.111-143.

SHMUELI, D.; SALOMON, I.;SHEFER, D. (1996). Neural network analysis of travel behavior: evaluating tools for prediction. Transportation Research C, v.4, n.3, p.151-166.

SILVA, A.N.R.; BRONDINO, N.C.M.; LIMA, R.S. (2000). O impacto de diferentes medidas de acessibilidade no valor do solo urbano. In: CONGRESSO DE PESQUISA E ENSINO EM TRANSPORTE, 14., Gramado. Programa Nacional da Pesquisa em Transporte 2000. p.349-359.

SIMONSEN, M.H. (1971). Teoria micro-econômica. Rio de Janeiro, Fundação Getúlio Vargas.

SMITH, G.C. (1992). The cognition of shopping centers by the central area and suburban elderly: an analysis of consumer information fields and evaluative criteria. Urban Geography, v.13, p.142-163.

SMITH, M.E. (1979). Design of small sample home interview travel surveys. Transportation Research Record, n.701, p.29-35.

SOBEL, K.L. (1980). Travel demand forecasting by using the nested multinomial logit model. Transportation Research Record, n.775, p.48-55.

SPEAR, B.D. (1977). Applications of new travel demand forecasting techniques to transportation: a study of individual choice models. Washington, Federal Highway Administration, US Department of Transportation. /Final Report to the Office of Highway Planning/.

SPIEGEL, M.R. (1984). Estatística. Trad.por Carlos A. Cruscius. 2.ed. São Paulo, Mcgraw-Hill.

STOPHER, P.R.; MEYBURG, A.H., eds. (1976). Behavioral travel-demand models. Massachusetts, Lexington Books; Heath and Company. p.3-53.

THAMIZ, V.; RENGARAJU, V.R.; KRISHNA RAO, K.V. (1996). Trip characteristics of travelers without vehicles. Journal of Transportation Engineering, p.76-81.

THRUN, S. (1995). Extracting rules from artificial neural networks with distributed representations. In:TESAURO, G.; TOURETZKY, D.; LEEN, T., eds. Advances in neural information processing systems. Cambridge, MIT Press. v.7, p.505-512.

TOWELL, G.G.; SHAVLIK, J.W. (1993). Extracting refined rules from knowledge-based neural networks. Machine Learning, v.13, p.71-101.

TVERSKY, A. (1972). Elimination-by-aspects: a theory of choice. Psychological Review, v.79, p.281-299.

VERLEYSEN, M. (1998). The explanatory power of artificial neural networks. In: METHODS: the explanatory power of models in the social sciences. Louvain-la-Neuve, s.ed.

WARNER, S.L. (1962). Strategic choice of mode in urban travel: a study of binary choice. Evanston, Northwestern University Press.

WATSON, P.L. (1974). The value of time: behavioral models of modal choice. Lexington, Lexington Books.

WEBBER, M.J. (1974). A theoretical analysis of aggregation in spatial choice models. Geographical Analysis, v.12, p.129-141.

WERMERSCH, F.G.; SILVA, A.N.R. (2001). Um problema potencial na obtenção de distâncias de viagem por transporte público urbano a partir de pesquisas origem-destino em ambiente SIG. In: CONGRESSO LATINOAMERICANO DE TRANSPORTE PUBLICO Y URBANO, 11., Havana - Cuba. Memorias... (em CD-ROM). Asociación de Investigación y Producción del Transporte - GRUPO IT.

WHITE, H. (1989). Learning in artificial neural networks: a statistical perspective. Neural Computation, v.1, p.425-461.

YANG, H. et al. (1993). Exploration of route choice behaviour with advanced traveler information using neural networks concepts. Transportation, v.20, p.199-223.

ZELL, A. et al (1995). SNNS Stuttgart neural network simulator – user manual, version 4.1. Germany, Universität Stuttgart/ I.P.V.R.. /Technical Report 6/

ZHANG, G.; PATUWO, E.; HU, M. (1998). Forecasting with artificial neural networks: the state of the art. International Journal of Forecasting, v.14, p.35-62.

APÊNDICE

Tabela com as subcategorias dos atributos utilizados para a composição do código de seleção de registros.

Algarismo associado à categoria	Atributos sócio-econômicos								Atributos de viagem		**Modo** de viagem escolhido
	n_autos	**n_idade**	**sit_fam**	**sexo**	**gr_inst**	**cart_mot**	**n_salar**	**tam_fam**	**motivo**	**dist**	
1	0	de 5 a 10	chefe de família	masc.	1ª a 4ª primário	possui	1a 5	1	trabalho	0 a 500	carro
2	1	de 10 a 15	cônjuge	femin.	primário completo ou ginásio incompleto	não possui	5 a 10	2	escola	500 a 1.000	carona
3	2 ou mais	de 15 a 18	filhos		primeiro grau completo ou segundo incompleto		10 a 20	3	compras	1.000 a 2.000	a pé
4		de 18 a 25	parentes		segundo grau completo ou terceiro incompleto		20 a 40	4	saúde	2.000 a 5.000	ônibus
5		de 25 a 40	agregados		terceiro grau completo		maior 40	5 ou mais	lazer	5.000 ou mais	
6		de 40 a 60	empregada ou residente						outros		