JOSÉ ALVES FERREIRA

Data Mining em Banco de Dados de Eletrocardiograma

Tese apresentada ao Instituto Dante Pazzanese de Cardiologia, Entidade Associada da Universidade de São Paulo para obtenção do título de Doutor em Ciências

Programa de Medicina, Tecnologia e Intervenção em Cardiologia

Orientador: Prof. Dr. Denys Emílio Campion Nicolosi

SÃO PAULO
2014

Dados Internacionais de Catalogação na Publicação (CIP)
Preparada pela Biblioteca do Instituto Dante Pazzanese de Cardiologia
©reprodução autorizada pelo autor

Ferreira, José Alves

Data Mining em Banco de Dados de Eletrocardiograma / José Alves Ferreira.—São Paulo, 2014.

Tese(doutorado)--Instituto Dante Pazzanese de Cardiologia
Universidade de São Paulo

Área de Concentração: Medicina, Tecnologia e Intervenção em Cardiologia

Orientador: Prof. Dr. Denys Emílio Nicolosi Campion

Descritores: 1.Data Mining 2. Apriori 3. Regras de Associação 4. Cardiologia 5. KDD 6. Eletrocardiograma

USP/IDPC/Biblioteca/025/14

DEDICATÓRIA

Essa tese é inspirada em minhas irmãs Joseane Alves Ferreira e Jorli Alves Ferreira, e em meus pais: Therezinha Alves Ferreira e Francisco Ferreira. Em minhas queridas filhas Carolina Velloza Ferreira e Lígia Velloza Ferreira, e, finalmente, em minha esposa Maria Lúcia Velloza Ferreira. Sempre tive apoio deles em todos os momentos da minha vida, apoio esse que nunca esperou recompensa. Assim, dedico a eles o que resultar de positivo desse trabalho.

AGRADECIMENTOS

No desenvolvimento, desta tese, tive o auxílio imprescindível de pessoas que contribuíram direta e significativamente para o andamento desse estudo. Estou seguro de que este apoio foi indispensável, sendo mais do que merecida a lembrança desses nomes: Ary Fagundes Bressane, na orientação sobre *data mining*, com a suite *Weka*; Cantídio Moura Campos Filho, na disponibilização e organização da base de dados; Anna Simene, na redação do meu trabalho; Dr. Francisco Faustino de Albuquerque Carneiro de França e Dra. Virginia Braga Cerutti Pinto, na orientação dos conceitos médicos; Fabiano Fernandes, nos conceitos técnicos sobre *data mining*, assim como todo o grupo da Dra Solange Oliveira Rezende, de São Carlos; Denys Emilio Campion Nicolosi, que vai muito além da orientação; Lígia Velloza Ferreira e Carolina Velloza Ferreira, no apoio da redação dessa tese, e Maria Lúcia Velloza Ferreira, por estar ao meu lado.

Não poderia também deixar de agradecer ao Instituto Butantan, ao Instituto Dante Pazzanese de Cardiologia e à Universidade de São Paulo pela oportunidade de realizar esse trabalho.

*A mente que se abre a uma nova **ideia** jamais*
voltará ao seu tamanho original.
Albert Einstein

NORMATIZAÇÃO ADOTADA

Esta tese está de acordo com as seguintes normas, em vigor no momento da sua publicação:

Referências: adaptado de *International Commitee of Medical Journals Editors* (Vancouver).

Universidade de São Paulo. Faculdade de Medicina. Divisão de Biblioteca e Documentação.

Guia de apresentação de dissertações, teses e monografias. Elaborado por Anneliese Carneiro da Cunha, Maria Julia de A.L. Freddi, Maria Fazanelli Crestana, Marinalva de Souza Aragão, Suely Campos Cardoso, Valéria Vilhena. 3. ed. São Paulo: Serviço de Biblioteca e Documentação – SBD/FMUSP; 2011.

Abreviatura dos títulos dos periódicos de acordo com "*List of Journals Indexed in Index Medicus*".

SUMÁRIO

LISTA DE ABREVIATURAS

bpm	Batimentos por minuto
CAD	Doenças das artérias coronárias
CNPQ	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CPU	Unidade Central de Processamento de um computador
CSV	Comma Separated Values – Valores Separados por Vírgulas
DBECG	Diretrizes Brasileira de ECG 2009
DCBD	Descoberta de Conhecimento em Banco de Dados
ECG	Eletrocardiograma
HAS	Hipertensão Arterial Sistêmica
IDPC	Instituto Dante Pazzanese de Cardiologia
GNU	Animal Gnu (é um sistema operacional elaborado por pessoas que trabalham em conjunto para a liberdade de todos os usuários de *software*)
GPL	*General Public License*
IAM	Infarto Agudo do Miocárdio
IMC	Índice de Massa Corporal
KDD	*Knowledge Discovery in Databases*
USP	Universidade de São Paulo
Weka	*Waikato Enviroment for Knowledge Analysis*

LISTA DE CÓDIGOS DE DIAGNÓSTICOS

D130	Ritmo sinusal
D139	Bradicardia sinusal
D175	Bloqueio atrioventricular do primeiro grau
D190	Ruído de artefato
D214	Sobrecarga Ventricular Esquerda
D220	Bloqueio de ramo direito
D223	Bloqueio de ramo esquerdo
D224	Bloqueio divisional ântero-superior esquerdo
D263	Eletrocardiograma normal
D290	Alteração da repolarização ventricular em parede inferior
D292	Alteração da repolarização ventricular em parede ântero-lateral
D300	Distúrbio de condução no ramo direito
D305	Alterações morfológicas
D312	Ausência de dados clínicos

LISTA DE MEDIDAS DO ELETROCARDIOGRAMA

P	Largura, em segundos, da onda “P” do eletrocardiograma
PRi	Largura, em segundos, do intervalo “PR” do eletrocardiograma
QRS	Largura, em segundos, do intervalo “QRS” do eletrocardiograma
QT	Largura, em segundos, do intervalo “QT” do eletrocardiograma
QTc	Largura corrigida, em segundos, do intervalo “QT” do eletrocardiograma
RR	Largura, em segundos, do intervalo “RR” do eletrocardiograma
sap	Ângulo da onda “P” do eletrocardiograma
saqrs	Ângulo do complexo “QRS” do eletrocardiograma
sat	Ângulo da onda “T” do eletrocardiograma
T	Largura, em segundos, da onda “T” do eletrocardiograma

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Ferreira JA. *Data Mining em Banco de Dados de Eletrocardiograma* [tese]. São Paulo: Instituto Dante Pazzanese de Cardiologia, Entidade Associada da Universidade de São Paulo; 2014.

Neste estudo, foi proposta a exploração de um banco de dados, com informações de exames de eletrocardiogramas (ECG), utilizado pelo sistema denominado Tele-ECG do Instituto Dante Pazzanese de Cardiologia, aplicando a técnica de *data mining* (mineração de dados) para encontrar padrões que colaborem, no futuro, para a aquisição de conhecimento na análise de eletrocardiograma. A metodologia proposta permite que, com a utilização de *data mining*, investiguem-se dados à procura de padrões sem a utilização do traçado do ECG. Três pacotes de *software* (*Weka*, *Orange* e *R-Project*) do tipo *open source* foram utilizados, contendo, cada um deles, um conjunto de implementações algorítmicas e de diversas técnicas de *data mining*, além de serem *softwares* de domínio público. Regras conhecidas foram encontradas (confirmadas pelo especialista médico em análise de eletrocardiograma), evidenciando a validade dessa metodologia.

Descritores: *Data mining*; *Apriori*; Regras de Associação; Cardiologia; KDD; Eletrocardiograma.

SUMMARY

Ferreira JA. *Data Mining in Electrocardiogram Databases* [thesis]. São Paulo: Instituto Dante Pazzanese de Cardiologia, Associated Entity of Universidade de São Paulo; 2014.

In this study, the exploration of electrocardiograms (ECG) databases, obtained from a Tele-ECG System of Dante Pazzanese Institute of Cardiology, has been proposed, applying the technique of data mining to find patterns that could collaborate, in the future, for the acquisition of knowledge in the analysis of electrocardiograms. The proposed method was to investigate the data looking for patterns without the use of the ECG traces. Three Data-mining open source software packages (Weka, Orange and R - Project) were used, containing, each one, a set of algorithmic implementations and various data mining techniques, as well as being a public domain software. Known rules were found (confirmed by medical experts in electrocardiogram analysis), showing the validity of the methodology.

Descriptors: Data mining; *Apriori*; Association rules; Cardiology; KDD; Electrocardiogram.

1 INTRODUÇÃO

1 INTRODUÇÃO

Doenças cardiovasculares são a principal causa de morte no Brasil[1]. O sexo masculino apresenta número maior de mortes (em 2011, representando 57%) devido, principalmente, a problemas cardíacos do que a qualquer outra doença.

Para prevenir ou diagnosticar precocemente o estado de saúde de pacientes, um dos métodos mais presentes é a realização do eletrocardiograma, tanto para o tratamento como na prevenção de alguma doença cardíaca[2].

O exame de eletrocardiograma é um método não invasivo, de baixo custo, muito difundido, de grande utilidade e amplo conhecimento. É uma ferramenta poderosa, quando aliado ao conhecimento médico, além de ser de fácil reprodutibilidade, podendo ser realizada em qualquer ponto no qual exista um eletrocardiógrafo.

Como em outros procedimentos, o eletrocardiograma avaliado por não cardiologistas é susceptível a falhas. Esses equívocos diagnósticos estão mais propensos à imprecisão, principalmente quando gerados por não especialistas.

Os serviços de telemedicina podem amenizar essa situação levando a locais distantes, em tempo hábil, a informação necessária para a tomada de decisão.

Infelizmente, há muitos locais em que esse tipo de serviço não está disponível ao médico. Em outras situações, devido à condição do paciente, não é possível esperar por intervenção, por isso é fundamental o diagnóstico em tempo hábil.

Para minimizar riscos aos pacientes, é importante realizar, rapidamente, e reduzir o número de falhas em análises de ECG.

É possível armazenar informações de eletrocardiogramas, em bases de dados, possibilitando encontrar relações importantes. Nessas bases, o

conjunto de padrões selecionados será denominado de conhecimento. O uso de ferramentas adequadas para identificar por padrões úteis pode se revelar forte aliado, auxiliando na geração de diagnósticos mais precisos.

O processo de descoberta de conhecimento em banco de dados (DCBD) consta de identificar padrões novos que estão ocultos, em bases de dados, no domínio da aplicação, válidos e significativos.

Os dados utilizados neste estudo foram coletados no momento da execução do ECG e, posteriormente, armazenados no banco de dados, sendo utilizado para a extração do conhecimento desejado. Na coleta desses dados, foram utilizadas as medidas do eletrocardiograma, a história clínica do paciente, além de dados pessoais do paciente que realizou o exame.

Há diversos locais em que os exames de ECG são coletados. Cada centro de atendimento possui um equipamento que se conecta, remotamente, com o Instituto Dante Pazzanese de Cardiologia (IDPC), via tecnologia de telefonia celular, enviando o traçado do eletrocardiograma do paciente, assim como as demais informações que compõem a base de dados. Esses dados são recebidos por uma equipe médica, de plantão permanente, que analisa o eletrocardiograma e retorna um laudo em poucos minutos, oferecendo informação confiável em qualquer parte do Estado de São Paulo em que esse serviço esteja disponível.

A solução proposta utilizou mineração de dados como técnica por busca de padrões. A definição mais referida na literatura sobre *data mining* foi elaborada por Usama Fayyad[3]: "o processo não trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis".

O processo completo é denominado *Knowledge Discovery in Databases* (KDD). O *data mining* é parte fundamental nessa busca para encontrar os padrões significativos, ou seja, conhecimento em grandes volumes de informações armazenados, sem que, necessariamente, haja hipótese ou ideia predefinida.[4]

O processo de *data mining* necessita de um computador (do tipo PC) que pode ser encontrado em, praticamente, todos os locais de atendimento a pacientes. Armazenar grandes volumes de dados, atualmente, é simples e de baixo custo. Esse quadro possibilita encontrar conhecimento de forma automática.

Na realização desse estudo, foram utilizados pacotes de *softwares* selecionados, reunindo grande quantidade de algoritmos, com interfaces amigáveis, de uso livre e capazes de manipular a base de dados.

2 OBJETIVOS

2 OBJETIVOS

1) Explorar a base de dados;
2) Encontrar relações novas ou válidas entre dados de ECG e os diagnósticos;
3) Destacar o volume de regras de associação entre os atributos contidos na base de dados;
4) Gerar novos conhecimentos ou constatar aqueles já consagrados, com as informações da base de dados para enriquecer o conhecimento em análise de ECG, na determinação de diagnósticos, sem a utilização do traçado do eletrocardiograma.

3 REVISÃO BIBLIOGRÁFICA

3 REVISÃO BIBLIOGRÁFICA

Carvalho[5] afirma que as bases de dados disponíveis continuam aumentando consideravelmente e que uma maneira de descobrir conhecimento é por meio da descoberta de regras de associação, via a técnica de taxonomia. Fujimoto[6] relata que a descoberta de conhecimento se tornou um fator primordial para tomadas de decisões nos diversos ramos de atuação humana e que técnicas de visualização de informação, com apoio de medidas de avaliação, se mostraram importantes para a compreensão e identificação de regras de associação generalizadas. Witten *et al*.[7] esclarecem que, embora as informações estejam disponíveis, estão escondidas, e são difíceis de serem encontradas e analisadas, justificando um processo automático para encontrar padrões que apresentem relevância para quem faz uso da informação. Navega[8] explica o processo de descoberta de padrões relatando a necessidade de analistas humanos no processo, introduzindo o pensamento por trás da mineração de dados. Rezende[9] investiga técnicas de exploração, dentre elas, as regras de associação, nas quais é importante identificar os dois principais fatores que podem torná-las subjetivamente interessantes para o usuário, ou seja, utilidade e *inesperabilidade*. Silva[10] propõe um novo algoritmo para gerar regras de associação para dados numéricos em intervalos não contínuos, concluindo que se obtêm resultados relevantes. Geng *et al*.[11] estudam o papel do uso das diversas medidas de interesse no processo de pós-processamento para criar um *ranking* de padrões, permitindo uma análise baseada na aplicação de cada caso de *data mining*. Zupan *et al*.[12] executam revisão da evolução dos *softwares* de *data* mining, comparando as ferramentas comerciais com as *open source* (uso livre e código aberto), importantes para a área biomédica. Becher *et al*.[13] investigaram um processo automático para seleção do conjunto ótimo de atributos para a montagem de um modelo de exploração, reduzindo o tempo de utilização de

CPU, aumentando a exatidão dos resultados encontrados. Park *et al*.[14] propõem um algoritmo, baseado em técnica *hash* (*hash* é um tipo de organização arquivo que fornece acesso rápido aos registros sob certas condições), obtendo redução do conjunto de *itemsets* e diminuindo o escopo do problema. Brossete *et al*.[15] utilizaram regras de associação obtidas de dados de vigilância em saúde para infecções hospitalares e demonstraram serem eficientes na identificação de padrões novos e inesperados. Yang *et al*.[16] sugeriram um modelo de detecção de fraudes e abusos financeiros, em sistemas de seguro de saúde, obtendo resultados eficientes em encontrar casos que não puderam ser alcançados de forma manual. Romão *et al*.[17] utilizaram o algoritmo *apriori* para descobrir regras de associação no banco de dados do Diretório dos Grupos de Pesquisa no Brasil (disponibilizado pelo CNPQ), constituindo-se em fonte estratégica de informações sobre a pesquisa brasileira, inventariando os pesquisadores e sua produção intelectual.

O *data mining* também pode ser utilizado em aplicações médicas. Konias *et al*.[18] sugeriram um novo algoritmo que busca por regras de associações em série de dados temporais periódicas, medindo o grau da correlação entre variáveis, independentes de variações da linha de base e escalas de amplitude do ECG, e, após testes em 60 pacientes com falha cardíaca congestiva, obtiveram regras de associação complementares para análise de eletrocardiogramas. Burn-Thorton *et al*.[19] investigaram uma metodologia para encontrar indicadores, no sinal de ECG, de candidatos a apresentarem infarto agudo do miocárdio (IAM), aplicado a um banco de dados com 2.730 eletrocardiogramas no qual obtiveram exatidão de 76,6%. Murugan[20] descreveu um método propondo a classificação de batimentos isquêmicos, no ECG, utilizando regras de associação, atingindo exatidão superior a 84%. Ordonez *et al*.[21] destacaram a possibilidade de se utilizar o *data mining* em dados médicos e como formatar esses dados para a forma apropriada de aplicação das ferramentas de exploração, concluindo que as regras de associação são adequadas para mineração de dados médicos. Cabral *et al*. [22], objetivando classificar os beneficiários de um plano de saúde

(denominado CELOS) “com indicativo” e “sem indicativo” de apresentar IAM, reduzindo custos do plano e aumentado a qualidade de vida dos associados, alcançaram 100% de aprovação de um especialista médico. Alizadehasani *et al*.[23], motivados pelos riscos e altos custos relativos a angiografia, aplicaram vários algoritmos a uma base de dados médicas, obtendo alta taxa de acurácia (94,08%), maior que a encontrada na literatura especializada. Cavalcante[24], objetivando identificar e avaliar a prevalência de fatores de risco para doenças cardiovasculares e associação entre esses fatores em pacientes com doenças das artérias coronárias (CAD), utilizando regras de associação, comprovou a prevalência de fatores de risco cardiovasculares. Nahar *et al*.[25] investigaram os fatores de saúde e de doença que contribuem para doenças cardíacas, usando regras de associação com diferentes algoritmos, encontrando fatores de risco característicos para cada sexo. Exarchos *et al*.[26] investigaram vários algoritmos de regras de associação para classificar “batimentos isquêmicos” ou “batimentos não isquêmicos”, em ECG de longa duração, alcançando 87% de sensitividade e 93% de especificidade em uma base de dados europeia denominada *European Society of Cardiology ST-T database.*

4 MATERIAIS E MÉTODOS

4 MATERIAIS E MÉTODOS

Trata-se de estudo observacional, não intervencionista. Situado no campo de extração de conhecimento a partir da base de dados utilizando *knowledge discovery in databases*.

4.1 Casuística

Este estudo utilizou o banco de dados formado com as informações de exames de eletrocardiograma do sistema Tele-ECG, do Instituto Dante Pazzanese de Cardiologia, contendo, atualmente, cerca de 550.000 eletrocardiogramas armazenados. Para a exploração, na busca de padrões úteis e importantes, uma amostra representativa da base de dados foi extraída do total de exames disponíveis.

Como *data mining* é uma técnica, por definição, não estatística, não há cálculo do tamanho mínimo de amostra, conforme orientação do Laboratório de Estatística e Epidemiologia do Instituto Dante Pazzanese de Cardiologia.

O projeto foi submetido ao Comitê de Ética, conforme Protocolo nº 4004.

4.2 K*nowledge-Discovery in Databases* KDD

4.2.1 Introdução ao KDD

Com as facilidades no desenvolvimento na área de Tecnologia da informação, grande quantidade de dados vem sendo acumulada em praticamente todas as áreas do conhecimento humano. Porém, o volume de informações é tão grande que um exame mais detalhado na busca de

padrões se tornou inviável. Por isso, é necessária a utilização de ferramentas adequadas para auxiliar nessa tarefa. Torna-se imprescindível o desenvolvimento de metodologias que ajudem o ser humano nessa tarefa, de forma automática e apropriada para analisar, interpretar e relacionar essas informações no contexto de cada aplicação[27].

Dessa forma, surgiu uma nova área, denominada *Knowledge Discovery in Databases*, que vem despertando grande interesse junto às comunidades científicas e industriais[28].

Há grande complexidade em realizar o KDD devido ao fato de se trabalhar com grandes quantidades de dados durante o processo e, também, na dificuldade em conjugar dinamicamente as interpretações de forma a decidir que ações devem ser realizadas, em cada caso, nas diversas fases do processo[27].

A tarefa de coordenação durante o processo de KDD é realizada por ser humano, o qual deve possuir o conhecimento necessário para tal tarefa, além de experiência para tomar as decisões[3].

Há dificuldades no processo de KDD, tais como o nível de detalhamento utilizado, que pode ser diferente e difuso, opiniões de especialistas que podem variar, além de conhecimentos prévios a serem utilizados, variando de acordo com cada especialista de KDD[28].

É comum entender o KDD como uma solução para realizar tarefa que combina conhecimento e experiência, a fim de alcançar conhecimento não aparente, para atingir a descoberta de algo que não se conhece a princípio[4,27].

4.2.2 O Processo de KDD, visão geral

O termo KDD foi formalizado em 1989 em referência ao amplo conceito de busca por conhecimento a partir de bases de dados[3].

Uma das definições mais populares é: “KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões

compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados”[3].

Na definição de Fayyad[3], interativo indica a necessidade do elemento que seja o responsável pelo controle do processo. Este elemento controlador representado pelo ser humano que realiza a análise e interpretação dos fatos na condução do processo.

Na mesma definição, iterativo alerta para a possibilidade de repetições do processo de KDD, parciais ou integrais, visando encontrar resultados satisfatórios por meio de refinamentos sucessivos.

A expressão “não trivial” reforça a complexidade presente na execução do KDD.

O KDD em bases de dados é composto por várias etapas sequenciais. A Figura 1 apresenta resumo das etapas operacionais executadas nesse processo. A etapa de pré-processamento visa preparar os dados, consolidando as informações de valor para os algoritmos na etapa seguinte pesquisarem por padrões nos dados (padrões são unidades de informação que se repetem, ou, então, são sequências de informações que dispõem de uma estrutura que se repete).

Nessa etapa, os dados são coletados, armazenados e organizados. Na sequência, serão realizadas as tarefas de limpeza, seleção dos dados, transformação ou codificação dos dados.[29]

Na próxima etapa, deve ser realizada a busca efetiva por padrões úteis no contexto da aplicação de KDD, é o denominado *data mining*.

A última etapa do KDD é o pós-processamento que trata do conhecimento obtido na mineração de dados. Tal tratamento objetiva avaliar a utilidade do conhecimento descoberto.[3]

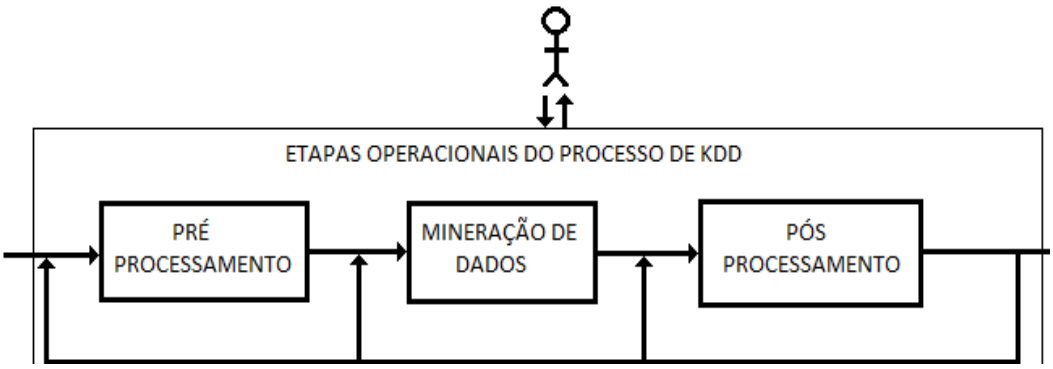


Figura 1 - Resumo das etapas operacionais de KDD, adaptada de Goldschmidt[27].

A complexidade inerente ao processo de KDD deriva de fatores operacionais e fatores de controle.[3,27,28]

Como exemplo de fatores operacionais, pode-se citar a exigência de manipular grandes volumes de dados heterogêneos e tratar resultados com diferentes formatos.

Fatores de controle são mais complexos e envolvem julgar como conduzir processos de KDD. Entre tais fatores, podem ser citados:

a) A dificuldade na formulação precisa dos objetivos a serem alcançados em processo de KDD;
b) A complexidade na escolha do algoritmo de *data mining* para busca de resultados satisfatórios. Isso é acentuado na medida em que surgem novos algoritmos[10,16,18-21,30] com o mesmo propósito, aumentando a diversidade de alternativas. Normalmente, a escolha dos algoritmos é limitada às opções conhecidas pelo analista de KDD;
c) A escolha da parametrização é complexa e deve ser adequada para os algoritmos em cada nova situação, podendo aumentar o número de iterações durante o processo de KDD;
d) A dificuldade inerente em controlar um processo longo com inúmeras alternativas;[7]

e) Ao limite da capacidade da memória humana para lembrar os eventos com o passar do tempo, afetando a capacidade de efetuar as comparações entre as alternativas e resultados, e, assim, comprometendo a tomada de decisões futuras;[7]

f) Há ainda a questão recorrente da necessidade de executar o processo de KDD para a descoberta de algo que não se sabe exatamente a princípio.

4.2.3 Pré-processamento

O pré-processamento deve cumprir as etapas a seguir:

a) A seleção dos dados na qual são identificadas as informações que deverão ser consideradas durante o processo de KDD. Por exemplo, o nome de um paciente é irrelevante em processo em que se deseja classificar as possíveis doenças, já a data de nascimento é importante;

b) Deve ser realizada a limpeza dos dados, na qual são efetuados os tratamentos (checagem de consistência, correções de erros, exclusão de valores nulos e redundantes) para propiciar a qualidade dos fatos por eles representados (integridade, veracidade e completude);

c) Não menos importante é o tratamento das informações ausentes, errôneas ou inconsistentes para não comprometer os resultados finais esperados;

d) A codificação dos dados deve ser executada a fim de permitir que nestes sejam aplicados os algoritmos escolhidos. Para tal, há duas possibilidades de algoritmos:

   i) numéricos, nos quais as informações a serem processadas devem ser convertidas do formato de dados reais para intervalos ou categorias. Esse processo é denominado

discretização (no Apêndice A, pode-se ver maiores esclarecimentos sobre esse procedimento), pois cria intervalos de faixas de valores para os dados originais;

ii) categórica, na qual a transformação deve representar numericamente os valores de atributos categóricos;

e) Finalmente, cabe a fase de enriquecimento dos dados, na qual novas informações podem integrar mais conhecimento no processo de descoberta de padrões. Essas novas informações são anexadas, com a ajuda do especialista do domínio de aplicação, com dados que não existem na base de dados, mas que são reconhecidos como úteis para o processo de descoberta de conhecimento.

4.2.4 *Data mining*

Na etapa de *data mining*, é realizada a busca propriamente dita dos padrões desejados no ambiente da aplicação do KDD. Muitos autores citam essa etapa como a principal dentro desse processo.

Na busca de padrões, são escolhidas técnicas (redes neurais, algoritmos genéticos, modelos estatísticos e probabilísticos, etc.) e os algoritmos a ser utilizados nos processos de descoberta de padrões[7]. A escolha da técnica apropriada depende do tipo de tarefa de KDD a ser realizada. Podemos citar algumas técnicas, como exemplo:

a) “Descoberta de Associação”, que procura por itens que ocorrem frequentemente em transações do banco de dados. O exemplo clássico é o de “cestas de compras” na área de marketing com a intenção de estimular a compra de itens que são comprados conjuntamente;

b) “Classificação”, que se fundamenta em descobrir conjuntos de registros com propriedades similares, denominadas classes;

c) “Regressão”, na qual se procura por uma função que possa mapear os registros na base de dados em valores numéricos reais;

d) “*Clusterização*”, é utilizada para separar os registros da base de dados em subconjuntos, ou *clusters*, de tal maneira que seus elementos compartilhem de características comuns, diferenciando-os dos elementos dos outros *clusters* (agrupamentos). Distingue-se da classificação por não possuir rótulos predefinidos e sim encontrar, de forma automática, os novos agrupamentos de dados[3];

e) “*Sumarização*”, que pretende encontrar características comuns entre conjunto de dados;

f) “Detecção de desvios”, na qual se examina os registros do banco de dados com características que sejam diferentes dos padrões considerados normais, dentro do contexto, são os denominados *outliers* (valores discrepantes);

g) “Descoberta de sequências”, na qual se procura as associações ocorridas ao longo de um período de tempo.

4.2.5 Pós-processamento

Após a etapa de *data mining*, ocorre a fase de pós-processamento, na qual o conhecimento obtido é refinado para facilitar a avaliação, pelo especialista na área de aplicação, da utilidade e novidade do conhecimento descoberto. Consta, nessa fase, a organização, simplificação das informações, geração de gráficos, diagramas ou relatórios para análise do especialista da área de aplicação.

As regras de associação são, usualmente, filtradas, eliminando as consideradas menos importantes, reduzindo-se o conjunto que será analisado pelo especialista no domínio da aplicação.

4.2.6 Componentes do processo

O processo de KDD é composto de:[27]

a) o problema no qual será realizado o processo de KDD;
b) os recursos disponíveis para a solução do problema; e
c) os resultados obtidos, a partir da aplicação, com os recursos disponíveis para a solução do problema.

Detalhamento dos componentes:

a) O problema no qual será realizado o processo de KDD. Pode ser caracterizado por três elementos:
   - a_1) O conjunto das características do conjunto de dados. O KDD espera que os dados estejam organizados em uma única estrutura tabular bidimensional contendo casos e características do problema. O tratamento e a consolidação dos dados são úteis e desejáveis no processo de KDD;
   - a_2) O especialista no domínio da aplicação que representa a pessoa ou o grupo de pessoas que domina o assunto no ambiente da aplicação de KDD. Ele possui o conhecimento prévio sobre o problema e a influência desde a definição dos objetivos do processo até a avaliação dos resultados;
   - a_3) Os objetivos da aplicação que visam atender os aspectos desejados para o modelo a ser implantado, respeitando os limites e expectativas dos especialistas no domínio da aplicação. Pode-se citar como exemplo a exatidão mínima aceita para o modelo de conhecimento, ou seja, o que será considerado apropriado para a aplicação. Normalmente, os objetivos do KDD levam em consideração os anseios dos especialistas no domínio, apesar de haver situações nas quais os objetivos não são claros no início do processo[4],

sendo refinados ao longo de todo o procedimento de KDD, em função dos resultados intermediários.

b) Recursos disponíveis para solução do problema. Pode-se ressaltar:

   b_1) O especialista em KDD é composto de um indivíduo ou mais, com experiência em processos de KDD. Ele age de acordo com o especialista no domínio de aplicação, e conduz o processo em cada fase. Executa todo o processo, além de avaliar os resultados encontrados;

   b_2) Ferramenta de KDD é o termo utilizado para se referir a todos os recursos computacionais que podem ser usados no processo de busca pelo conhecimento. Pode ser constituído por *software* integrando as funcionalidades de análise e tratamento dos dados, e, também, algoritmos isolados que possam fazer parte do processo de KDD;

   b_3) Plataforma computacional é o conjunto de recursos de *hardware* que possam ser utilizados na execução do KDD. Varia desde uma máquina isolada até um conjunto de computadores trabalhando em paralelo.

c) Os resultados obtidos na aplicação são constituídos pelos modelos usados para descoberta do conhecimento, assim como o histórico das atividades realizadas.

Os modelos de conhecimento são avaliados em relação ao cumprimento das metas estabelecidas nos objetivos da aplicação. A expressão “modelo de conhecimento” indica qualquer abstração de conhecimento, expresso em alguma linguagem, que descreva algum conjunto de dados[27].

Os históricos das atividades realizadas fazem parte dos resultados e sua utilização permite uma análise crítica.

4.2.7 Modelagem da metodologia

Para superar a complexidade dos processos de KDD, utiliza-se metodologia com planejamento das atividades a serem desenvolvidas para atender os objetivos propostos (Figura 2). Dessa maneira, pode-se dividir essa metodologia em:

a) levantamento inicial;
b) definição dos objetivos.

Para atender esses itens, devem ser realizadas as tarefas:

a) Definir pessoas e as áreas envolvidas;
b) Buscar pelo *hardware* e *software* mais adequados;
c) Detalhar as bases de dados disponíveis;
d) Averiguar os bancos de dados para definir o mérito dos atributos, seu significado, sua qualidade e a quantidade disponível;
e) Delinear as necessidades e expectativas nas áreas envolvidas;
f) Assinalar o conhecimento prévio.

Um requisito fundamental é definir os objetivos, e, para tal, é necessário o entendimento da situação na qual o processo será realizado. Exame cuidadoso da natureza dos dados é imprescindível. Não menos importante é a interação entre o especialista do KDD com o especialista no domínio da aplicação.

O analista do KDD, então, deve listar as expectativas no domínio da aplicação e agrupá-las de acordo com suas naturezas. Assim, é possível definir o tipo de tarefa de KDD deve ser aplicada.

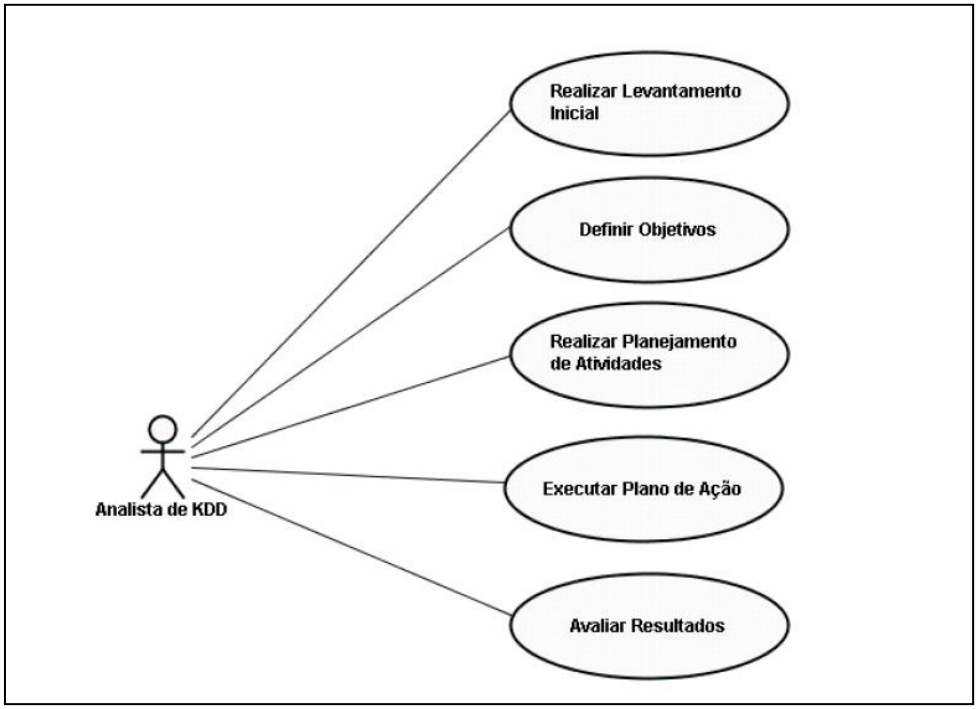


Figura 2 - Atividades a serem desenvolvidas no KDD.[27]

É importante anotar os passos intermediários, ou resultados parciais, no processo, permitindo a comparação dos resultados entre si e com o conhecimento prévio acumulado.

Posteriormente, a escolha do tipo de algoritmo apropriado é, então, realizada.

O KDD deve iniciar com a aplicação dos tipos de busca de padrões, testando os diversos valores nos parâmetros oferecidos para os algoritmos a fim de se encontrar o conhecimento embutido na base de dados.

Como é um processo iterativo, pode ser repetido parcial ou totalmente na busca de melhores resultados. É nessa hora que a documentação realizada ao longo do processo é extremamente útil.

É praticamente impossível testar todas as alternativas disponíveis, tanto para técnicas, algoritmos, como também para os parâmetros de uma dada aplicação, o que reforça a importância do planejamento anterior. O processo é refinado ao longo de toda a execução, buscando aprimorar os resultados para atender os objetivos, como na Figura 3.

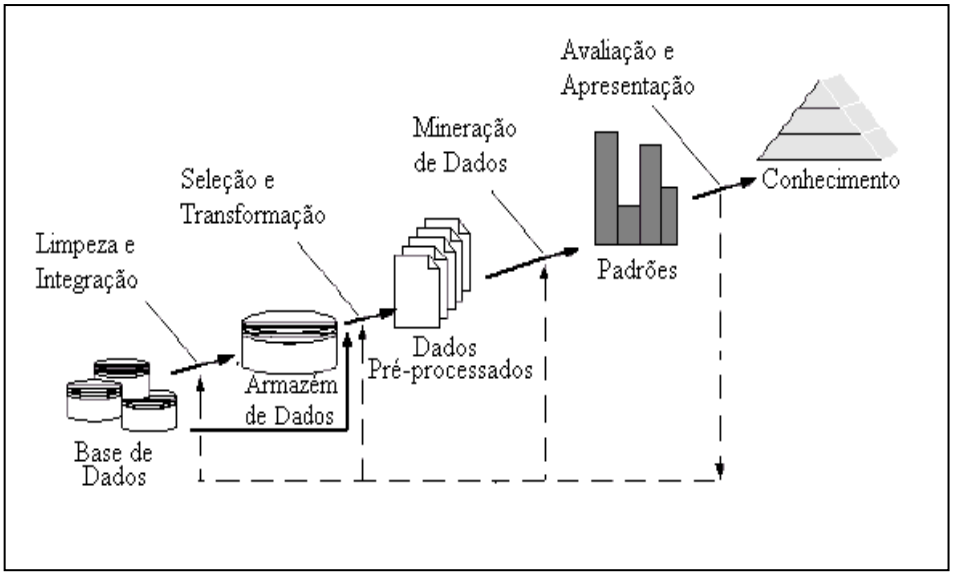


Figura 3 - Etapas do processo de KDD (copiada).[10]

4.3 Montagem da base de dados

Nesse estudo, o banco de dados utilizado para exploração foi montado com as informações de eletrocardiogramas. Em cada linha, há dados relativos de um ECG, sendo composta pelo conjunto de características definidas por valores denominados atributos.

Tipicamente, recorre-se à métrica de frequência de ocorrências para encontrar quais são as associações significantes entre os atributos.

As regras de associação são representativas das relações entre os atributos e são selecionadas de acordo com métricas de interesse, tipicamente, *suporte* e *confiança* (para mais detalhes, ver Apêndice B).

4.4 Ferramentas para *Data Mining*

4.4.1 Extração dos padrões

Nessa etapa, são escolhidas as ferramentas para a extração dos padrões propriamente dita e a configuração para extração do conhecimento. Para esta pesquisa, foram utilizadas as ferramentas *Weka*[7], *Orange* e *R* com pacote *arules*.

4.4.2 O pacote *Weka*

Weka (*Waikato Environment for Knowledge Analysis*) é uma ferramenta gráfica que agrega algoritmos para mineração de dados. Foi desenvolvida pelo Departamento de Ciência da Computação da Universidade Waikato, na Nova Zelândia.

Foi escrita na linguagem Java, o que permite que seu código seja executado em diferentes plataformas, dando a esse *software* boa portabilidade. Além de ser distribuído sob a licença *General Public License* (GPL é a designação da licença para *software* livre idealizada por Richard Matthew Stallman, em 1989, no âmbito do projeto GNU da *Free Software Foundation*), o que lhe confere a possibilidade de se alterar o código fonte.

A versão utilizada foi a 3.6.9 e está disponível na Web (http://www.cs.waikato.ac.nz/ml/weka), com uma interface amigável. O ambiente utilizado foi o "*Explorer*", visualizado nas Figuras 4, 5 e 6. Seus algoritmos fornecem relatórios com dados analíticos e estatísticos no domínio minerado. Possui diversas técnicas de mineração de dados, incluindo classificação, seleção de atributos, agrupamento e busca por regras de associação.

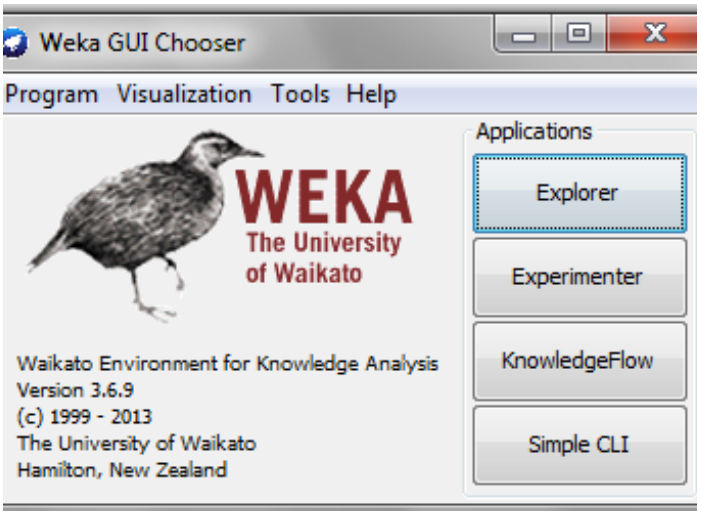


Figura 4 - Tela inicial do *Weka* para escolher a aplicação: Explorer.

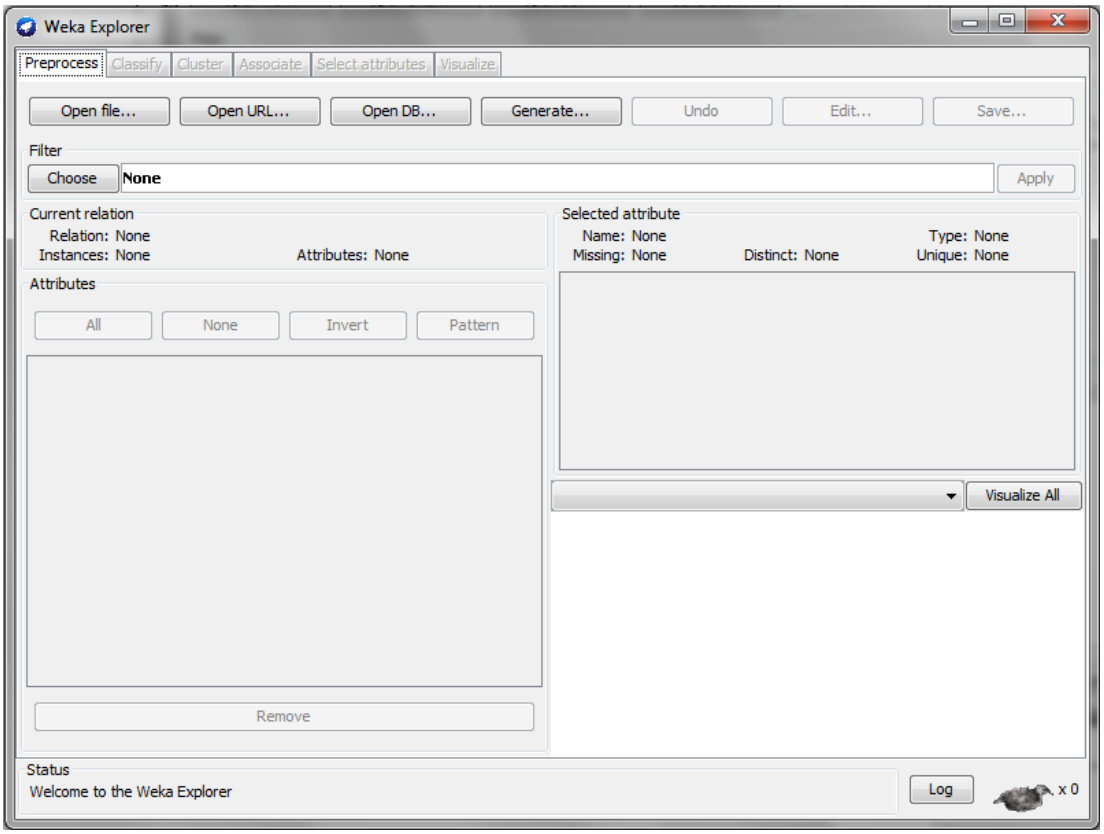


Figura 5 - Página inicial do ambiente *Weka*. Aqui é carregada a base de dados e aplicados os filtros desejados.

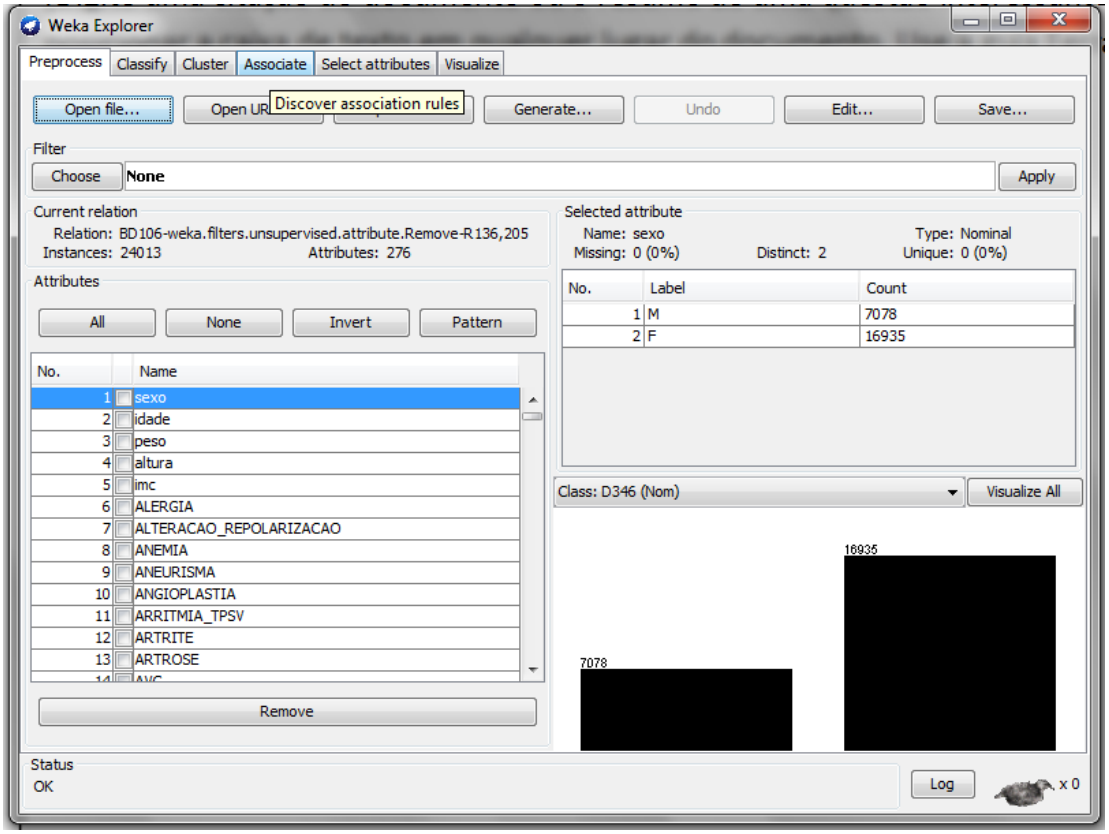


Figura 6 - Tela do *Weka* após carga da base de dados

4.4.3 *Orange*

Orange é uma ferramenta de *data mining* baseada em componentes e, também, um conjunto de programas de máquinas de aprendizado voltados para programação visual do tipo *front-end* (*front-end* é uma interface com o usuário que coleta os valores de entradas nas várias formas e executa o processamento para aplicar na ferramenta) usada para analisar os dados. Possui ainda uma biblioteca em linguagem *Python*. Inclui um conjunto de componentes para processar recursos, e tratar escores, filtrar, modelar e várias técnicas para explorar os dados. Foi implementada em linguagem *C++* e *Python* (*Python* é uma linguagem de programação de alto nível de uso geral, realçando a legibilidade do código. Frequentemente utilizada como uma linguagem de *script*, mas também é usada em contextos *não script*).

Orange é distribuído como *software* livre sob o GPL, e é mantido pelo Laboratório de Bioinformática da Faculdade de Computação e Ciência da Informação da Universidade de Ljublajana, Slovenia.

O *Orange* procura atender:

a) usuário desde iniciante em *data mining,* que pode desenhar suas próprias análises *pipelines* (técnica de encadeamento de um conjunto de processos a serem executados por meio de seus fluxos padrão, de tal forma em que a saída de um processo é utilizada como entrada para o processo seguinte) sem qualquer *script* ou programação Python;
b) até programadores, que preferem acessar a ferramenta por meio de uma interface de *script*.

Orange pode ser executado em diversas plataformas, *Linux, Mac OS X* e *Windows*.

Utiliza da técnica de *pipeline* para ler os dados, visualizar o modelo, explorá-lo e interpretar os resultados.

Orange pode ser utilizado com diferentes combinações para os seus *widgets* para realizar desde o pré-processamento até a aplicação de algoritmos de exploração. Um *widget* é um componente de interface gráfica do usuário (GUI), o que inclui janelas, botões, menus, ícones, barras de rolagem, etc. Cada *widget* oferece algumas funcionalidades básicas, tais como a leitura dos dados, mostrados em uma tabela de dados, escolhendo os recursos, seja manualmente ou com base em alguns escores, preditores de treinamento, opções para validar os dados, e assim por diante. O usuário conecta os *widgets* pela comunicação de canais. A força e flexibilidade do *Orange* estão nas diferentes maneiras nas quais os *widgets* podem ser combinados.

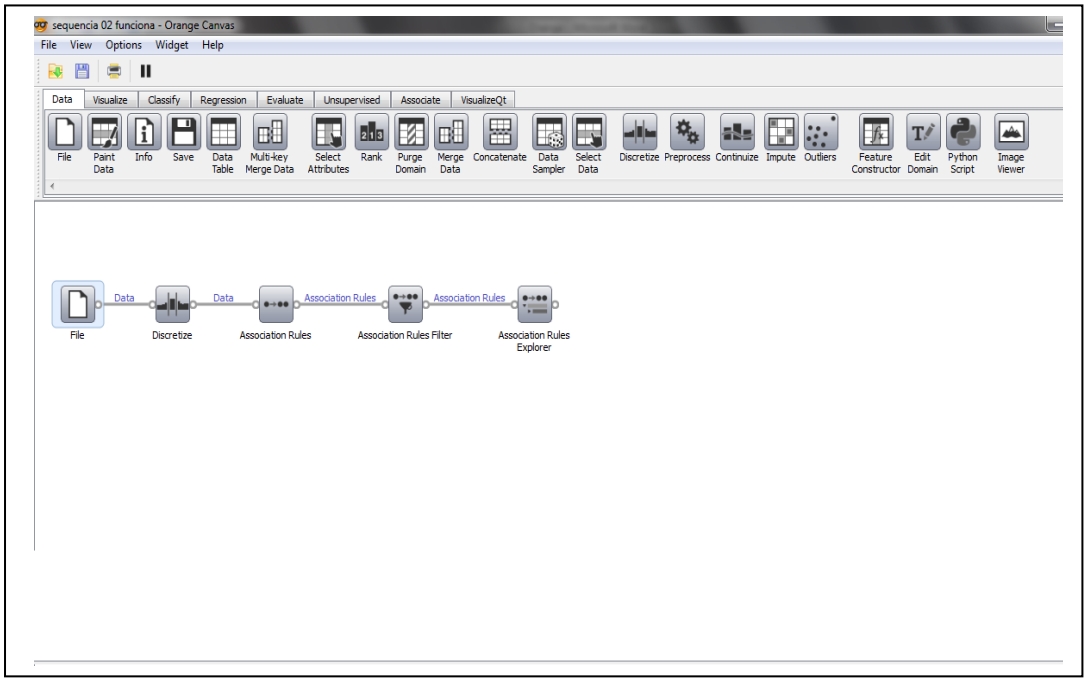


Figura 7 - Esquema que explora instâncias de dados no *Orange.*

A Figura 7 mostra um esquema de exploração. Enquanto a filosofia geral do *Orange* é que *widget* deveria ser simples, o poder da ferramenta deriva das diferentes formas de conectá-los.

Nesse estudo, foi utilizada a versão 2.6.1, disponível em http://orange.biolab.si.

4.4.4 *R-Project – arules*

Foi utilizada a versão 2.15.3 do *R-Project*. Pode ser obtido em: www.r-project.org.

O *R* é um *software* livre para computação estatística e construção de gráficos que pode ser baixado e distribuído gratuitamente de acordo com a licença GNU, sendo, atualmente, um programa multiuso. É uma ferramenta de linha de comando (com interface gráfica disponível) que possui linguagem própria para criar a interface entre o comando e a sua execução, por isso também pode ser utilizado com *script*. Pode ler um arquivo e, internamente, resolver como executá-lo e quais programas utilizar.

O *R* Funciona com um sistema de pacotes ("*library*" ou "*package*"). Qualquer pessoa pode fazer um pacote para ele e liberá-lo para *download*, automatizando quase todo tipo de tarefa. O ambiente de trabalho pode ser visualizado na Figura 8. Como a comunidade é muito grande, há pacotes para muitos tipos de problemas.

Ao trabalhar com um processo dentro do *R*, pode-se utilizar o resultado obtido em um pacote como entrada para outro pacote. Está disponível para as plataformas *UNIX*, *Windows* e *MacOS*.

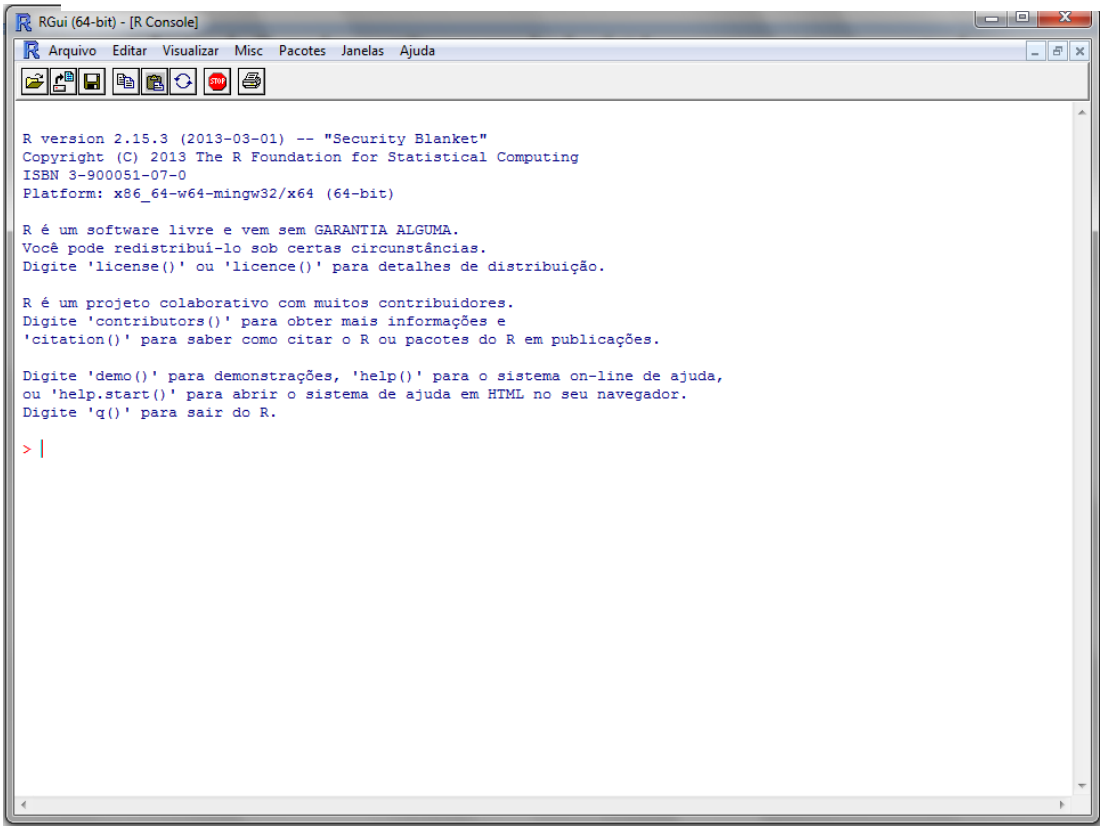


Figura 8 - Ambiente de trabalho do *R-Project.*

As análises feitas no *R* são digitadas diretamente na linha de comandos, na qual são digitados os comandos e funções que se deseja utilizar. Uma das maiores potencialidades e virtudes do *R* é a sua programação. É um programa leve (ocupa pouco espaço em memória) e, geralmente, é rápido, até em computadores com *hardwares* mais antigos. Isso ocorre porque, ao instalar o *R*, apenas as configurações mínimas para

seu funcionamento básico são instaladas. Para realizar tarefas mais complexas, pode ser necessário instalar pacotes adicionais. O *workspace* é sua área de trabalho do *R*.

Uma maneira que aperfeiçoa o uso do *R* e que poupa tempo é usar um *script* (um arquivo .txt) para executar as tarefas desejadas. Neste caso, os comandos não são digitados diretamente na linha de *prompt,* mas em um editor de texto (por exemplo: *Bloco de notas*). *Script* do *R* é apenas um "arquivo.R", no qual se digitam todos os comandos, sendo facilmente alterado, para cada tarefa.

Arules é uma extensão para *R* que fornece a infraestrutura necessária para criar e manipular o conjunto de dados de entrada para o algoritmo de mineração, e analisar os resultados de regras de associação e *itemsets*. A infraestrutura fornecida pelo pacote foi também criada para *interfacear* novos algoritmos, e para adicionar novos tipos de medidas de interesse e associações.

4.5 Exploração da base de dados

Procurou-se encontrar regras com os atributos da base de dados criada que implicassem em um dos possíveis diagnósticos disponíveis no sistema Tele-ECG.

4.5.1 Construção da base de dados

A etapa de *data mining* consiste em aplicar o algoritmo escolhido para explorar a base de dados com o intuito de encontrar padrões que possam resultar em conhecimento util. Nessa etapa, é importante assegurar que a base de dados seja cuidadosamente explorada para extrair o conhecimento embutido nos dados em um processo exaustivo. Nesse sentido, o algoritmo

foi executado várias vezes nas bases de dados, alterando-se os parâmetros de configuração até que se encontrem padrões desejados.

Para a montagem da base de dados de trabalho, as informações foram obtidas, inicialmente, por meio de uma cópia do banco de dados do Tele-ECG. A base de dados está originalmente dividida em tabelas com as informações do traçado do ECG, assim como dados relativos aos pacientes os quais realizaram os exames. Essas tabelas contêm dados relativos ao paciente, tais como:

a) número do laudo que identifica o exame de maneira exclusiva;
b) número do equipamento que realizou o exame;
c) identificação do paciente;
d) sexo do paciente;
e) idade completa do paciente, em anos;
f) peso, em kg do paciente;
g) altura, em cm do paciente;
h) um campo identifica o hospital em que foi realizado o exame;
i) ganho do aparelho que realizou o ECG;
j) velocidade de deslocamento do papel de registro do ECG;
k) filtro definido no aparelho de ECG;
l) um campo de observação no qual estão anotadas medidas do eletrocardiograma, as quais são realizadas com o auxílio do *software* de apoio do sistema Tele-ECG, ou seja;
    l.1) frequência cardíaca em bpm (FC);
    l.2) largura do pulso da onda P, em segundos (P);
    l.3) intervalo de tempo PR, em segundos (PRi);
    l.4) intervalo de tempo do complexo QRS, em segundos;
    l.5) intervalo de tempo QT, em segundos;
    l.6) valor QTc calculado em segundos;
    l.7) intervalo de tempo RR, em segundos;
    l.8) intervalo de tempo do complexo QRS, em segundos;
    l.9) intervalo de tempo, da onda T, em segundos;

l.10) valor da medida dos ângulos dos eixos elétricos;

l.10.1) valor do ângulo da onda P, em graus (sap);

l.10.2) valor do ângulo da onda T, em graus (sat);

l.10.3) valor do ângulo da onda QRS, em graus (saqrs);

l.11) identificação do médico que realizou o laudo desse exame;

l.12) códigos de medicamentos que o paciente relatou estar utilizando;

l.13) data de realização do exame;

l.14) data de alteração (caso haja alteração do laudo);

l.15) um campo de texto livre, no qual o técnico insere os dados relativos à descrição do paciente sobre seu estado de saúde.

As demais tabelas do sistema contêm os códigos dos medicamentos e dos possíveis laudos, além de dados de uso do sistema.

Esse conjunto de tabelas foi extraído com a ajuda do *software* Microsoft Access 2003. Com a base de dados aberta, as informações foram separadas, aglutinadas e esse conjunto convertido pelo *Access* para o formato “xls”, para posterior processamento com o *software Microsoft Excel* 2003.

Para tratar a base de dados no *Excel*, deve-se limitar o número de entradas (exames de eletrocardiograma, também denominados registros na base de dados) para o valor máximo de 65.536 (linhas de arquivo), pois esse é o limite imposto pelo próprio programa.

Foram amostrados do banco de dados original, aproximadamente, 46.000 registros de exames de eletrocardiograma, considerado um subconjunto representativo de toda a base de dados do sistema Tele-ECG.

Para uso das ferramentas de *data mining,* cada coluna da base de dados no *Excel* deve representar somente um atributo para cada exame de eletrocardiograma, no qual a ocorrência é preenchida com sim (Y na base de dados) e a sua ausência foi deixada em branco.

Posteriormente, a ausência de um atributo foi preenchida, na base de dados, por um sinal de “?”, para que a ferramenta de *data mining* interprete realmente como ausência, pois, caso contrário, seria contabilizada na geração de regras de associação, o que não é desejável.

A seleção dos registros que irão compor a amostra foi realizada por sorteio simples baseado em tabela de números causais[31]. Esse conjunto de dados representa a base de dados inicial.

Essa base de dados amostrada foi tratada com o auxílio do *software Excel*. Inicialmente, os campos de registros (atributos) foram separados em colunas. Posteriormente, as colunas foram tratadas separadamente para validar os registros.

A primeira coluna escolhida foi “dados clínicos”. Este campo é constituído por texto livre e, por isso, é difícil separá-la por qualquer método automático. Foi realizada a separação manual das informações (atributos) nele contidas. Seus dados foram realocados em outras colunas separadas das demais no *Excel*.

A coluna “dados clínicos” é relatada pelo paciente e é inserida pelo técnico que realiza o procedimento, na base de dados original do Tele-ECG, ou seja, não representa o resultado de avaliação médica. Esse campo foi manualmente analisado e dividido em colunas, as quais também compõem o banco de dados para fins desse.

Ao final da fase de separação de dados clínicos, 120 novas colunas foram adicionadas ao banco de dados inicial. Cada nova coluna representa um atributo do paciente, como, por exemplo, HAS (Hipertensão Arterial Sistêmica), etilismo, tabagismo, entre outros.

O segundo campo de registro escolhido foi “observações”. Nessa coluna do banco de dados do Tele-ECG, há composição de várias medidas aglutinadas relativas ao exame de eletrocardiograma, a saber: frequência cardíaca, intervalos relativos à onda P, ao complexo QRS, ao intervalo PR, ao intervalo RR, ao intervalo QT e ao valor de QTc (QT corrigido), além dos valores dos ângulos dos eixos elétricos sap, saqrs e sat, relativos à onda P, ao complexo QRS e a onda T.

Para a extração das informações com a utilização da ferramenta de *data mining*, cada valor foi separado em uma coluna independente na base de dados em formação, acrescentando 10 colunas em substituição à coluna de “observações” no banco de dados concluindo a segunda fase de pré-processamento.

A terceira coluna de atributos selecionada foi relativa aos diagnósticos na qual são visualizados os respectivos códigos das doenças e estes (códigos criados pelo sistema Tele-ECG) fazem parte da base de dados original sendo utilizados para identificar os diagnósticos realizados em cada eletrocardiograma.

Posteriormente, a base de dados foi cuidadosamente analisada e notou-se que um eletrocardiograma pode apresentar mais de um código de resultado. Na amostra coletada, observou-se que, em um determinado eletrocardiograma, podem existir até 12 códigos de “resultados” distintos.

As ferramentas de exploração pedem que, em cada coluna da base de dados, haja somente um atributo, e, para contornar essa situação, foi necessária, novamente, a separação dos diagnósticos em colunas diferentes, resultando em 143 novas colunas de diagnósticos. Novamente, o símbolo “Y” indicou a presença do respectivo diagnóstico na respectiva coluna.

Para próxima etapa, a limpeza dos dados, inicialmente, foram removidas todas as linhas com dados de ECG que não representaram efetivamente um eletrocardiograma válido (durante a fase de implantação do sistema Tele-ECG alguns registros de dados continham a palavra “teste”, utilizadas somente nos testes de implantação do sistema). Também foram removidos atributos e símbolos não pertinentes ao escopo desse trabalho, como pode ser visto no Apêndice C.

Continuando a limpeza de dados, valores inconsistentes, tais como idade superior a 105 anos, altura superior a 2,5 metros, altura inferior a 30 cm ou peso do paciente superior a 500 kg, foram considerados como erro de digitação ou, talvez, inserção em campo incorreto, não importando qual for o

caso, esses registros foram removidos, pois não correspondem a elementos confiáveis para a busca de padrões.

Para os parâmetros médicos, um especialista foi consultado para definir as faixas de valores aceitáveis, por exemplo, foram aceitos os valores de frequência cardíaca entre 1 bpm e 300 bpm, dentre outros. Para cada valor inconsistente, todo o registro foi removido da base de dados. O Apêndice D apresenta o conjunto de regras utilizadas nesse processo de seleção de atributos.

Com a intenção de enriquecer a informação para cada registro, foi calculado e adicionado à base de dados em processo o valor do *imc* (peso em kg/[altura em metros]^2, calculado por meio de uma fórmula inserida dentro do programa *Excel*). Na análise de consistência para esse valor, adotou-se a faixa de 9 kg/m^2 (inclusive) até 80 kg/m^2 (inclusive) para o *imc*, invalidando e, consequentemente, excluindo os demais registros da base de dados sob análise.

Para o processo de exploração, foram selecionados somente os atributos pertinentes às doenças cardiovasculares, auxiliado pelo especialista na área de domínio, eliminando-se 27 atributos do total amostrado (no Apêndice E, pode-se ver a lista dos itens excluídos). Assim, a preparação da base de dados foi considerada concluída.

Após a separação dos atributos em colunas, o arquivo resultante apresentou 278 colunas que representam o banco de dados a ser utilizado para a exploração efetivamente.

Após essa limpeza e seleção dos atributos a serem utilizados na exploração da base de dados, restaram 24.012 eletrocardiogramas e foi denominada por “base 1”. Todas as células em branco foram preenchidas manualmente com o sinal de “?”, conforme justificado anteriormente. No Apêndice F, pode-se ver uma amostra da base de dados pronta.

O banco de dados ficou pronto para ser convertido para o formato de dados exigido pelas ferramentas *Weka*, *Orange* e *R,* é o denominado formato de arquivo “.arff”.

O início da geração do arquivo.arff foi executado com auxílio do programa *Excel*. Após a base de dados estar verificada e pronta para as análises, foi salva no formato ".csv" (*Comma Separated Values* – Valores Separados por Vírgulas). O resultado desse arquivo salvo pode ser aberto com o programa "Bloco de Notas" do W*indows* (nesse caso, *Windows 7*), pois é arquivo do tipo não documento, ou seja, não há formatações, seu conteúdo é puramente composto pelos dados, e isso é necessário para o correto trabalho das ferramentas de mineração de dados.

Após o processo de correções no arquivo ".csv", executado com o apoio de um programa de edição hexadecimal, (pois pareceu mais simples e rápido corrigir as falhas geradas no processo de salvamento pelo programa *Excel*. Essas falhas ocorrem devido à inserção de espaços em branco na base de dados original e troca de "," por ";"). Para esse propósito, foi utilizado o programa *Hexeditor*.

Para completar a conversão para o formato ".arff", foi necessário incluir as *diretivas* utilizadas pelas ferramentas de exploração. Utilizando o programa "Bloco de Notas", foram inseridas todas as *diretivas* necessárias.

A primeira *diretiva* contém o nome da base de dados precedido pelo símbolo "@". Nas linhas que se sucedem, é definido cada atributo (um por linha), na mesma sequência em que foram utilizados no banco de dados. Cada linha inicia com o símbolo @*attribute*, seguida por um espaço em branco, o nome e o tipo do atributo, podendo ser numérico ou nominal, com sua lista de possibilidades ou sim/não (Y/N), entre chaves.

Na sequência, em nova linha, deve ser inserido o comando @*data* para definir o início da relação das linhas com os dados e todos os seus respectivos valores de atributos relativos aos exames de ECG.

Nas linhas de dados do arquivo.arff, devem ser listados os atributos na mesma sequência em que foram relacionados na base de dados, definidos anteriormente, e separados por vírgula. No Apêndice G, há um exemplo de arquivo ".arff" para melhor visualização e compreensão.

Dessa maneira, a base de dados foi manualmente ajustada para atender todos os requisitos das ferramentas de *data mining.*

Com o arquivo convertido, pode-se iniciar a exploração da base de dados. As ferramentas de mineração de dados possibilitam o uso de vários tipos de algoritmos para serem executados com diferentes parâmetros em cada um deles.

O processo de iteração na busca de padrões foi realizado alterando-se os parâmetros oferecidos pelas ferramentas para o algoritmo escolhido no caso denominado *apriori* (no Apêndice B, pode-se ver maiores detalhes).

O modo escolhido para utilização da ferramenta Weka foi o "*Explorer*", escolhido após a inicialização do programa, no menu principal da ferramenta.

O algoritmo para regras de associação escolhido não foi desenvolvido para trabalhar com valores numéricos, no entanto, alguns atributos da base de dados estudada nesse trabalho são desse tipo, a saber: idade, peso, altura, *imc*, sap, sat, saqrs, FC, P, PRi, QRS, QT, QTc e RR.

Para preparar os atributos numéricos, foi utilizado o filtro denominado *discretize*, da ferramenta *Weka* a fim de transformá-los em faixas de valores. Após aplicar esse filtro, o algoritmo *apriori* pode ser executado pela ferramenta *Weka*.

O processo de ajuste e exploração da base de dados foi feito com o apoio de programas de computador, no entanto, uma parcela significativa do tratamento foi feita de forma manual, o que levou cerca de 1.700 horas de trabalho para preparar a base de dados para a exploração (base 1). A coleta de dados, nessa base de dados, foi realizada exclusivamente pelo autor, por possuir o conhecimento necessário sobre a base de dados e o processo de KDD necessários.

Na ferramenta *Weka*, foi escolhido o algoritmo para encontrar regras de associação, assim como os parâmetros desejados.

Dentre os esses parâmetros possíveis para o *Weka*, temos:

a) *lowerBoundMinSupport*: valor do *suporte* mínimo;
b) *metricType*: define os tipos de métricas permitidas pela ferramenta, ou seja, *lift*, *leverage*, *conviction* e *confidence*;
c) *minMetric*: considera somente válidas as regras com valor acima do definido nesse parâmetro;
   numRules: número máximo de regras a serem encontradas;
d) *upperBoundMinSupport*: inicia iterativamente, reduzido desse valor até o valor mínimo definido;
e) *delta*: iterativamente, diminui o *suporte* desse valor até ser atingido o valor máximo ou o número de regras de associação definidas.

Após explorar a base de dados, observou-se que há muitos resultados de diagnósticos normal e/ou ritmo sinusal nos padrões encontrados e, portanto, com o intuito de gerar novo conhecimento, a base de dados foi refeita com alterações. Assim, foram eliminados os registros nos quais o diagnóstico gerado foi exclusivamente normal, ritmo sinusal, ou combinação dos dois anteriores, gerando uma nova base de dados, denominada base 2.

Posteriormente, foram executadas novas explorações com a ferramenta *Weka* com o objetivo de encontrar outras regras de associação. Essa nova base de dados resultante contém 9.271 registros (dados de exames de eletrocardiograma).

Utiliza-se o termo exames de eletrocardiograma e não paciente, pois o mesmo paciente pode realizar mais de um exame de eletrocardiograma e compor mais de um registro na base de dados.

Na primeira fase desse estudo, 21 análises foram realizadas utilizando a ferramenta *Weka* obtendo-se 3500 regras de associação.

É comum gerar-se conjuntos muito grandes de regras de associação dificultando o pós-processamento. Por esse motivo, foi criada uma metodologia de compilação de resultados com finalidade de facilitar a

visualização e as avaliações (o Apêndice H ilustra esse tipo de compilação de resultados). A ferramenta *Weka* não gera gráfico para esse tipo de algoritmo.

Após pós-processamento das regras, não foram encontrados resultados considerados como conhecimento novo nas 3.500 regras de associação anteriores. Assim, novas explorações foram realizadas como descritas a partir desse ponto.

Em nova etapa, foi realizada a contagem de ocorrências em cada coluna na base de dados (base 1). Foram separadas as colunas nas quais o número de ocorrências de antecedentes contados foi superior a 100 (valor adotado). A intenção foi de explorar regras de associação com os atributos mais prováveis.

O resultado obtido pode ser visto na Tabela 1 abaixo:

Tabela 1 - Antecedentes mais frequentes na base de dados.

Item	Atributo	Nº de Ocorrências
1	ARRITMIA	104
2	DIABETE	845
3	DISPNEIA	125
4	PRECORDIALGIA	536
5	HAS	7810
6	PALPITAÇÃO	114
7	PERIÓDICO, CONTROLE	1207
8	PÓS OP	100
9	PRÉ OP	2447
10	FC > 100 bpm	457

Para cada item da Tabela 1, foi criada nova base de dados e preparada para *data mining*.

O processo de separação de cada item utilizou de filtro, no *Excel.* Por exemplo, no item 5 (HAS), aplicou-se o filtro e selecionou-se somente aquelas linhas de registro nas quais o valor de HAS é “Y”, eliminando-se as demais linhas.

A nova base de dados selecionada recebeu novo nome e foi aplicada a ferramenta *Weka* em nova exploração. Dessa forma, o item HAS foi prevalente esperando-se encontrar regras de associação com esse atributo.

Como os atributos PRE OP, POS OP e PERIÓDICO CONTROLE não podem ser considerados determinantes para doenças cardíacas, foram descartados e a exploração foi repetida para todos os demais itens pertencentes à Tabela 1.

Em cada caso, o processamento de conversão para o formato basededados.arff, para cada base de dados, foi realizado antes de se aplicar à ferramenta *Weka*.

Para explorar os diagnósticos (consequentes) que apresentam contagem de ocorrência na base de dados (base 1) superior a 800 (número adotado), foi montada nova base de dados para cada item da Tabela 2. As duas últimas (D220 e D223) linhas dessa tabela foram sugeridas pelo especialista médico.

Tabela 2 - Diagnósticos mais frequentes na base de dados.

Arquivo	Diagnóstico	Descrição	ocorrências
Sem uso	D263	Eletrocardiograma normal	15033
Sem uso	D130	Ritmo sinusal	22079
BD08.xlsx	D312	Ausência de dados clínicos	1886
BD09.xlsx	D224	Bloqueio divisional ântero-superior esquerdo	1414
BD10.xlsx	D305	Alterações morfológicas	1268
BD11.xlsx	D292	Alteração repolarização ventricular parede ântero-lateral	1200
BD12.xlsx	D190	Ruído de artefato	1093
BD13.xlsx	D139	Bradicardia sinusal	1064
BD14.xlsx	D290	Alteração repolarização ventricular parede inferior	993
BD15.xlsx	D214	Sobrecarga ventricular esquerda	809
BD16.xlsx	D220	Bloqueio de ramo direito	594
BD24.xlsx	D223	Bloqueio de ramo esquerdo	288

Com as bases de dados de diagnósticos montadas, novamente, a ferramenta *Weka* foi utilizada para explorar cada uma delas individualmente,

exceto para os diagnósticos D130, D263, D190, D305 e D312, pois está fora dos objetivos do trabalho (o Apêndice I mostra os diagnósticos excluídos).

Foi gerada uma nova base de dados para todos os itens das Tabelas 1 e 2 em arquivo. Cada uma dessas novas bases de dados foram preparadas para a exploração, convertendo-as para o formato ".arff". Esse processo foi realizado com o auxílio do programa *Excel*, Bloco de Notas do *Windows* e *Hexeditor*, e, posteriormente, explorados com a ferramenta *Weka*.

Com esses resultados obtidos dos atributos filtrados das Tabelas 1 e 2 pela ferramenta *Weka,* novamente, foi necessário pós-processamento dos resultados, separando as colunas antecedentes com os consequentes. A finalidade foi permitir o uso de filtros no *Excel* para facilitar o pós-processamento de forma manual.

Com o arquivo de resultados formatado, foi realizada a seleção das regras de associação antes de enviar para o especialista médico avaliá-las quanto à novidade e importância.

Foram utilizados os critérios a seguir para filtrar as regras antes de serem enviadas para o especialista médico. O objetivo é eliminar a(s) regra(s) que, provavelmente, são menos importantes para atender os objetivos do trabalho:

a) Apresentar um código de diagnóstico no antecedente;
b) Consequente apresentar D130 (Ritmo sinusal), D139 (Bradicardia sinusal), D263 (Eletrocardiograma normal) ou D305 (Alterações Morfológicas), pois, de acordo com o especialista médico, não definem diagnósticos desejados;
c) Consequente apresentar atributos diferentes de códigos de diagnósticos;
d) Atendidas todas as regras anteriores, foram escolhidas as regras com maior valor de *lift*(), sup() ou conf().

O especialista médico sugeriu pesquisar diagnósticos: bloqueio de ramo esquerdo e bloqueio de ramo direito. Por isso, foram criadas duas novas bases de dados para testar essas hipóteses.

Em nova exploração com a ferramenta *Weka*, foi definido novo conjunto de 10.000 regras, nas quais D130 e D263 foram eliminados. Em nenhuma regra, foi encontrado diagnóstico no consequente e, por isso, foram descartadas.

Em nova tentativa de encontrar regras significativas, com a ferramenta *Weka*, foram geradas 100.000 regras de associação. Selecionadas 82 delas, impondo no antecedente somente medidas do ECG (P, PRi, FC, QT, QTc, sap, sat, saqrs ou RR) e, posteriormente, enviadas para análise do especialista médico. A seleção foi baseada no valor do *lift*.

Utilizando o documento denominado Diretrizes Brasileira de ECG 2009[2] (DBECG), o qual contém as faixas utilizadas pelos médicos para definir os diagnósticos em eletrocardiogramas, foi montada uma nova base de dados *discretizada* manualmente. O Apêndice J mostra as faixas utilizadas nesse processo de *discretização*. Após a exploração com a ferramenta *Weka*, não foram encontradas regras interessantes (nenhum consequente constituído por diagnóstico) para enviar para o especialista na área de aplicação.

Continuando a exploração da base de dados original (base 1), utilizou-se uma nova ferramenta, denominada *Orange*. As regras com diagnóstico D130 (diagnóstico de ritmo sinusal), D263 (diagnóstico normal) ou combinação de ambas foram eliminadas no processo de exploração buscando explorar ainda mais a base de dados. Foram encontradas 136 regras interessantes e enviadas para análise.

O processo de exploração prevê a iteração exaustiva até se obter os resultados desejados e, assim, uma nova tentativa foi executada para explorar a base de dados com a ferramenta *R* (com o pacote *arules*) para gerar novas regras interessantes. Foram encontradas 1.075 regras de associação. Após a seleção dessas regras, foram separadas 543 regras de associação para nova análise do especialista médico.

Com a ferramenta *R*, nova exploração com os atributos *discretizados* manualmente foi realizada, obtendo-se 7.121 regras de associação. Foi realizada nova seleção de regras nesse conjunto baseada no valor do *lift,* separando 137 regras enviadas para análise do especialista médico.

Procurando reavaliar as regras encontradas (no último conjunto de regras encontradas pelo *R*, ou seja, 7.121 regras de associação), foi realizada de forma manual a separação de novas regras de associação mais genéricas, logo mais úteis e importantes, utilizando o seguinte procedimento:

a) Foi criada nova planilha *Excel*, na qual são anotadas as regras selecionadas;
b) Utilizando o filtro em todas as colunas buscando os atributos comuns para um determinado código de diagnóstico;
c) Posteriormente, em análise visual, foram excluídas as regras redundantes.

Foram encontradas 7 regras gerais, dentre cinco diagnósticos diferentes (D139, D175, D220, D223 e D224), e também foram enviadas para a análise do especialista médico.

5 RESULTADOS E DISCUSSÃO

5 RESULTADOS E DISCUSSÃO

A descrição dos resultados a ser utilizada adota sequência cronológica de valores obtidos.

A base de dados estabelecida após o pré-processamento resultou em 24.012 exames de eletrocardiogramas (base 1). Essa base foi composta, em sua maioria, por informações de exames de eletrocardiograma com diagnóstico normal, ritmo sinusal, ou combinação de ambos (62%). A consequência dessa característica foi a prevalência desses diagnósticos nos resultados e, consequentemente, nas regras de associação geradas pelo processo de *data mining*. Por esse motivo, essa base de dados, posteriormente, foi filtrada a fim de excluir os registros com esses diagnósticos e, assim, gerar uma nova base de dados denominada “base 2”, com 9.217 exames.

5.1 Exploração inicial com a ferramenta *Weka*

Após preparadas, as duas bases de dados foram exploradas com o pacote *Weka* para obter os padrões escondidos, utilizando o algoritmo “*apriori*”.

Nesse estudo, a *discretização* geral dos dados foi executada por tentativa e erro, equalizando os valores em cada faixa criada, definindo o valor de *bins* (*bin* é um parâmetro do *Weka,* mais detalhes no Apêndice K) para todos os atributos numéricos, ou seja, idade, altura, peso, FC, P, PRi, sap, sat, saqrs, *imc*, QRS, QT, QTc e RR.

Foram, inicialmente, realizadas 21 aplicações de *data mining*, com o algoritmo *apriori*, variando-se os parâmetros de mineração. Essas análises estão resumidas na Tabela 3 a seguir:

Tabela 3 - Resumo dos resultados com *Weka.*

Análise	Nº de regras geradas	Nº *bins*	*minsup*	Métrica	Base de dados
1	100	14	0.5	*Confiança* = 0.9	Base 1
2	100	14	0.55	*Confiança* = 0.8	Base 1
3	200	14	0.4	*Confiança* = 0.9	Base 1
4	200	14	0.45	*Convicção* = 1.1	Base 1
5	200	14	0.5	*Confiança* = 0.8	Base 1
6	200	14	0.55	*Lift* = 1.0	Base 1
7	200	14	0.4	*Confiança* = 0.8	Base 1
8	100	14	0.6	*Lift* = 1.0	Base 1
9	100	14	0.1	*Leverage* = 0.1	Base 1
10	100	14	0.5	*Convicção* = 1.1	Base 1
11	200	14	0.45	*Confiança* = 0.7	Base 1
12	200	5	0.75	*Confiança* = 0.7	Base 1
13	200	14	0.4	*Confiança* = 0.9	Base 1
14	100	14	0.15	*Confiança* = 0.9	Base 2
15	100	14	0.2	*Confiança* = 0.8	Base 2
16	200	14	0.15	*Confiança* =0.9	Base 2
17	200	14	0.15	*Convicção* =1.1	Base 2
18	200	14	0.15	*Lift* = 1.1	Base 2
19	200	14	0.15	*Confiança* = 0.8	Base 2
20	200	14	0.45	*Confiança* =0.9	Base 2
21	200	14	0.4	*Confiança* = 0.6	Base 2
total regras	3500				

Como se pode ver na Tabela 3, o conjunto de regras de associação geradas foi de 3.500 pelo pacote de *software Weka* (um exemplo desse tipo de resultado pode ser visualizado no Apêndice L).

Como a ferramenta *Weka* não possui visualização gráfica para esse algoritmo, adotou-se, nesse trabalho, resumir os resultados em planilha *Excel* para as regras de associação geradas, com a finalidade de facilitar a visualização, já que o número de regras foi grande. Como foram realizadas 21 análises de *data mining*, o mesmo número de planilhas de resumos foi construído para extrair os padrões.

Essa mesma metodologia foi aplicada na base de dados completa (base 1) e, também, para a base de dados filtrada (base 2).

Esse conjunto de regras de associação foi filtrado para posterior análise do especialista médico, com o objetivo de separar aquelas que serão úteis e importantes em aplicações futuras. Foram separadas 1.732 regras de associação e enviadas para a apreciação do especialista médico.

Não foram encontradas regras de associação consideradas novas por esse especialista. Assim, o estudo de mineração de dados prosseguiu buscando por novas estratégias de exploração.

5.2 Exploração baseada nos atributos mais frequentes

Uma nova maneira utilizada para encontrar padrões foi utilizar os antecedentes com frequência maior que 100 ocorrências na base de dados (base 1). O resultado dessa contagem foi extraído da Tabela 1 e pode ser visto na Tabela 4 a seguir.

Esse procedimento foi executado e os resultados podem ser visualizados na Tabela 5.

Tabela 4 - Antecedentes explorados.

ITEM	ATRIBUTO	OCORRÊNCIAS
1	ARRITMIA	104
2	DIABETE	845
3	DISPNEIA	125
4	PRECORDIALGIA	536
5	HAS	7810
6	FC > 100 bpm	457
7	PALPITAÇÃO	114

A Tabela 5 abaixo resume os resultados parciais encontrados para esse procedimento acima descrito:

Tabela 5 - Resumo dos resultados para exploração de antecedentes

Nº	Atributo	Nº regras	Nº *bin*	sup	conf	Nº instâncias
1	Arritmia	2761	10	0,1	0,9	104
2	Arritmia	1381	14	0,1	0,9	104
3	Arritmia	10.000	5	0,1	0,9	104
4	Diabetes	6135	14	0,1	0,9	845
5	Diabetes	10.000	10	0,1	0,6	845
6	Diabetes	8291	14	0,1	0,6	845
7	Diabetes	10.000	5	0,25	0,6	845
8	Diabetes	10.000	10	0,1	0,9	845
9	Dispneia	3205	10	0,1	0,9	125
10	Dispneia	1562	14	0,1	0,9	125
11	Dispneia	10.000	5	0,15	0,9	125
12	FC > 100 bpm	902	10	0,1	0,9	420
13	FC > 100 bpm	376	14	0,1	0,9	420
14	FC > 100 bpm	5625	5	0,1	0,9	420
15	HAS	10.000	10	0,2	0,9	7810
16	HAS	10.000	5	0,4	0,9	7810
17	HAS	10.000	14	0,1	0,9	7810
18	Palpitação	2974	10	0,1	0,9	114
19	Palpitação	1713	14	0,1	0,9	114
20	Palpitação	10.000	5	0,1	0,9	114
21	Precordialgia	2501	10	0,1	0,9	536
22	Precordialgia	1226	14	0,1	0,9	536
23	Precordialgia	10.000	5	0,15	0,9	536
	Total regras	138.652				

Após o exame dos resultados, não foi encontrada regra na qual o consequente representasse um diagnóstico e, por isso, não foram enviadas para a análise do especialista médico.

Como as medidas obtidas no ECG (P, PRi, QT, QTc, QRS, FC, RR, sap, sat, saqrs) não fizeram parte das regras de associação encontradas, novas explorações foram realizadas aumentando o parâmetro que define o número máximo de regras a serem encontradas. A Tabela 6 lista as explorações executadas com esses valores:

Tabela 6 - Exploração com número de regras elevado

Nº	Nº regras	Nº bin	sup	conf	Nº instâncias
1	10.000	10	0,3	0,5	24.012
2	100.000	10	0,15	0,5	24.012
3	20.000	10	0,25	0,5	24.012
4	73.868	14	0,1	0,5	24.012
5	100.000	5	0,35	0,5	24.012
Total regras	303.868				

Analogamente ao que ocorreu com os atributos anteriores, nenhuma regra foi enviada para o especialista médico, pois também não foram geradas regras de associação nas quais o consequente foi composto por um diagnóstico.

5.3 Exploração de diagnósticos

Como não foram encontradas regras de associação com a ferramenta *Weka*, consideradas como novo conhecimento, então, novas bases de dados foram montadas utilizando os diagnósticos mais frequentes na base de dados completa (base 1). A Tabela 7 extraída da Tabela 2 mostra a contagem para os diagnósticos mais frequentes na base de dados.

Tabela 7 - Contagem de diagnósticos mais frequentes.

Item	Descrição / código do atributo	Ocorrências
1	Bradicardia sinusal / D139	1064
2	Sobrecarga ventricular esquerda / D214	809
3	Bloqueio divisional ântero-posterior esquerdo / D224	1414
4	Alteração repolarização ventricular parede inferior / D290	993
5	Alteração repolarização ventricular parede ântero-lateral / D292	1200
6	Alterações morfológicas / D305	1268
7	Ausência de dados clínicos / D312	1886
8	Ruído de artefato / D190	1093
	Total de ocorrências	41.545

As três últimas linhas da Tabela 7 (D305, D312 e D190) podem ser eliminadas, pois não definem diagnósticos.

Assim, foram explorados os demais diagnósticos conforme resumidos na Tabela 8 a seguir:

Tabela 8 - Diagnósticos explorados.

Nº	Atributo	Nº regras	Nº *bin*	*sup*	*conf*	Nº instancias
1	D139	723	10	0,1	0,5	1064
2	D214	3143	10	0,1	0,5	809
3	D224	10.000	10	0,1	0,5	1414
4	D290	5336	10	0,1	0,5	993
5	D292	6767	10	0,1	0,5	1200
	Total regras	25.969				

Desse conjunto, foram separadas 1.757 regras de associação e enviadas para o especialista médico a fim de avaliar o conhecimento encontrado.

Não foram encontradas regras novas nesse conjunto de regras de associação obtidas por esse processo, do ponto de vista do especialista médico.

Para investigar os diagnósticos bloqueio de ramo esquerdo (D223) e de ramo direito (D220), foram montadas duas novas bases de dados.

A Tabela 9 mostra o número de ocorrências para esses diagnósticos.

Tabela 9 - Ocorrências de bloqueios de ramos direito e esquerdo.

Item	Descrição / código do atributo	Ocorrências
1	Bloqueio de ramo direito / D220	594
2	Bloqueio de ramo esquerdo / D223	288

A Tabela 10 mostra os resultados da exploração para essas novas bases de dados.

Tabela 10 - Exploração para bloqueio de ramos direito e esquerdo.

Nº	Atributo	Nº regras	Nº *bin*	sup	conf	Nº instâncias
1	D220	2908	10	0,1	0,9	594
2	D223	1118	10	0,1	0,9	288

Apesar de algumas regras de associações encontradas serem válidas, nenhuma regra de associação foi considerada nova para aplicação médica.

5.4 Exploração com a ferramenta *Orange*

Para continuar buscando por novo conhecimento, o *Orange* foi utilizado na base 1 obtendo-se 136 regras de associação com *suporte* = 0,01 e *confiança* = 0,01. Para selecionar regras de associação, foi utilizada a métrica de interesse denominada *lift,* com valor mínimo 0,7. Desse total de regras de associação, foram selecionadas 35 regras, as quais foram enviadas para o especialista médico.

O especialista no domínio da aplicação considerou todas as regras de associação verdadeiras, apesar de não serem novas, exceto a regra de associação HAS → D139, pois ficou em dúvida quanto a sua validade.

De maneira geral, o especialista médico também concordou com a medida *lift* ser boa medida para apontar se uma regra de associação foi favorável ou contrária a um determinado diagnóstico. Houve concordância com o formato no qual as regras de associação foram apresentadas, e,

também, com os segundo e terceiro diagnósticos para o conjunto de atributos (esse formato pode ser visto no Apêndice M). Nesse formato, o mesmo atributo pode implicar em mais de um diagnóstico, com diferentes valores de *lift*.

5.5 Explorando com maior números de regras de associação.

Em nova exploração, foi definido gerar 10.000 regras de associação na ferramenta *Weka* e, para a métrica de interesse, foram escolhidos o *lift* e *confiança*. Eliminando-se as colunas referentes aos diagnósticos D130 (ritmo sinusal) e D263 (normal) para a base 1, foram obtidos os resultados listados na Tabela 11:

Tabela 11 - Exclusão de D130 e D263 com número elevado de regras.

Nº	Nº regras geradas	Nº *bin*	sup	métrica	Nº instâncias
1	1268	10	0,1	lift = 1,2	24.012
2	10.000	10	0,15	lift = 1,1	24.012
3	10.000	10	0,2	lift = 1,2	24.012
4	10.000	10	0,25	conf = 0,9	24.012
5	10.000	10	0,3	lift = 1,1	24.012
6	10.000	10	0,2	lift = 1,1	24.012

Não foram encontradas regras de associação nas quais o consequente foi representado por diagnóstico e, por isso, não foram enviadas para a análise do especialista médico.

5.6 Exploração com as medidas do ECG

Como nas regras de associação anteriores não ocorreram valores de medidas do ECG, nos antecedentes, foi selecionado um conjunto de 82

regras de associação a partir das regras encontradas na Tabela 8 e enviadas para análise.

Não foi encontrada nenhuma regra de associação nova na opinião do especialista médico.

5.7 Exploração alterando o processo de *discretização*.

Até esse ponto, as faixas utilizadas para *discretização* foram realizadas de forma automática pelas ferramentas de exploração (*Weka* e *Orange*). Então, em nova tentativa de se obter outras regras de associação relevantes, foi montada nova base de dados com os valores de faixas baseados na Diretriz Brasileira de ECG 2009 (DBECG)[2].

O resultado parcial obtido pode ser visualizado na Tabela 12 abaixo:

Tabela 12 - Faixas de medidas do ECG *discretizados* manualmente.

Nº regras	Nº *bin*	*sup*	*lift*	Nº instâncias
10.000	10	0,25	1,1	24.012

Não foi encontrada nenhuma regra na qual um código de diagnóstico estivesse presente (nem no antecedente, nem no consequente), tanto na ferramenta *Weka* quanto na ferramenta *Orange*, e, por isso, essas regras de associação foram descartadas.

5.8 Explorando com o *R*

Nesse ponto, haviam sido esgotadas as explorações com a base de dados com as ferramentas *Weka* e *Orange*, e, por isso, uma nova ferramenta de exploração foi adotada com a finalidade de encontrar novas regras de associação. Essa ferramenta foi o *R*.

Os resultados podem ser vistos na Tabela 13 abaixo:

Tabela 13 - Resultados obtidos utilizando-se o *R.*

Item	Nº regras	Nº *bin*	sup	conf	Nº instâncias
1	1074	*default*	0,01	0,01	24.012
2	7120	DBECG	0,01	0,01	24.012

Nesse caso, o número de *bin* foi definido pela ferramenta de exploração para o item 1, da Tabela 13, enquanto para o item 2 o valor de *bin* foi definido no *script* (o Apêndice N tem o *script* utilizado) que também coordenou a execução com o *R*. No primeiro caso, foram selecionadas 543 regras de associação nas quais o valor do *lift* foi o critério de seleção. Para o segundo caso, 137 regras de associação foram selecionadas, também utilizando o *lift* como métrica de interesse. A diferença entre esses dois casos foi que, no primeiro caso, a *discretização* realizada pela ferramenta e, no segundo caso, adotou-se a *discretização* de acordo com a DBECG[2].

Apesar de haver concordância para alguns casos, o especialista médico não encontrou regras de associação consideradas novas para a área médica.

5.9 Explorando no pós-processamento as regras de associação generalizadas.

Buscando obter um conhecimento mais generalizado e que reflita de fato o que está representado na base de dados (base 1), foi executada uma busca manual dentre as regras de associação encontradas (os dois conjuntos obtidos com o *R*), ou seja, com atributos comuns nas diferentes regras. Novamente, a planilha *Excel* foi utilizada para se encontrar os elementos comuns nas diversas regras de associação de um código de diagnóstico.

Foram selecionadas 7 regras de associação em mais um resultado parcial, e enviadas para análise pelo especialista médico que concordou com seis delas, mas não considerou como novo conhecimento.

5.10 Resumo de todas as explorações realizadas

A fim de facilitar a visualização dos resultados parciais encontrados ao longo do estudo, foi montada a Tabela 14 com o resumo geral desses resultados abaixo:

Tabela 14 - Resumo geral de todos os resultados.

Item	Atributo de interesse	Regras geradas	*bin*	Sup	Valor da métrica de interesse	Nº de instâncias
1	Todos	100	14	0.5	*Confiança*= 0.9	24.012
2	Todos	100	14	0.55	*Confiança*= 0.8	24.012
3	Todos	200	14	0.4	*Confiança*= 0.9	24.012
4	Todos	200	14	0.45	*Convicção*= 1.1	24.012
5	Todos	200	14	0.5	*Confiança*= 0.8	24.012
6	Todos	200	14	0.55	*lift* = 1.0	24.012
7	Todos	200	14	0.4	*Confiança*= 0.8	24.012
8	Todos	100	14	0.6	*lift* = 1.0	24.012
9	Todos	100	14	0.1	*Leverage* = 0.1	24.012
10	Todos	100	14	0.5	*Convicção*= 1.1	24.012
11	Todos	200	14	0.45	*Confiança*= 0.7	24.012
12	Todos	200	5	0.75	*Confiança*= 0.7	24.012
13	Todos	200	14	0.4	*Confiança*= 0.9	24.012
14	Todos	100	14	0.15	*Confiança*= 0.9	9217
15	Todos	100	14	0.2	*Confiança*= 0.8	9217
16	Todos	200	14	0.15	*Confiança* =0.9	9217
17	Todos	200	14	0.15	*Convicção* =1.1	9217
18	Todos	200	14	0.15	*lift* = 1.1	9217
19	Todos	200	14	0.15	*Confiança*= 0.8	9217
20	Todos	200	14	0.45	*Confiança* =0.9	9217
21	Todos	200	14	0.4	*Confiança*= 0.6	9217
22	Arritmia	2761	10	0,1	*Confiança*= 0,9	104
23	Arritmia	1381	14	0,1	*Confiança*= 0,9	104
24	Arritmia	10.000	5	0,1	*Confiança*= 0,9	104
25	Diabetes	6135	14	0,1	*Confiança*= 0,9	845
26	Diabetes	10.000	10	0,1	*Confiança*= 0,6	845
27	Diabetes	8291	14	0,1	*Confiança*= 0,6	845
28	Diabetes	10.000	5	0,25	*Confiança*= 0,6	845
29	Diabetes	10.000	10	0,1	*Confiança*= 0,9	845
30	Dispneia	3205	10	0,1	*Confiança*= 0,9	125
31	Dispneia	1562	14	0,1	*Confiança*= 0,9	125
32	Dispneia	10.000	5	0,15	*Confiança*= 0,9	125
33	FC > 100 bpm	902	10	0,1	*Confiança*= 0,9	420
34	FC > 100 bpm	376	14	0,1	*Confiança*= 0,9	420
35	FC > 100 bpm	5625	5	0,1	*Confiança*= 0,9	420
36	HAS	10.000	10	0,2	*Confiança*= 0,9	7810

(continua)

(conclusão)

Item	Atributo de interesse	Regras geradas	*bin*	Sup	Valor da métrica de interesse	Nº de instâncias
37	HAS	10.000	5	0,4	*Confiança*= 0,9	7810
38	HAS	10.000	14	0,1	*Confiança*= 0,9	7810
39	Palpitação	2974	10	0,1	*Confiança*= 0,9	114
40	Palpitação	1713	14	0,1	*Confiança*= 0,9	114
41	Palpitação	10.000	5	0,1	*Confiança*= 0,9	114
42	Precordialgia	2501	10	0,1	*Confiança*= 0,9	536
43	Precordialgia	1226	14	0,1	*Confiança*= 0,9	536
44	Precordialgia	10.000	5	0,15	*Confiança*= 0,9	536
45	Medidas do ECG	10.000	10	0,3	*Confiança*= 0,5	24.012
46	Medidas do ECG	100.000	10	0,15	*Confiança*= 0,5	24.012
47	Medidas do ECG	20.000	10	0,25	*Confiança*= 0,5	24.012
48	Medidas do ECG	73.868	14	0,1	*Confiança*= 0,5	24.012
49	Medidas do ECG	100.000	5	0,35	*Confiança*= 0,5	24.012
50	D139	723	10	0,1	*Confiança*=0,5	1064
51	D214	3143	10	0,1	*Confiança*=0,5	809
52	D224	10.000	10	0,1	*Confiança*=0,5	1414
53	D290	5336	10	0,1	*Confiança*=0,5	993
54	D292	6767	10	0,1	*Confiança*=0,5	1200
55	D220	2908	10	0,1	*Confiança*=0,9	594
56	D223	1118	10	0,1	*Confiança*=0,9	288
57	Todos	1268	10	0,1	*lift* = 1,2	24.012
58	Todos	10.000	10	0,15	*lift* = 1,1	24.012
59	Todos	10.000	10	0,2	*lift* = 1,2	24.012
60	Todos	10.000	10	0,25	*Confiança*= 0,9	24.012
61	Todos	10.000	10	0,3	*lift* = 1,1	24.012
62	Todos	10.000	10	0,2	*lift* = 1,1	24.012
63	Todos	10.000	10	0,25	*lift* = 1,1	24.012
64	Todos	1074	*default*	0,01	*Confiança*=0,01	24.012
65	Todos	7120	DBECG	0,01	*Confiança*=0,01	24.012
66	Todos	70	*default*	0,01	*Confiança*=0,01	24.012
	Total regras	545.547				

Na tabela 14, pode-se confirmar que o processo de exploração da base de dados (base 1) foi amplo, abordando diferentes alternativas de busca pelo conhecimento. Foram encontradas muitas relações entre os atributos disponíveis na área de Cardiologia.

6 CONCLUSÕES E DISCUSSÕES

6 CONCLUSÕES E DISCUSSÕES

O principal objetivo desse estudo foi explorar a base de dados de EGC formada com informações dos pacientes, medidas dos eletrocardiogramas e laudos gerados pelo grupo de especialistas do Tele-ECG do Instituto Dante Pazzanese de Cardiologia. Na busca do conhecimento, regras de associação foram geradas a partir de uma base de dados real, em um processo inovador, de forma ampla e podendo inspirar trabalhos futuros. Desse modo, conclui-se que:

A base de dados apresentou características próprias que serviram de orientação para todo o processo de exploração:

a) A base de dados é multirótulos (múltiplas saídas, ou seja, 143 diagnósticos diferentes, representados em uma saída diferente, na base de dados), isso é incomum na literatura sobre uso de *data mining* em bases de dados;
b) A base apresenta dados esparsos (93,6% das células sem valores de atributos), o que limita o processo de busca por conhecimento, já que há relativamente pouca informação disponível;
c) A base é desbalanceada (há muito mais exames de pacientes normais e com ritmo sinusal do que o restante, correspondendo a 62% do total de registros na base de dados), o que dificulta a busca por padrões novos, já que o algoritmo utilizado se baseia na frequência de ocorrência dos atributos presentes na base de dados;
d) Apenas nove atributos em antecedentes do total de 135 possíveis (FC, RR, PRi, QRS, saqrs, sap, P, QTc e HAS), no banco de dados, foram utilizados nas regras encontradas, ou seja, poucos atributos estão envolvidos na definição dos diagnósticos.

A falta de medidas de amplitude das ondas do sinal de ECG (T, QRS, P e também do segmento ST), a não utilização do traçado (impossibilitando a análise morfológica do eletrocardiograma), a total ausência de informações sobre uso de drogas pelo paciente (por exemplo, betabloqueadores diminuem a frequência cardíaca), a presença de diagnósticos raros (já que os resultados dependem implicitamente da frequência de ocorrência deles na base de dados), são vieses que dificultaram na descoberta de mais diagnósticos e, por isso, são fatores limitantes na busca do conhecimento.

Em alguns casos, não houve concordância, para determinados atributos que definem um diagnóstico, entre os diferentes especialistas médicos. Outras vezes, as definições utilizadas pelos médicos não foram suficientemente claras para um processo realizado por computador. Exemplificando, a classificação “idade escolar” pode não ser suficientemente clara, aumentando a complexidade para se encontrar resultado que depende desse fator.

A análise dos resultados depende da interpretação do especialista médico, utilizando-se de medidas subjetivas levando em consideração seu ponto de vista, sua experiência, seu conhecimento e suas preferências pessoais, para decidir sobre a importância das regras de associação em avaliação. Por exemplo, um médico que trabalha em uma região na qual há grande incidência para uma determinada enfermidade tende a pensar primeiro nas doenças em que está habituado a tratar.

Existe a possibilidade de perda de informações na seleção das regras de associação encontradas. Isso ocorre por falta de conhecimento na área médica do especialista em KDD, o que pode gerar seleção tendenciosa das regras de associação a serem submetidas ao especialista médico.

O sistema Tele-ECG apresenta uma falha importante quanto à “idade de bebês” utilizada para armazenar as informações na base de dados. Isso se deve à unidade utilizada para idade ser “*anos*”, o que, em alguns casos, não é apropriada. Por exemplo: um bebê de um mês é muito diferente de outro bebê de nove meses, mas o sistema não os diferencia.

A área de análise de ECG dispõe de muitos anos de estudo e experiência, e, por isso, há a dificuldade em encontrar regras novas para os especialistas médicos.

Informações relatadas pelos pacientes e inseridas por um técnico da coleta do ECG devem ser consideradas como problema crítico quando se deseja uma base de dados consistente.

O processo de *discretização* executado de forma automática pelas ferramentas de busca de regras de associação apresenta distorção. As faixas definidas podem separar os atributos de forma desbalanceada (são pacotes de extração de regras de associação fechados), implicando, em alguns casos, a dificuldade de se encontrar o conhecimento desejado.

O processo de descoberta realizado nessa pesquisa foi significativo, pois demonstrou a possibilidade de lidar com grande quantidade de informações de eletrocardiograma (mais de 545.000 regras de associação geradas), sem o uso do traçado do ECG, oferecendo nova abordagem, permitindo explorar diversos aspectos da base de dados.

Foi possível a obtenção de uma base de dados tratada com 24.012 registros de ECG, com dados confiáveis, caracterizando um resultado significativo.

Foram encontradas regras válidas, mas não foi encontrada regra nova e válida, na opinião dos especialistas médicos.

Dentre as medidas de interesse utilizadas ao longo do estudo, a medida de interesse *lift* demonstrou ser apropriada para mensurar os valores dos atribuídos, tanto no caso de reforçar o valor da regra quanto para o caso de afastar um diagnóstico.

A descoberta automática de regras de associação válidas, em uma base de dados esparsa e desbalanceada, demonstra o potencial dessa técnica.

Em toda essa tese, foram encontradas regras já consagradas pelos especialistas médicos na análise de eletrocardiogramas, confirmando a validade do KDD:

a) **fc(46.2bpm-48.1bpm) rr(1.228s a 1.306s)** → **D139:**

Indica que, quando a frequência cardíaca está entre “46,2 bpm e 48,1 bpm” e o intervalo RR está entre “1,228 segundos e 1,306 segundos”, acarreta em diagnóstico D139 (bradicardia sinusal). Não define patologia, é uma consequência matemática;

b) **PRi > 0,2s** → **D175:**

Informa que, para intervalos PR com valores maiores que 0,2 segundos, implica em diagnóstico D175 (Bloqueio atrioventricular do primeiro grau). Essa característica também pode ser induzida pelo uso de drogas (não consideradas nesse trabalho): Betabloqueadores de cálcio – amiodarona, digoxina e propafenona;

c) **qrs > 0,12s** → **D220:**

Significa que, para intervalos “QRS” com valores maiores que 0,12 segundos, implica em diagnóstico D220 (Bloqueio de ramo direito);

d) **qrs > 0,12s** → **D223:**

Pode ser entendido que intervalos “QRS” com valores maiores que 0,12 segundos implicam no diagnóstico D223 (Bloqueio de ramo esquerdo).

Para efetuar a diferenciação entre os dois resultados anteriores (nos itens “c” e “d”), deve ser utilizada a morfologia do sinal de ECG.

e) **saqrs < 45˚, qrs < 0,12s** → **D224:**

Pode ser entendido que ângulos do complexo “QRS” com valores menores que 45 graus e larguras do intervalo “QRS” menores que 0,12 segundos implicam em diagnóstico D224 (Bloqueio divisional ântero-superior esquerdo);

f) **HAS=Y → D224:**

Afirma que, se o paciente que realizou o ECG apresenta hipertensão, implica em diagnóstico D224 (Bloqueio divisional ântero-superior esquerdo);

g) **HAS=Y → D214:**

Informa que, se o paciente que realizou o ECG apresenta hipertensão, implica em diagnóstico D214 (Sobrecarga Ventricular Esquerda);

h) **saqrs(-48˚ a -33˚) → D214:**

Informa que, para valores de ângulos do complexo “QRS” entre -48 e -33 graus, implica em diagnóstico D214 (Sobrecarga Ventricular Esquerda);

6.1 Sugestões futuras

a) Criar consistência de verificação para os dados de entrada de ECG, ou seja, verificar se os dados são pertinentes a determinados campos de entrada, como, por exemplo: idade maior 105 anos pode ser aceita somente após confirmação. Esse procedimento pode melhorar a confiabilidade das informações armazenadas na base de dados;
b) Inserir função para medir a amplitude dos parâmetros do ECG (P, QRS, T, ST), enriquecendo as informações na base de dados;
c) Repetir o trabalho de tese, utilizando as drogas utilizadas pelos pacientes;
d) Para os pacientes com idade até um ano, alterar a unidade de tempo para *meses*, ao invés de *anos*.

7 REFERÊNCIAS

7 REFERÊNCIAS

1 Mansur AP, Favarato D. Mortalidade por doenças cardiovasculares no Brasil e na região metropolitana de São Paulo: atualização 2011. *Arq Bras Cardiol.* 2012;2:755-61.

2 Nicolau JC, Polanczyk CA, Pinho JA, Bacellar MSC, Ribeiro DGL, Darwich RN et al. Diretriz de interpretação de eletrocardiograma de repouso. Arquivos Brasileiros de Cardiologia, 2003;80(2):1-18.

3 Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: an overview. In: Fayyad U, Piatetsky-Shapiro G, Amith Smyth P, Uthurusamy R. (eds.). *Advances in knowledge discovery and data mining*, MIT Press, Cambridge, 1996, 1-36.

4 Engels R. *Planning tasks for knowledge discovery in databases; performing task-oriented user-guidance.* In Proc. of the 2nd Int. Conf. on KDD, 1996.

5 Carvalho VO. *Generalização de regras de associação utilizando conhecimento de domínio e avaliação do conhecimento generalizado* [Dissertação]. Universidade de São Paulo. Instituto de Ciências Matemáticas e de Computação; São Carlos, 2007.

6 Fujimoto ML. *Uma metodologia para exploração de regras de associação generalizadas integrando técnicas de visualização de informação de medidas de avaliação do conhecimento* [Dissertação]. Universidade de São Paulo. Instituto de Ciências Matemáticas e de Computação; São Carlos, 2008.

7 Witten IH, Frank E, Hall MA. *Data mining: practical machine learning tools and techniques.* 3rd ed. Burlington Morgan Kaufmann; 2011.

8 Navega S. *Princípios essenciais do data mining.* [cited 2013 Feb 2]. Available from: http://www.intelliwise.com/snavega.

9 Rezende SO. *Sistemas inteligentes. Fundamentos e aplicações.* Ed Manole. 2003.

10 Silva AP. *Geração de regras de associação quantitativas com intervalos não contínuos* [Dissertação]. Universidade Federal de Minas Gerais. Instituto de Ciências Exatas Departamento de Ciência da Computação; Belo Horizonte, 2004.

11 Geng L, Hamilton HJ. Interestingness measures for data mining: a survey. *ACM Computing Surveys*. 2006;38(3):1-32.

12 Zupan B, Demsar J. Open-source tools for data mining. *Clin Lab Med.* 2000;28:37-54.

13 Becher JD, Berkhin P, Freeman E. *Automating exploratory data analysis for efficient data mining.* KDD ACM, 2000. p. 424-9.

14 Park JS, Chen MS, Yu PS. An effective hash based algorithm for mining association rules. Proceeding SIGMOD '95 Proceedings of the 1995 ACM SIGMOD international conference on management of data p. 175-86.

15 Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT, Moser SA. Association rules and data mining in hospital infection control and public health surveillance. *J Am Med Inform Assoc.* 1998;5:373-81.

16 Yang WS, Hwang SY. A process-mining framework for the detection of healthcare fraud and. *Expert Systems with Applications.* 2006;31:56-68.

17 Romão W, Niederauer CAP, Martins A, Tcholakian A, Pacheco RCS, Barcia RM. Extração de regras de associação em C&T: o algoritmo apriori. In: XIX Encontro Nacional em Engenharia de Produção, 1999, Rio de Janeiro. XIX Encontro Nacional em Engenharia de Produção, 1999.

18 Konias S, Maglaveras N. A rule discovery algorithm appropriate for electrocardiograph signals. *Computers in Cardiology*. 2004;31:57-60.

19 Burn-Thornton KE, Edenbrandt L. Myocardial Infarction: in pointing the key indicators in the 12-lead ECG using data mining. *Computers and Biomedical Research.* 1998;31:293-303.

20 Murugan S, Radhakrishnan S. Rule based classification of ischemic ecg beats using antminer. *Int J Eng Sci Technol*. 2010;8:3929-35.

21 Ordonez C, Santana CA, Braal L de. Discovering interesting association rules in medical data. [cited 2013 Dec 30]. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.5403&rep=rep1&type=pdf.

22 Cabral AO, Rocha FJ. *Descoberta de padrões para a identificação de beneficiários com indicativos a infarto agudo do miocárdio* [trabalho]. Pontifícia Universidade Católica do Paraná; Curitiba, 2012.

23 Alizadehsania R, Habibia J, Hosseinia MJ, Mashayekhia H, Boghrati R, Ghandeharioun A, Bahadorian B, Sani ZA. A data mining approach for diagnosis of coronary artery disease. *Computer Methodos and Programs in Biomedicine III*. 2013;52-61.

24 Cavalcante PF. *A importância dos fatores de risco na obstrução das artérias coronárias utilizando técnicas de mineração de dados* [Dissertação]. Universidade Católica de Goiás. Pós Graduação em Ciências Ambientais e Saúde. Goiânia, 2009.

25 Nahar J, Imam T, Tickle KS Chen YPP. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*. 2013;1086-93.

26 Exarchos TP, Papaloukas C, Fotiadis DI. An Association Rule Mining-Based Methodology for Automated Detection of Ischemic ECG Beats. *IEEE Trans Biomed Eng.* 2006;53(8):1531-40.

27 Goldschmidt R, Passos E. *Data Mining: um guia prático – conceitos, técnicas, ferramentas, orientações.* Ed Campus Elsevier. 4ª ed. 2005. p. 261.

28 Boente ANP, Goldschmidt RR, Estrela VV. Uma Metodologia de suporte ao processo de descoberta de conhecimento em bases de dados. SEGeT – Simpósio de Excelência em Gestão e Tecnologia. [cited 2013 Dec 30]. Available from: http://s3.amazonaws.com/academia.edu.documents/30383697/seget2008kdd.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1388414708&Signature=0GzSNLySiF8OnF00Jk4oHDNW%2BNw%3D&response-content-disposition=inline.

29 Galvão ND, Marin HF. Data mining: a literature review. *Acta Paul Enferm.* 2009;22(5):686-90.

30 Zaki MJ, Parthasarathy S, Ogihara M. Parallel algorithms for discovery of association rules. *Data Mining and Knowledge Discovery.* 1997;1:343-73.

31 Haddad N. *Metodologia de estudos em Ciências da Saúde. Como planejar, analisar e apresentar um trabalho científico*. Roca; 2004. p.250-5.

32 Srikant R, Agrawal R. Mining generalized association rules. In: Proceedings of the 21st VLDB Conference Zurich, Swizerland, 1995.

33 Agrawal R, Imielinski T, Swami A. Mining association rules between set of itens in large databases. In: ACM SIGMOD INT'L Conference on Management of Data. Proceedings. Washington, 1993. p. 207-16.

34 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: 20th INT'L Conference on Very Large Databases, 1994. Proceedings. Santiago, 1994.

35 Motta CGL. *Metodologia para mineração de regras de associação multiníveis incluindo pré e pós-processamento* [Doutorado]. Universidade Federal do Rio de Janeiro, Programa de Pós-Graduação em Engenharia Civil. Rio de Janeiro, 2010.

APÊNDICES

APÊNDICES

APÊNDICE A – Discretização.

É comum a presença de atributos numéricos, em bases de dados o que obriga a manipulação deste tipo de dado. Porém, nem todas as técnicas de mineração de dados são capazes de trabalhar esses tipos de dados adequadamente, como é o caso do algoritmo *apriori*.

O tratamento de valores numéricos leva a divisão em intervalos representativos. Por isso é comum realizar esse procedimento para os atributos no pré-processamento, para obter melhores resultados.

Discretização é a codificação dos valores contínuos em intervalos discretos, os quais podem ser interpretados como um conjunto de valores. É uma forma de diminuir a quantidade de valores diferentes que um atributo pode apresentar.

Através da *discretização* convertem-se atributos numéricos em atributos categóricos. No entanto a qualidade dos resultados torna-se dependente da *discretização* dos dados.

Na *discretização,* um intervalo gerado funciona como conjunto para todos os elementos nele contidos e passa a apresentar uma representação comum.

Na geração de regras de associação, a definição de intervalos muito grandes pode gerar regras muito amplas que irão conter pouca informação.

Em outra situação intervalos muito pequenos podem não ser frequentes o suficiente para a identificação de padrões. Percebe-se que o número de elementos contidos em cada intervalo pode interferir na qualidade dos resultados.

Por exemplo, seja a função para cálculo de rendimento financeiro obtido de compras com cartão de crédito, dado pela equação representativa como a seguir: Rendimento(5000 - 10000) → Compras_com_cartão(1000 -

1800) [*sup* = 3%, *conf* = 85%], com o seguinte significado: dado um cliente com rendimento entre R$ 5.000,00 e R$ 10.000,00, existe a chance de 85% (*confiança*) de que ele gaste entre R$ 10.00,00 a R$ 1.800,00 de compras por mês, no cartão de crédito. Tem-se essa regra presente em 3% (*suporte*) das transações na base de dados.

APÊNDICE B - Algoritmo Apriori e métricas de interesse.

Motivação

Uma das maneiras objetivas de se encontrar conhecimento é através da descoberta de relações entre diferentes atributos. É uma técnica fácil de ser compreendida e caracterizada por apresentar tendências embutidas nos dados analisados pretendendo encontrar quais itens ou instâncias estão relacionadas, ou seja, ocorrem de forma conjunta em uma mesma transação.

Para a geração de regras de associação, o algoritmo *apriori* é considerado como o estado da arte. Ele pode trabalhar com grande número de atributos, gerando grande número de combinações de atributos e é considerado aprendizado não supervisionado.

Esse algoritmo realiza sua tarefa em duas fases. Na primeira etapa gera todas as combinações possíveis entre atributos e na segunda descarta as regras que não atendam os valores de *suporte* e *confiança* mínimos estipulados pelo analista de KDD.

Introduzido por Srikant[32], para um conjunto de transações em banco de dados, uma transação é um conjunto de itens na forma X (antecedente) $\rightarrow$ Y (consequente), informa que se X ocorre nessa transação, então há boa chance de ocorrer também o item Y.

Geralmente grande número de regras são encontradas pelos algoritmos e por isso medidas de interesse são utilizadas para reduzir o número de regras a serem avaliadas, facilitando a etapa de pós-processamento.

Para a geração do conjunto de regras, usualmente define-se valores limiares mínimos para os valores de *suporte* (*minsup*) e *confiança* (*minconf*) e dessa forma são utilizadas somente as regras que atendem a esses valores mínimos.

Algoritmo *apriori*

É importante entender alguns conceitos antes de tratar do algoritmo propriamente dito:

a) O modelo "cesta de mercado"

Para descrever o relacionamento entre dois objetos é comum se utilizar o denominado modelo de "cesta de mercado de dados". Nesse modelo há os itens e as cestas, também chamadas de transações. Cada cesta contém os itens comprados, por um cliente, em um mercado de compras, ou seja, há um conjunto de itens (*itemsets*). Pensando em termos de banco de dados pode-se atribuir cada cesta a um registro contendo o conjunto dos atributos.

b) *itemsets* frequentes

Com as informações de *itemsets* frequentes, por exemplo, um varejista pode compreender quais itens são consumidos conjuntamente. Isso é importante em pares ou conjuntos maiores de itens que ocorrem muito mais frequentemente do que esperado para os itens comprados individualmente.

Para esse estudo foi utilizada a idéia de *itemsets* frequentes com o interesse de se encontrar conjunto de itens (aqui atributos) que podem ocorrer juntos. Para as entradas, são esperados os atributos de ECG e para as saídas os diagnósticos. Encontrar um conjunto de *itemsets* frequentes, com dados dos eletrocardiogramas e diagnósticos associados, é considerado uma relação importante para a tomada de decisões futuras na análise de ECGs.

c) Regras de associação

As regras encontradas pelo algoritmo *apriori* são ordenadas nos resultados, de acordo com a medida de *confiança* encontrada para cada regra.

d) Monotonicidade dos *itemsets*

Os algoritmos de busca por *itemsets* frequentes se utilizam da propriedade da monotonicidade que afirma que “se um conjunto *I* de itens é frequente então cada subconjunto de *I* também é frequente. Se um *itemset* é frequente, então todos os seus subconjuntos também são frequentes. Ou, o *suporte* de um *itemset* nunca é maior que o *suporte* de seus subconjuntos.

e) Mineração de regras de associação

Poderíamos listar o conjunto de todas as possíveis regras de associação, calcular o *suporte* e a *confiança* de cada regra e podar aquelas que não atendem os limiares definidos, mas essa solução é inviável, pois é computacionalmente proibitiva.

Para o algoritmo *apriori* a monotonicidade permite compactar informações sobre *itemsets* frequentes. Para o limiar de *suporte* listam-se somente os *itemsets* frequentes nos quais todos os subconjuntos são frequentes. A monotonicidade diz-nos que, se houver uma tripla frequente, então existem três pares frequentes contidos nela. Pode haver pares frequentes contidos em uma tripla não frequente. Espera-se encontrar mais pares frequentes que triplas frequentes, mais triplas frequentes que quádruplas frequentes, e assim por diante.

O algoritmo *apriori* - descrição

Apriori é muito utilizado pelas comunidades de bancos de dados e aprendizado de máquina. Assume que os atributos nas bases de dados são categóricos e, portanto não pode ser aplicado a dados numéricos.

Descrevendo o algoritmo:

a) seja $k = 1$.

b) Obtenha conjuntos frequentes de tamanho 1.

c) Repita enquanto novos *itemsets* frequentes forem obtidos.

   c.1) Obtenha *itemsets* candidatos de tamanho $k+1$ a partir de *itemsets* de tamanho k.

c.2) Elimine *itemsets* candidatos contendo subconjuntos de tamanho *k* não frequentes.

c.3) Conte o *suporte* de cada candidato varrendo o banco de dados.

c.4) Elimine candidatos não frequentes deixando só os frequentes.

Definição formal do algoritmo

Formalizando a mineração de regras de associação.[29,33,34]

Seja a base de dados $T = \{t_1, t_2, ..., t_n\}$ contendo n transações.

Seja o conjunto de m itens $I = \{i_1, i_2, ..., i_m\}$ disponíveis para compor cada transação $t_i \in T$, tal que $t_i \subseteq I$.

Um conjunto de itens é denominado *itemset*. Se ele possuir *k* itens, então é um *k-itemset*.

Sejam *A* e *B* dois *itemsets*, tais que $A \subseteq I$ e $B \subseteq I$ e que não possuam itens em comum, ou seja, $A \cap B = \varphi$. Uma transação t_i contem o *itemset A* se e somente se $A \subseteq t_i$.

Regra de associação é uma implicação da forma: $A \rightarrow B$, na qual $A \subseteq I$, $B \subseteq I$ e $A \cap B = \varphi$.

Deve ser lida como: *A* implica em *B*, no qual *A* é chamado antecedente e *B* é o consequente da regra. Tanto o antecedente quanto o consequente podem ser formados por mais de um item.

A mineração de dados objetiva encontrar todas as regras que associem a presença de um *itemset A* com qualquer outro (*B, C …*), no conjunto de transações *T.* Como *I* conta com m itens, o espaço de busca para todas as regras é teoricamente 2^m, pois todos os itens podem constituir *itemsets*.

De fato nem todos os itens de *I* ocorrem nas transações de *T* e outros ocorrem em poucas transações. Essa dispersão de itens é usada durante a geração de regras de associação, tornando os métodos viáveis e eficientes.

A frequência de um *itemset* conhecida como *suporte*, denotada por *s*, é o número de transações em *T* que contêm este *itemset*.

O *suporte s* de uma regra de associação $A \rightarrow B$ é a porcentagem de transações que contêm $A \cup B$ (ambos A e B) em relação ao total de n transações de T. O *suporte* é a probabilidade de ocorrência do *itemset* $A \cup B$ em T.

O *suporte* indica a frequência relativa das regras e pode ser usado para compará-las, ou seja, regras com altos valores de *suporte* podem ser interessantes (uteis, ocorrem com frequência, são confiáveis e podem ser utilizadas para fazer previsões), por se distinguirem quantitativamente das demais. Por outro lado, as de baixo *suporte*, podem representar somente ocorrência ao acaso. *Suporte* representa a aplicabilidade da regra.

A *confiança c* de uma regra de associação, $A \rightarrow B$, é a porcentagem de transações que contêm $A \cup B$ com relação a todas as transações de T que contêm A.

Para uma regra apresentar qualquer interesse elevado, significa que a presença de A em uma cesta, de alguma maneira implica na presença de B, na mesma cesta, ou interesse negativo significando que a presença de A desencoraja a presença de B, na mesma cesta.

A *confiança* aponta a capacidade de predição das regras. Regras com altos valores de *confiança* se destacam qualitativamente das demais, pelo nível de certeza de ocorrência do consequente da regra, dado que o seu antecedente ocorre. Regras com baixa *confiança* não fornecem segurança de predição, são de uso limitado.

Regra de associação interessante é aquela com *suporte* e *confiança* maiores ou iguais, aos limites pré-estabelecidos de *minsup* e *minconf*. Esse modelo é denominado Modelo *Suporte*/*Confiança*.

Após a descoberta de regras de associação é comum apresentar as regras interessantes com o seguinte formato:

$A \rightarrow B$ [*suporte*, *confiança*]

Um *itemset* frequente é aquele cuja frequência é maior ou igual ao produto do *minsup* pelo total de transações de T.

Denota-se por C_k um conjunto de *k-itemsets* candidatos, denota-se por L_k um conjunto de *k-itemsets* frequentes.

Fases do algoritmo *apriori*

O objetivo de descobrir regras de associação pode ser decomposto em duas etapas:

a) Na fase 1 o algoritmo busca encontrar todos os conjuntos de itens com *suporte* maior que o mínimo estabelecido pelo analista de KDD. Os *itemsets* que atender essa exigência são chamados *itemsets* frequentes;
b) Na fase 2 o algoritmo utiliza os *itemsets* frequentes obtidos para selecionar as regras de associação finais.

Apriori encontra todos os conjuntos de itens frequentes, denominados *itemsets* frequentes (L_k) e faz uso de duas funções: a função *Apriori_gen*, para gerar os candidatos e eliminar os que não são frequentes e a função *Genrules* é utilizada para selecionar as regras de associação.

Inicialmente realiza a contagem de ocorrências dos itens a fim de determinar os *itemsets* frequentes de tamanho unitário (*1-itemsets* frequentes). Os *k* passos posteriores consistem-se das duas fases acima. Primeiro, os *itemsets* frequentes L_{k-1}, encontrados no passo anterior (*k-1*) são usados para gerar os conjuntos de itens potencialmente frequentes, os chamados *itemsets* candidatos (C_k).

Em seguida é realizada nova varredura contando o *suporte* de cada candidato em C_k.

Pode-se ver abaixo (Figura 9) um exemplo para compreender a geração de *itemsets* frequentes:

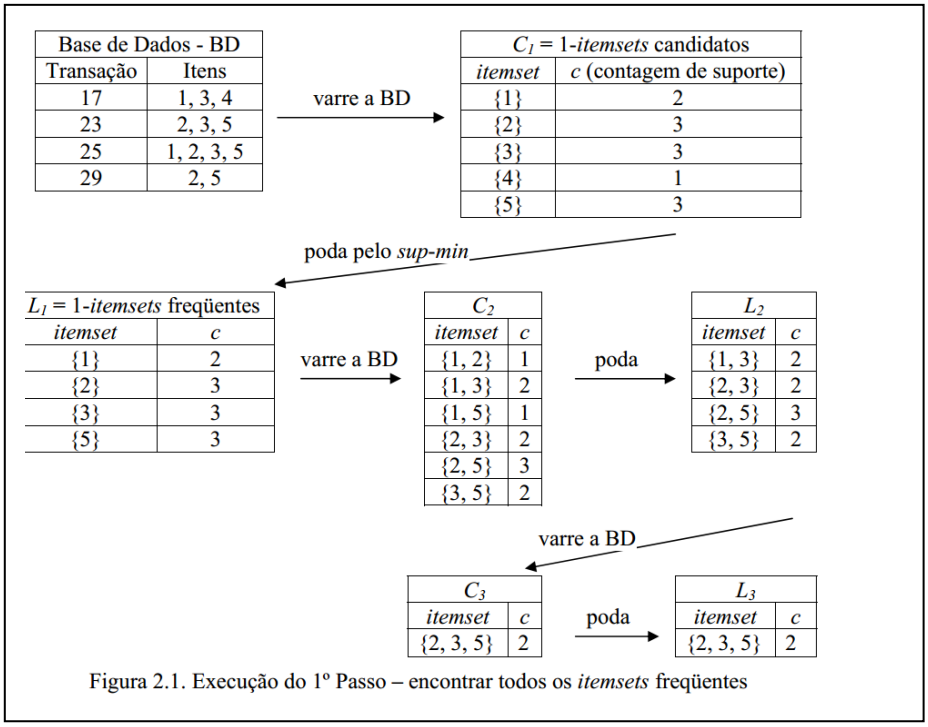


Figura 9 - Exemplo de obtenção de regras de associação. Modificado de Motta CGL, 2010[35].

Considerações Importantes

a) *minsup*

Usar o mesmo valor de *minsup* obriga o algoritmo assumir que todos os itens possuem frequências similares. Isso não é verdadeiro em muitos domínios, já que alguns itens aparecem frequentemente enquanto outros são raros.

b) *suporte* muito baixo

Alguns problemas podem exigir *suporte* mínimos muito baixos e.g. caviar → champanha.

c) Perda de regras

Suporte e *confiança* mínimas altas, podem perder regras interessantes.

d) Regras não relacionadas

Confiança pode atribuir alto interesse a regras não correlacionadas.

Por exemplo

i) Para um supermercado, clientes compram panelas e frigideiras muito menos frequentemente do que pão e leite.

ii) Em um hospital, pacientes com doenças de alta complexidade ocorrem com maior frequência do que os de baixa complexidade.

e) Para a escolha de *minsup*

Diminuir o valor de *minsup* acarreta em maior quantidade de *itemsets* frequentes o que pode aumentar o número de candidatos e a dimensão máxima dos *itemsets* frequentes (gerando regras com mais condições/conclusões).

f) Número de itens no conjunto de exemplos

i) Quanto mais itens, maior o espaço necessário para armazena-los.

ii) Se o número de *itemsets* frequentes também for grande, o tempo de computação tende a aumentar.

iii) O tempo de execução pode aumentar com o aumento do número de transações, uma vez que o algoritmo *apriori* efetua múltiplas varreduras.

g) Para grandes variações na frequência dos itens, há dois problemas:

i) Quando *minsup* apresenta valor alto às regras que envolvem itens raros não são encontradas. Para encontrar regras que envolvem itens frequentes e raros, *minsup* deve apresentar valor pequeno.

ii) Causar a geração muito grande de regras, já que os itens frequentes (*1-itemsets*) serão associados entre si de todas as maneiras possíveis, tornando menos eficaz à poda das regras.

Medidas de Interesse

Além de gerar grandes conjuntos de regras de associação grande parte dos resultados encontrados costuma ser composta por regras óbvias, redundantes e às vezes contraditórias.

A fim de auxiliar na avaliação da importância das regras, são utilizadas medidas de interesse que buscam indicar se a regra apresenta fatores desejáveis ou não.

Nos dois principais passos da geração de regras de associação percebe-se que *suporte* e *confiança* são as medidas de interesse aplicadas às regras de associação, representando propriedades diferentes. Outras medidas de interesse podem ser aplicadas para avaliar às regras de associação.

Para que se possa exemplificar cada métrica, será utilizada a seguinte base de dados denominada *I*: supondo um conjunto de transações *T* com 100 transações, com três *itemsets* I_1, I_2 e I_3 presentes em 20, 10 e 40 operações respectivamente e que I_1 e I_2 ocorrem juntos em 5 transações, I_1 e I_3 ocorrem simultâneos em 11 transações.

a) *Suporte*

O *suporte* representa o percentual de transações em que todos os itens contidos na regra estão presentes: $sup(I_1 \rightarrow I_2) = sup(I_2 \rightarrow I_1) = P(I_1, I_2)$.

Na base de dados *I* pode-se dizer que o *suporte* de $(I_1 \rightarrow I_2)$ é 5/100 = 0.05, ou seja, o *suporte* é de 5%.

b) *Confiança*

A *confiança* expressa à força da regra, ou a chance de acerto da regra, indicando a probabilidade do lado direito da regra ocorrer dado que o lado esquerdo da regra também ocorrer: $conf(I_1 \rightarrow I_2) = P(I_2 \mid I_1) = P(I_1, I_2) / P(I_1) = sup(I_1 \rightarrow I_2) / sup(I_1)$.

Utilizando I temos que $sup(I_1, I_2) = 0.05$ e $sup(I_1) = 0.2$, então $conf(I_1 \rightarrow I_2) = 0.05/0.2 = 0.25$, ou a regra $I_1 \rightarrow I_2$ apresenta *confiança* de 25%.

O *suporte* é utilizado para realizar podas na geração dos *itemsets* frequentes, já *confiança* é usada para "filtrar" as regras, permitindo somente as regras que possuem *confiança* superior ao limiar predefinido. Um dos problemas com a *confiança* é que ela é muito sensível com relação à frequência do lado direito da regra (I_2) porque um *suporte* muito alto de I_2 pode fazer com que a regra possua uma *confiança* alta, ainda se não houver uma associação entre os *itemsets* I_1 e I_2 da regra.

c) *Lift*

O *lift* (varia entre 0 e +∞) de uma regra de associação é a razão da *confiança* pelo percentual de transações cobertas pelo lado direito da regra. Dessa forma mostra o quão mais frequente é o lado direto da regra quando o lado esquerdo está presente: $Lift(I_1 \rightarrow I_2) = P(I_1, I_2)/(P(I_1)P(I_2)) = conf(I_1 \rightarrow I_2)/sup(I_2)$. Como vimos anteriormente $conf(I_1 \rightarrow I_2) = 0.25$, lembrando que $sup(I_2)=0.1$, o $lift(I_1 \rightarrow I_2) = 0.25/0.1 = 2.5$.

Quando o *lift* é maior que 1, o lado direito da regra ocorre com mais frequência, nas transações em que o lado esquerdo ocorre. Já quando é menor que 1, o lado direito é mais frequente nas transações em que o lado esquerdo não ocorre. Para o valor igual a 1, o lado direito ocorre com a mesma frequência independente do lado esquerdo ocorrer ou não. Assim as regras que possuem *lift* maior que 1 são mais interessantes que as demais e maior deverá ser a relação entre os dois lados da regra.

d) *Leverage*

O *leverage* (varia entre -0.25 e 0.25) representa o número transações adicionais cobertas pelos lados direito e esquerdo, além do esperado, caso os dois lados fossem independentes um do outro: $leverage(I_1 \rightarrow I_2) = P(I_1, I_2) - (P(I_1)P(I_2))$

Sabendo que $P(I_1, I_2) = 0.05$, $P(I_1) = 0.2$ e $P(I_2) = 0.1$. Pode-se calcular o $leverage(I_1 \rightarrow I_2) = (0.05)-(0.2 * 0.1) = (0.05 – 0.02) = 0.03$, que representa 30 transações.

Leverage maior que 0 indica que os dois lados da regra ocorrem juntos, em número de transações maior que o esperado, caso os itens encontrados nas regras fossem completamente independentes. Para o *leverage* menor que 0, os dois lados da regra ocorrem juntos, menos que o esperado. No *leverage* igual a 0, os dois lados da regra ocorrem juntos, exatamente o esperado, indicando que os dois lados provavelmente são independentes. Dessa maneira quanto maior o *leverage* mais interessante será a regra.

Um problema do *leverage* é não levar em consideração as proporções de uma regra para a outra. Supondo a regra $I_3 \rightarrow I_1$, teríamos o *leverage* ($I_3 \rightarrow I_1$) = 0.03, representando 30 transações, porém, na primeira regra $I_1 \rightarrow I_2$ isto é muito mais interessante, já que dizer que 30 transações ocorrem além do esperado, no conjunto de 50 transações, é mais significativo que 30 transações ocorrendo além do esperado em 110 transações. Por essa razão, quando este é positivo usa-se dividi-lo pelo *suporte* da regra. Assim pode-se obter o percentual do *suporte* que não ocorre por acaso e teríamos que para a regra $I_1 \rightarrow I_2$, 60% do seu *suporte* ocorre além do esperado, enquanto isso, apenas 27% do *suporte* da regra $I_3 \rightarrow I_1$ ocorrem além do esperado.

e) *Convicção*

A *convicção* (varia entre 0.5 e $+\infty$ e é direcional, isto é, conv($A \rightarrow C$) $\neq$ conv($C \rightarrow A$)) parte da ideia de que logicamente $I_1 \rightarrow I_2$ pode ser reescrito como $\neg(I_1 \wedge \neg I_2)$, então a *convicção* verifica o quanto ($I_1 \wedge \neg I_2$) está distante da independência:

$$\text{conv}(I_1 \rightarrow I_2) = P(I_1)P(\neg I_2)/P(I_1, \neg I_2)=(1\text{-sup}(I_2)/(1\text{-conf}(I_1 \rightarrow I_2))$$

Pode-se calcular conv($I_1 \rightarrow I_2$) = (1 – 0,1) / (1-0.25) = 3.6. Diferente da *confiança*, a *convicção* apresenta valor 1 quando os *itemsets* da regra não possuem nenhuma relação e quanto maior a *convicção,* maior a relação entre I_1 e I_2, já quando o valor é menor que 1 a relação entre os itens é negativa, ou seja, quando I_1 ocorre, I_2 tende a não ocorrer.

Um dos objetivos desta métrica é resolver uma falha da *confiança*, encontrando a relação entre dois itens diferentes na qual somente um deles

apresenta frequência alta. Nesse caso o item de alta frequência poderia mascarar a relação quando olhássemos para a *confiança*.

APÊNDICE C - Critérios para seleção e limpeza de registros.

1) Devido a proposta desse estudo algumas colunas da base de dados original do Tele-ECG foram eliminadas a saber:

a) IDLaudo_ECGLAUDO;
b) GanhoECG_ECGLAUDO;
c) Paciente_ECG;
d) Identificacao_ECGLAUDO;
e) GanhoECG_ECGLAUDO;
f) velocidadeECG_ECGLAUDO;
g) FiltroECG_ECGLAUDO;
h) DataHora_ECGLAUDO;
i) IDDiagnostico, ID_medico;
j) FC_ECGLAUDO;
k) Medicamentos_ECGLAUDO;
l) ID_MedicoAlteracao_ECGLAUDO;
m) DataAlteracao_ECGLAUDO;
n) Data_coleta.

Obs: essas colunas são visualizadas no *Excel* quando a base original completa é aberta. Os diagnósticos estão em uma tabela separada.

2) Rejeitar linhas de registro (exame de ECG) quando encontrar as frases:

a) Eletrodo de membro solto;
b) Repetindo exame;
c) Laudo já dado;
d) Eletrodo com ruído;
e) Eletrodo solto;
f) Código de barra;
g) Eletrodo solto;
h) Ganho 2N.

3) A coluna Observação_ECGLAUDO contém 10 campos mesclados (FC, P, PRi, QRS, QT, QTc, RR, sap, sat, saqrs) separados com as seguintes regras:

a) Eliminar as unidades: “sec” (segundos);
b) Eliminar as “,” (vírgulas);
c) Eliminar os símbolos “II” (visualizados no Excel);
d) Eliminar os comentários (tais como exemplo médico Kenji...);
e) Eliminar símbolo “=” (igual);
f) Eliminar *string* graus (º).

APÊNDICE D - Critérios para validar atributos de eletrocardiograma.

Item	Parâmetros	mínimo	máximo	unidade
1	sap	-30	+115	graus
2	sat	-30	+115	graus
3	saqrs	-30	+115	graus
4	FC	1	300	bpm
5	PRi	0,07	0,40	segundos
6	QT	0,30	0,60	segundos
7	QTc	0,30	0,60	segundos
8	P	0,15	0,5	segundos
9	RR	0,19	2,39	segundos
10	idade	0	105	anos
11	peso	0,5	500	kg
12	altura	40	250	cm
13	*imc*	9	80	Kg/cm^2

APÊNDICE E - Critérios para exclusão de atributos antes da exploração.

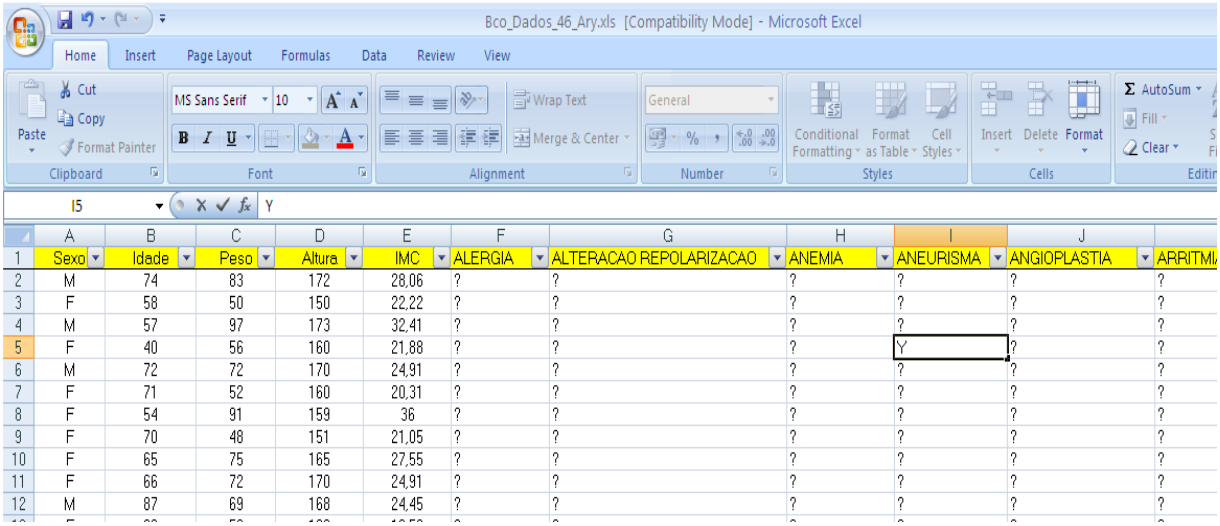
Sob orientação do especialista médico foram eliminados os atributos considerados não relevantes para essa pesquisa:

1) Ansiedade;
2) Antecedentes;
3) Ascite;
4) Congênito;
5) Derrame pericárdico (poucos casos de ocorrência na base de dados);
6) Disfagia;
7) Dispepsia;
8) Difteria;
9) Displasia;
10) Diurex;
11) Enalapril;
12) Febre;
13) Fluxetina;
14) Hipercalemia (poucos casos de ocorrência na base de dados);
15) Hiper esportiva;
16) Holter;
17) Hormônio;
18) Isordil;
19) Inapetência;
20) Lítio psiquiátrico;
21) Metilvita;
22) Mieloma mixoma intracardio (poucas ocorrências);
23) Operado PCR;
24) Pré-excitado ventricular(poucas ocorrências);
25) Prótese;

26) Sono;
27) Tuberculose.

Obs: poucas ocorrências na base de dados implica em número menor que 4 ocorrências.

APÊNDICE F - Amostra parcial da base de dados.



Bco_Dados_46_Ary.xls [Compatibility Mode] - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	
1	Sexo	Idade	Peso	Altura	IMC	ALERGIA	ALTERACAO REPOLARIZACAO	ANEMIA	ANEURISMA	ANGIOPLASTIA	ARRITMI
2	M	74	83	172	28,06	?	?	?	?	?	?
3	F	58	50	150	22,22	?	?	?	?	?	?
4	M	57	97	173	32,41	?	?	?	?	?	?
5	F	40	56	160	21,88	?	?	?	Y	?	?
6	M	72	72	170	24,91	?	?	?	?	?	?
7	F	71	52	160	20,31	?	?	?	?	?	?
8	F	54	91	159	36	?	?	?	?	?	?
9	F	70	48	151	21,05	?	?	?	?	?	?
10	F	65	75	165	27,55	?	?	?	?	?	?
11	F	66	72	170	24,91	?	?	?	?	?	?
12	M	87	69	168	24,45	?	?	?	?	?	?

APÊNDICE G - Arquivo .arff.

Arquivo arff

Formato e exemplo.

É a maneira padrão de representação de dados, para o *Weka*. Esse formato é melhor explicado usando um exemplo (tabela 15). Esse exemplo foi retirado do livro "Data Mining Practical Machine Learning Tools and Techniques"[7], encontrado na página 11. Trata-se de uma tabela de dados fictícios, com relação à condições de tempo para definir se pode ser realizado um evento esportivo.

Tabela 15 - Dados fictícios com relação à condições de tempo para definir se pode ser realizado um evento esportivo.

Tabela 15 **Weather data with some numeric attributes.**

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Pode se observar nessa tabela que estão anotados: aspecto de tempo (ensolarado, nublado, chuvoso - sunny, overcast, rainy), valor da temperatura, em fahrenheit (variando de aproximadamente 18ºC até 29ºC, ou 64ºF até 85ºF), o valor da umidade relativa do ar, a existência de vento e a ultima coluna representando o resultado (realizar ou não o determinado evento).

Assim para essa tabela 15 o arquivo arff correspondente seria como visto no quadro abaixo:

```
% arff file for the weather data with some numeric
features
%

@relation weather

@attribute outlook {sunny, overcast, rainy}

@attribute temperature numeric

@attribute humidity numeric

@attribute windy {TRUE, FALSE}

@attribute play {yes, no}



@data

%
% 14 instances
%

sunny,85,85,FALSE,no

sunny,80,90,TRUE,no

overcast,83,86,FALSE,yes

rainy,70,96,FALSE,yes

rainy,68,80,FALSE,yes

rainy,65,70,TRUE,no

overcast,64,65,TRUE,yes

sunny,72,95,FALSE,no

sunny,69,70,FALSE,yes

rainy,75,80,FALSE,yes

sunny,75,70,TRUE,yes

overcast,72,90,TRUE,yes

overcast,81,75,FALSE,yes
```

O sinal “%” indica uma linha de comentário e a linha será ignorada pelo *Weka*. Após esse comentário há o nome da relação (nome do arquivo, nesse caso weather) e a definição dos atributos, representados em colunas

separadas (outlook, temperature, humidity, windy, play), nos quais cada um deles é precedido pelo *@attribute* nome e o tipo real ou a lista de possibilidades, entre chaves, para atributos nominais (por exemplo yes, no).

Não há indicação de qual atributo estará no consequente dos resultados encontrados.

A seguir há uma linha com *@data*, indicando que a partir da próxima linha virão os valores para cada atributo, na mesma ordem em que foram definidos e serão separados por vírgula. Para os valores ausentes deve-se utilizar um ponto de interrogação.

APÊNDICE H - Um exemplo de tabela parcial resumo de resultados.

Resultados gerados pelo *Weka*

		Análise 01							
fc(25-70.5)	pri(-inf-11.475)	qt(0.375714-0.497143)	qtc(0.34625-0.4619)	sap(53.571429-72.857143)	Sexo(F)	D130	D263	rule	Confianc
	o		i		i	i		1	1
	o		i				i	2	1
	o		i			i	i	3	1
	o		i	i		i		4	1
	o		i			i		5	1
	o	i				i		6	1
	o			i		i		7	1
	o	i	i			i		8	1
i	o		i			i		9	1
i	o	i				i		10	1
	o				i	i		11	1
	o						i	12	1
	o					i	i	13	1
	o				i	i		14	1
	o					i		15	1
	o		i	i				16	1
	o			i				17	1
i	o	i						18	1
i	o		i					19	1
	o	i	i					20	1
			i			o	i	21	1
	i		i			o	i	22	1
			i			o	i	23	1
i	o	i						24	1
						o	i	25	1
	i					o	i	26	1
	o	i			i			27	1
	o					o	i	28	1
	o		i		i			29	1
	o		i					30	0.99
	o	i						31	0.99

APÊNDICE I - Diagnósticos excluídos nesse trabalho.

Alguns diagnósticos são frequentes na base de dados, mas não atendem os objetivos do estudo e foram excluídos no processo de filtragem antes do envio para a análise do especialista no domínio.

1) D305 (alterações morfológicas), pois ele é utilizado pelos médicos como “outros”, ou aqueles que podem apresentar muitas possibilidades e que as vezes não se sabe, portanto não deve ser encarado como um diagnóstico.
2) D190 = ruído de artefato
3) D312 = ausência de dados clínicos
4) D300 (Distúrbio de condução no ramo direito) obtido pela morfologia do ECG, por isso as conclusões obtidas através das regras não podem ser afirmadas e não funcionaram segundo o especialista médico.
5) D130 (Ritmo sinusal) não usado por não representar uma doença, como é pretendido nesse estudo.
6) D263 (Eletrocardiograma normal) não representa um código de doença, não atende os objetivos dessa tese.

APÊNDICE J - Intervalos de *discretização* de acordo com a DBECG.

A base de dados utilizada nesse estudo contêm 10 medidas relativas ao sinal de ECG a saber:

1) Frequência Cardíaca (FC), medida em batimentos por minuto (bpm);
2) Eixo elétrico da onda P (sap), medido em graus(º);
3) Valor de QTc calculado pela fórmula: $QTc = \frac{QT}{\sqrt{RR}}$, medido em segundos;
4) Eixo elétrico da onda QRS (saqrs), medido em graus(º);
5) Intervalo QRS, medido em segundos;
6) Intervalo PR, medido em segundos (PRi);
7) Largura da onda P, medida em segundos;
8) Eixo elétrico da onda T (sat), medido em graus(º);
9) Intervalo QT, medido em segundos
10) Intervalo RR, medido em segundos

Na exploração foram *discretizados* sete desses valores pelo fato de possuírem intervalos bem definidos.

Para o tratamento das medidas foram adotadas as seguintes faixas etárias:

a) idade de adulto > 16 anos;
b) idade de criança ≤ 16 anos.

As faixas de valores adotados para a *discretização*, em processo manual, serão descritos a seguir:

1) A medida frequência cardíaca (FC) pode está resumida na tabela 16

Tabela 16 - Faixas de FC utilizadas na *discretização* de adultos

FC adulto	Estado	Nome adotado
FC < 50 bpm	bradicardia	fc<50bpm
50 bpm ≤ FC ≤ 100 bpm	normal	fc(50bpm_100bpm)
FC > 100 bpm	taquicardia	fc>100bpm

Por exemplo, para a primeira linha, da tabela 16 indica que para as faixas de valores de frequência cardíaca menores que 50 bpm o atributo é considerado como “bradicardia” e na base de dados o nome adotado foi “fc<50bpm”. E assim por diante.

Tabela 17 - Faixas de FC utilizadas na *discretização* de crianças

criança normal	criança bradicardia	criança taquicardia	convenção para FC
fc(90bpm_182bpm)	fc<90bpm	fc>182bpm	0-1 ano → fc1
fc(89bpm_152bpm)	fc<89bpm	fc>152bpm	1-3 anos → fc2
fc(73bpm_137bpm)	fc<73bpm	fc>137bpm	3-5 anos → fc3
fc(65bpm_133bpm)	fc<65bpm	fc>133bpm	5-8 anos → fc4
fc(62bpm_130bpm)	fc<62bpm	fc>130bpm	8-12 anos → fc5
fc(60bpm_120bpm)	fc<60bpm	fc>1200bpm	12-16 anos → fc6

A tabela 17 mostra os valores utilizados para crianças e como exemplo, a segunda linha relata que para valores de frequência cardíaca entre 89 bpm e 152 bpm o atributo é considerado “normal”, já para valores menores que 90 bpm é considerado como “bradicardia” e para valores maiores que 152 bpm é considerado como “taquicardia”. Nesse caso a faixa é denominada fc2 e se refere a crianças com idade entre 1 e 3 anos de idade. Todos os intervalos de idade são abertos para o valor maior, ou seja, no exemplo descrito “1 ano” pertence a faixa e 3 anos não pertence a essa faixa.

2) Para o eixo elétrico da onda P (sap) a tabela 18 resume as faixas adotadas.

Tabela 18 - Valores utilizados para *discretização* do atributo SAP.

adulto	criança
sap<-30g	sap<60g
sap(-30g_90g)	sap(60g_120g)
sap>90g	sap>120g

A tabela 18 mostra que para adulto foram adotadas três possíveis faixas de valores para adulto e também para crianças. Esses nomes foram utilizados para os atributos, por exemplo, todos os registros de adulto com valor do eixo elétrico da onda P menor que – 30º, receberam o nome de "sap<-30g".

3) O valor de QTc calculado segue a tabela 19.

Tabela 19 - Valores adotados para o atributo QTc

Valor de QTc	Estado
QTc<0.33s	QTC curto
QTc(0.33s_0.45s)	normal
QTc[0.45s_0.47s)F	normal para mulher
QTc[0.45s_0.46s)C	normal crianças
QTc>0.45sM	QTc longo para homem
QTc>0.47sF	QTc longo para mulher
QTc>0.46sC	QTc longo para criança

Na tabela 19 pode-se ver que os valores adotados de QTc variam de acordo com o sexo e idade (criança ou adulto). Para exemplificar na ultima linha o valor de QTc>0.46 segundos foi considerado como QTc longo para criança.

4) Eixo elétrico da onda QRS (saqrs) é resumida na tabela 20.

Tabela 20 - Valores adotados para o eixo elétrico do complexo QRS

Adulto	criança normal	criança anormal	convenção para saqrs
saqrs<-67g	saqrs=120g	saqrs<>120g	≤ 1 ano
saqrs(-67g_-45g]	saqrs(60g_120g)	saqrs<60g	1 < saqrs≤ 4 anos
saqrs(-45g_-30g]		saqrs>120g	1 < saqrs≤ 4 anos
saqrs(-30g_90g]	saqrs=60g	saqrs<>60g	5 anos ≤ saqrs ≤ 16 anos
saqrs(-30g_90g]			
saqrs>=90g			

Na tabela 20 a letra “g” significa graus(º) e o símbolo “<>” representa a palavra “diferente”, assim “saqrs<>120g” indica eixo elétrico da onda QRS diferente de 120º. A primeira coluna a esquerda se refere a adulto e as demais para crianças.

5) Intervalo QRS, medido em segundos;

Tabela 21 - Valores adotados para *discretização* do intervalo QRS

Adulto	Criança
qrs<0.08s	qrs<0.09s
qrs(0.08s_0.10s]	qrs>=0.09s
qrs(0.10s_0.12s)	sem valor
qrs>0.12s	sem valor

Na tabela 21 encontram-se os valores adotados para os intervalos do complexo QRS de adultos e crianças, nos quais “qrs<0.08s” significa valores de QRS menores que 0,08 segundos. Para crianças há somente duas faixas definidas.

6) O intervalo PR está dividido em duas tabelas. A tabela 22 mostra as faixas adotadas para ECG de adultos e a tabela 23 mostra as faixas adotadas para ECG de crianças

Tabela 22 - Valores adotados para o intervalo PR para ECG de adultos

PRi	Estado
PRi(0.12s_0.20s)	normal
PRi<0.12s	menor
PRi>0.20s	maior

A tabela 22 informa que, por exemplo, o valor de PRi, adulto com valores entre 0,12 segundos e 0,20 segundos foi considerada normal e o valor anotado na base de dados *discretizada* foi “PRi(0.12s_0.20s)”.

Tabela 23 - Valores adotados para o intervalo PR para ECG de crianças

Abaixo	normal	acima	convenção para PRi
Pri<0.07s	Pri(0.07s_0.16s)	Pri>0.16s	(0-1) ano → PRi1
Pri<0.08s	Pri(0.08s_0.15s)	Pri>0.15s	[1-3) anos → PRi2
Pri<0.08s	Pri(0.08s_0.16s)	Pri>0.16s	[3-5)anos → PRi3
Pri<0.09s	Pri(0.09s_0.16s)	Pri>0.16s	[5-8) anos → PRi4
Pri<0.09s	Pri(0.09s_0.17s)	Pri>0.17s	[8-12) anos →PRi5
Pri<0.09s	Pri(0.09s_0.18s)	Pri>0.18s	[12-16) anos → PRi6

Na tabela 23 por exemplo, pode-se notar que na primeira linha o valor de intervalo PR entre 0,07 segundos e 0,16 segundos, para criança é considerado normal e o nome na base de dados *discretizada* foi “Pri(0.07s_0.16s)”.

7) Largura da onda P, medida em segundos;

Tabela 24 - Valores adotados para o intervalo da onda P

Adulto	criança
p<0.08s	sem valor
p(0.08s_0.12s)	p<0.09s
p>0.12s	p>=0.09s

A tabela 24 revela que foram adotadas três faixas de valores para adulto e duas faixas para crianças. Por exemplo, a faixa considerada normal o nome adotado para adulto foi “p(0.08s_0.12s)” e para criança “p<0.09s”.

APÊNDICE K - Parâmetro *Bin.*

Parâmetro *bin é* utilizado no processo de *discretização*, alterando um atributo numérico para categórico e define o número de intervalos a serem utilizados no processo de *discretização*. É oriundo do processo denominado *binning.*

Binning consta de um método que ordena os valores dos atributos, da base de dados, utilizando o conceito de vizinhança entre os dados. Após a ordenação, os valores são distribuídos por grupos (*bins* ou *buckets*) nos quais cada um deverá manter a representatividade do elementos. Em cada grupo aplica-se um critério na escolha de uma medida para ajusta-los, tais como a média aritmética, a mediana ou um valor de limite. Assim substituem-se os valores pelas medidas calculadas em cada grupo, ajustando assim os valores da série. Diversos métodos podem ser utilizados para ajustar os valores dos grupos. Exemplo: o atributo “idade” poderia ser dividido em:

criança: 0-12 anos
adolescente: >12-17 anos
jovem: >17-35 anos
adulto: >35-59 anos
idoso: >59 anos

No processo de *discretização* os *bins* podem dividir a faixa de valores para um determinado atributo em sub-faixas com as possibilidades:

a) mesmo número de valores dentro da faixa. Exemplo: com 10 valores para os atributos:

5, 7, 12, 35, 65, 82, 84, 88, 90, 95.

Para criar 5 *bins*, pode-se dividir a faixa de valores com dois valores em cada faixa:

[5,7], [12,35], [65,82], [84,88], [90,95]

b) mesma largura de faixa. Exemplo: para valores observados entre 0 - 100, poder-se-iam criar 5 *bins*: largura = (100 – 0)/5 = 20

faixas:

[0-20], (20-40], (40-60], (60-80], (80-100]

Tipicamente o primeiro e último *bins* são estendidos para todos os valores fora de faixa, assim:

(-infinito-20], (20-40], (40-60], (60-80], (80-infinito)

Discretização* no *Weka

- No *Weka*, a *discretização* de um atributo é realizada aplicando-se o filtro apropriado, após a carga da base de dados, na aba pré-processamento escolhendo-se o botão “filtro”.
- Escolhe-se a opção *discretize* em filtros não supervisionados.
- Por *default*, utiliza-se igual largura de *binning*.
- Para usar igual frequência clicar no nome do filtro e nas propriedades dele definir “use *equal-frequency parameter*” como verdadeiro (*true*).

APÊNDICE L - Um resultado *Weka* de *data mining.*

Exemplo de um resultado de análise, pelo *Weka*, com *confiança* de 0.9 e geração de 100 regras de associação.

Análise 01

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 100 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     ECG-weka.filters.unsupervised.attribute.Discretize-B14-M-1.0-Rfirst-last
Instances:    24021
Attributes:   278
              [list of attributes omitted]
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.5 (12010 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10
Generated sets of large itemsets:
Size of set of large itemsets L(1): 8
Size of set of large itemsets L(2): 20
Size of set of large itemsets L(3): 20
Size of set of large itemsets L(4): 7
Best rules found:
 1. sexo=F qtc='(0.34625-0.4619]' D130=Y 14546 ==> pri='(-inf-11.475]' 14546    conf:(1)
 2. qtc='(0.34625-0.4619]' D263=Y 14451 ==> pri='(-inf-11.475]' 14451    conf:(1)
 3. qtc='(0.34625-0.4619]' D130=Y D263=Y 14386 ==> pri='(-inf-11.475]' 14386    conf:(1)
 4. sap='(53.571429-72.857143]' qtc='(0.34625-0.4619]' D130=Y 13949 ==> pri='(-inf-11.475]' 13949    conf:(1)
 5. qtc='(0.34625-0.4619]' D130=Y 20242 ==> pri='(-inf-11.475]' 20241    conf:(1)
 6. qt='(0.375714-0.497143]' D130=Y 15586 ==> pri='(-inf-11.475]' 15585    conf:(1)
 7. sap='(53.571429-72.857143]' D130=Y 15173 ==> pri='(-inf-11.475]' 15172    conf:(1)
 8. qt='(0.375714-0.497143]' qtc='(0.34625-0.4619]' D130=Y 14193 ==> pri='(-inf-11.475]' 14192    conf:(1)
 9. fc='(25-70.5]' qtc='(0.34625-0.4619]' D130=Y 13634 ==> pri='(-inf-11.475]' 13633    conf:(1)
10. fc='(25-70.5]' qt='(0.375714-0.497143]' D130=Y 12425 ==> pri='(-inf-11.475]' 12424    conf:(1)
11. sexo=F D130=Y 15874 ==> pri='(-inf-11.475]' 15871    conf:(1)
12. D263=Y 15036 ==> pri='(-inf-11.475]' 15033    conf:(1)
13. D130=Y D263=Y 14966 ==> pri='(-inf-11.475]' 14963    conf:(1)
14. fc='(25-70.5]' D130=Y 14565 ==> pri='(-inf-11.475]' 14562    conf:(1)
15. D130=Y 22082 ==> pri='(-inf-11.475]' 22075    conf:(1)
16. sap='(53.571429-72.857143]' qtc='(0.34625-0.4619]' 14807 ==> pri='(-inf-11.475]' 14786    conf:(1)
17. sap='(53.571429-72.857143]' 16172 ==> pri='(-inf-11.475]' 16122    conf:(1)
18. fc='(25-70.5]' qt='(0.375714-0.497143]' qtc='(0.34625-0.4619]' 12897 ==> pri='(-inf-11.475]' 12850    conf:(1)
19. fc='(25-70.5]' qtc='(0.34625-0.4619]' 14834 ==> pri='(-inf-11.475]' 14778    conf:(1)
20. qt='(0.375714-0.497143]' qtc='(0.34625-0.4619]' 15301 ==> pri='(-inf-11.475]' 15235    conf:(1)
21. qtc='(0.34625-0.4619]' D263=Y 14451 ==> D130=Y 14386    conf:(1)
```

22. pri='(-inf-11.475]' qtc='(0.34625-0.4619]' D263=Y 14451 ==> D130=Y 14386 conf:(1)
23. qtc='(0.34625-0.4619]' D263=Y 14451 ==> pri='(-inf-11.475]' D130=Y 14386 conf:(1)
24. fc='(25-70.5]' qt='(0.375714-0.497143]' 13562 ==> pri='(-inf-11.475]' 13501 conf:(1)
25. D263=Y 15036 ==> D130=Y 14966 conf:(1)
26. pri='(-inf-11.475]' D263=Y 15033 ==> D130=Y 14963 conf:(1)
27. sexo=F qt='(0.375714-0.497143]' 12158 ==> pri='(-inf-11.475]' 12101 conf:(1)
28. D263=Y 15036 ==> pri='(-inf-11.475]' D130=Y 14963 conf:(1)
29. sexo=F qtc='(0.34625-0.4619]' 15432 ==> pri='(-inf-11.475]' 15355 conf:(1)
30. qtc='(0.34625-0.4619]' 21833 ==> pri='(-inf-11.475]' 21686 conf:(0.99)
31. qt='(0.375714-0.497143]' 16841 ==> pri='(-inf-11.475]' 16725 conf:(0.99)
32. fc='(25-70.5]' 15942 ==> pri='(-inf-11.475]' 15825 conf:(0.99)
33. sexo=F 16942 ==> pri='(-inf-11.475]' 16778 conf:(0.99)
34. pri='(-inf-11.475]' D130=Y D263=Y 14963 ==> qtc='(0.34625-0.4619]' 14386 conf:(0.96)
35. pri='(-inf-11.475]' D263=Y 15033 ==> qtc='(0.34625-0.4619]' 14451 conf:(0.96)
36. D130=Y D263=Y 14966 ==> qtc='(0.34625-0.4619]' 14386 conf:(0.96)
37. D130=Y D263=Y 14966 ==> pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 14386 conf:(0.96)
38. D263=Y 15036 ==> qtc='(0.34625-0.4619]' 14451 conf:(0.96)
39. D263=Y 15036 ==> pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 14451 conf:(0.96)
40. pri='(-inf-11.475]' D263=Y 15033 ==> qtc='(0.34625-0.4619]' D130=Y 14386 conf:(0.96)
41. D263=Y 15036 ==> qtc='(0.34625-0.4619]' D130=Y 14386 conf:(0.96)
42. D263=Y 15036 ==> pri='(-inf-11.475]' qtc='(0.34625-0.4619]' D130=Y 14386 conf:(0.96)
43. fc='(25-70.5]' pri='(-inf-11.475]' qt='(0.375714-0.497143]' 13501 ==> qtc='(0.34625-0.4619]' 12850 conf:(0.95)
44. fc='(25-70.5]' qt='(0.375714-0.497143]' 13562 ==> qtc='(0.34625-0.4619]' 12897 conf:(0.95)
45. fc='(25-70.5]' qt='(0.375714-0.497143]' 13562 ==> pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 12850 conf:(0.95)
46. sexo=F pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 15355 ==> D130=Y 14546 conf:(0.95)
47. sexo=F pri='(-inf-11.475]' 16778 ==> D130=Y 15871 conf:(0.95)
48. sap='(53.571429-72.857143]' pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 14786 ==> D130=Y 13949 conf:(0.94)
49. sexo=F qtc='(0.34625-0.4619]' 15432 ==> D130=Y 14546 conf:(0.94)
50. sexo=F qtc='(0.34625-0.4619]' 15432 ==> pri='(-inf-11.475]' D130=Y 14546 conf:(0.94)
51. sap='(53.571429-72.857143]' qtc='(0.34625-0.4619]' 14807 ==> D130=Y 13949 conf:(0.94)
52. sap='(53.571429-72.857143]' qtc='(0.34625-0.4619]' 14807 ==> pri='(-inf-11.475]' D130=Y 13949 conf:(0.94)
53. sap='(53.571429-72.857143]' pri='(-inf-11.475]' 16122 ==> D130=Y 15172 conf:(0.94)
54. sap='(53.571429-72.857143]' 16172 ==> D130=Y 15173 conf:(0.94)
55. sap='(53.571429-72.857143]' 16172 ==> pri='(-inf-11.475]' D130=Y 15172 conf:(0.94)
56. sexo=F 16942 ==> D130=Y 15874 conf:(0.94)
57. sexo=F 16942 ==> pri='(-inf-11.475]' D130=Y 15871 conf:(0.94)
58. fc='(25-70.5]' pri='(-inf-11.475]' D130=Y 14562 ==> qtc='(0.34625-0.4619]' 13633 conf:(0.94)
59. fc='(25-70.5]' D130=Y 14565 ==> qtc='(0.34625-0.4619]' 13634 conf:(0.94)
60. fc='(25-70.5]' D130=Y 14565 ==> pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 13633 conf:(0.94)
61. fc='(25-70.5]' pri='(-inf-11.475]' 15825 ==> qtc='(0.34625-0.4619]' 14778 conf:(0.93)
62. pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 21686 ==> D130=Y 20241 conf:(0.93)
63. pri='(-inf-11.475]' qt='(0.375714-0.497143]' 16725 ==> D130=Y 15585 conf:(0.93)
64. pri='(-inf-11.475]' qt='(0.375714-0.497143]' qtc='(0.34625-0.4619]' 15235 ==> D130=Y 14192 conf:(0.93)
65. pri='(-inf-11.475]' 23707 ==> D130=Y 22075 conf:(0.93)
66. fc='(25-70.5]' 15942 ==> qtc='(0.34625-0.4619]' 14834 conf:(0.93)
67. qt='(0.375714-0.497143]' qtc='(0.34625-0.4619]' 15301 ==> D130=Y 14193 conf:(0.93)
68. qt='(0.375714-0.497143]' qtc='(0.34625-0.4619]' 15301 ==> pri='(-inf-11.475]' D130=Y 14192 conf:(0.93)
69. qtc='(0.34625-0.4619]' 21833 ==> D130=Y 20242 conf:(0.93)
70. qtc='(0.34625-0.4619]' 21833 ==> pri='(-inf-11.475]' D130=Y 20241 conf:(0.93)

71. fc='(25-70.5]' 15942 ==> pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 14778 conf:(0.93)
72. qt='(0.375714-0.497143]' 16841 ==> D130=Y 15586 conf:(0.93)
73. qt='(0.375714-0.497143]' 16841 ==> pri='(-inf-11.475]' D130=Y 15585 conf:(0.93)
74. fc='(25-70.5]' pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 14778 ==> D130=Y 13633 conf:(0.92)
75. fc='(25-70.5]' pri='(-inf-11.475]' qt='(0.375714-0.497143]' 13501 ==> D130=Y 12424 conf:(0.92)
76. fc='(25-70.5]' pri='(-inf-11.475]' 15825 ==> D130=Y 14562 conf:(0.92)
77. sap='(53.571429-72.857143]' pri='(-inf-11.475]' D130=Y 15172 ==> qtc='(0.34625-0.4619]' 13949 conf:(0.92)
78. sap='(53.571429-72.857143]' D130=Y 15173 ==> qtc='(0.34625-0.4619]' 13949 conf:(0.92)
79. sap='(53.571429-72.857143]' D130=Y 15173 ==> pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 13949 conf:(0.92)
80. fc='(25-70.5]' qtc='(0.34625-0.4619]' 14834 ==> D130=Y 13634 conf:(0.92)
81. fc='(25-70.5]' qtc='(0.34625-0.4619]' 14834 ==> pri='(-inf-11.475]' D130=Y 13633 conf:(0.92)
82. sap='(53.571429-72.857143]' pri='(-inf-11.475]' 16122 ==> qtc='(0.34625-0.4619]' 14786 conf:(0.92)
83. pri='(-inf-11.475]' D130=Y 22075 ==> qtc='(0.34625-0.4619]' 20241 conf:(0.92)
84. D130=Y 22082 ==> qtc='(0.34625-0.4619]' 20242 conf:(0.92)
85. D130=Y 22082 ==> pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 20241 conf:(0.92)
86. sexo=F pri='(-inf-11.475]' D130=Y 15871 ==> qtc='(0.34625-0.4619]' 14546 conf:(0.92)
87. sexo=F D130=Y 15874 ==> qtc='(0.34625-0.4619]' 14546 conf:(0.92)
88. sexo=F D130=Y 15874 ==> pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 14546 conf:(0.92)
89. fc='(25-70.5]' qt='(0.375714-0.497143]' 13562 ==> D130=Y 12425 conf:(0.92)
90. fc='(25-70.5]' qt='(0.375714-0.497143]' 13562 ==> pri='(-inf-11.475]' D130=Y 12424 conf:(0.92)
91. sap='(53.571429-72.857143]' 16172 ==> qtc='(0.34625-0.4619]' 14807 conf:(0.92)
92. sexo=F pri='(-inf-11.475]' 16778 ==> qtc='(0.34625-0.4619]' 15355 conf:(0.92)
93. pri='(-inf-11.475]' 23707 ==> qtc='(0.34625-0.4619]' 21686 conf:(0.91)
94. sap='(53.571429-72.857143]' 16172 ==> pri='(-inf-11.475]' qtc='(0.34625-0.4619]' 14786 conf:(0.91)
95. fc='(25-70.5]' 15942 ==> D130=Y 14565 conf:(0.91)
96. fc='(25-70.5]' 15942 ==> pri='(-inf-11.475]' D130=Y 14562 conf:(0.91)
97. pri='(-inf-11.475]' qt='(0.375714-0.497143]' 16725 ==> qtc='(0.34625-0.4619]' 15235 conf:(0.91)
98. sexo=F 16942 ==> qtc='(0.34625-0.4619]' 15432 conf:(0.91)
99. qt='(0.375714-0.497143]' D130=Y 15586 ==> qtc='(0.34625-0.4619]' 14193 conf:(0.91)
100. pri='(-inf-11.475]' qt='(0.375714-0.497143]' D130=Y 15585 ==> qtc='(0.34625-0.4619]' 14192 conf:(0.91)

APÊNDICE M - Formato de regras descobertas com o *Orange* resumido.

	A	B	C	D	E	F	G	H	I	J
1	Afasta	(0.722)	(0.733)	(0.743)	(0.828)	(0.838)	(0.876)	(0.944)	(0.981)	(0.984)
2	sexo=F	D300	D139	D175	D220	D224	D214	D292	D290	D305
3		(0.983)								
4	HAS=Y	D139							discorda	
5	aponta	(1.010)	(1.242)	(1.152)	(1.297)					
6	sexo=F	D291	D296	D253	D294					
7		(1.038)	(1.045)	(1.134)	(1.296)	(1.387)	(1.411)	(1.615)	(1.639)	(1.666)
8	sexo=M	D305	D290	D292	D214	D224	D220	D175	D139	D300
9		(1.069)	(1.084)	(1.340)	(1.392)	(1.550)	(2.415)			
10	sexo=F HAS	D305	D224	D290	D292	D214	DIABETE			
11		(1.111)	(1.269)	(1.295)	(1.366)	(1.438)	(1.727)	(2.256)	(2.389)	
12	HAS	D305	D224	D175	D290	D292	D214	DIABETE	sexo=F DIABETE	
13		(2.389)								
14	sexo=F DIABETE	HAS								
15		(9.293)	(9.674)							
16	D292	sexo=F D290	D290							
17		(2.415)	(2.256)							
18	DIABETE	sexo=F HAS	HAS							
19		(9.656)								
20	sexo=F D292	D290								
21		(9.656)	(9.674)							
22	D290	sexo=F D292	D292							
23		(9.293)								
24	sexo=F D290=Y	D292								
25		(4.020)								
26	fc='(76.8-101.2]'	pri=(-inf-16.045] rr=(0.631-0.818]								
27		(4.020)								
28	pri=(-inf-16.045] rr=(0.631-0.818]	fc='(76.8-101.2]'								

APÊNDICE N – Um exemplo de *script* utilizado para gerenciar a execução de exploração no ambiente *R-Project*

Obs.: é um arquivo em formato não documento, por isso não deve conter acentuação.

```
##############################################################################################
# Script para geracao de regras de associacao a partir do arquivo arff de entrada gerado pelo                    #
# arquivo excel do banco de dados de exames                                                                      #
##############################################################################################

#-----------------------------------------------
# Pre-processamento dos dados
#-----------------------------------------------
# Aqui, serao carregadas as bibliotecas do R-project para o trabalho e serao realizadas
# algumas transformacoes para adequar a entrada ao formato de trabalho interno do programa.
# Esse pre-processamento na getcw() nao altera nada nos valores dos dados, serve apenas para preparar
# o arquivo de entrada para ser utilizado no processo.
#-----------------------------------------------------------------------------------------------
# Biblioteca do R-project para tratar arquivos arff
library(foreign)

# Biblioteca do R-project para geracao de regras de associacao
# (Utiliza a implementacao do Borgelt)
library(arules)

# Carrega o arquivo arff na memoria
# (Para utilizar e preciso modificar o caminho do arquivo para o do sistema.)
#  Exemplo: c:/dados/arquivo_entrada.arff )
arff_file <- read.arff('c:/users/ferreira/doutorado/workspace01/BD106.arff')

# Converte o formato de representacao interna do R-project para adequar ao processo.
arff_file_factor <- as.data.frame(lapply(arff_file,factor))

# Cria uma representacao em formato de transacoes do arquivo de entrada
arff_transactions <- as(arff_file_factor, "transactions")

# Cria um arquivo txt com as transacoes no formato do programa de geracao de regras
# (Apenas para inspecao visual. Nao e necessario para o processo)
# (Para utilizar e preciso modificar o caminho do arquivo para o do sistema.)
# (Dica: No windows, utilize a barra invertida -> /
```

```
#  Exemplo: c:/dados/arquivo_entrada.arff )
write(arff_transactions,file="c:/users/ferreira/doutorado/workspace01/transactions.txt")

#-----------------------------------------------------------------------------------------
# Processamento dos dados e geracao das regras
#-----------------------------------------------------------------------------------------
# Agora as regras de associacao serao geradas
#-----------------------------------------------------------------------------------------

# Executa o programa para geracao de regras de associacao a partir das transacoes.
# supp - valor do suporte minimo. 1% -> 0.01
# conf - valor de confianca minima. 1% -> 0.01
# target - apenas informa ao programa que e para gerar regras de associacao.
arff_rules <- apriori(arff_transactions, parameter = list(supp=0.03, conf=0.05, target="rules"))

#-----------------------------------------------------------------------------------------
# Pos-processamento das regras geradas
#-----------------------------------------------------------------------------------------
# Aplica  filtros para deixar no conjunto apenas as regras de associacao de interesse
#-----------------------------------------------------------------------------------------

# Filtra as regras que tem no consequente apenas os valores no formato DXXX=Y, ou seja,
# aquelas regras que implicam em algum diagnostico positivo.
arff_rules_subset <- subset(arff_rules, subset = rhs %pin% "=Y")

# Filtra as regras com lift maior que 1.5, ou seja, que aumenta em 1.5X a chance do consequente acontecer
arff_rules_subset <- subset(arff_rules_subset, subset = lift > 1.5)

# Ordena as regras filtradas pelo valor de lift
arff_rules_subset <- sort(arff_rules_subset, by="lift")

# Cria um arquivo txt de saida com as regras geradas e ordenadas pelo lift
write(arff_rules_subset, file="c:/users/ferreira/doutorado/workspace01/rules_subset.txt")

#-----------------------------------------------------------------------------------------
# Daqui para baixo eu refaz o processo de geracao de regra, mas tentando discretizar os atributos.
#-----------------------------------------------------------------------------------------

# Discretizacao - Aqui, tenta discretizar os atributos numericos para aumentar as chances deles aparecerem nas regras
arff_transactions_categorize <- arff_file
arff_transactions_categorize[,'idade']  <-  discretize(arff_transactions_categorize[,'idade'],  "fixed",  categories=c(0,20,60,Inf),
labels=c("(0-20_anos)", "(20-60_anos)", "(60+_anos)"))
arff_transactions_categorize[,'altura'] <- discretize(arff_transactions_categorize[,'altura'], "fixed", categories=c(0,150,180,Inf),
labels=c("(0-150_cm)", "(150-180_cm)", "(180+_cm)"))
```

```
arff_transactions_categorize[,'imc'] <- discretize(arff_transactions_categorize[,'imc'], "fixed", categories=c(0,17,18.49,24.99,29.99,34.99,39.99,Inf), labels=c("(0-17)", "(17-18.49)", "(18.5-24.99)", "(25-29.99)", "(30-34.99)", "(35-39.99)", "(40+)"))
arff_transactions_categorize[,'peso'] <- discretize(arff_transactions_categorize[,'peso'], "fixed", categories=c(0,60,80,100,Inf), labels=c("(0,60_kg)", "(60,80_kg)", "(80,100_kg)", "(100+_kg)"))
arff_transactions_categorize[,'sat'] <- discretize(arff_transactions_categorize[,'sat'], "interval", categories=4)
arff_transactions_categorize[,'rr'] <- discretize(arff_transactions_categorize[,'rr'], "interval", categories=4)
arff_transactions_categorize[,'qt'] <- discretize(arff_transactions_categorize[,'qt'], "interval", categories=4)# Uma vez discretizados, eu gerei novamente um conjunto de transacoes, mas com os novos atributos discretos
arff_transactions_categorize <- as(arff_transactions_categorize, "transactions")
# Cria o arquivo txt das transacoes para conferir
write(arff_transactions_categorize,file="c:/users/ferreira/doutorado/workspace01/transactions_categorize.txt")
# Geracao das novas regras de associacao considerando os atributos discretizados
arff_rules_categorize <- apriori(arff_transactions_categorize, parameter = list(supp=0.03, conf=0.05, target="rules"))

# Filtra as regras que tem no consequente apenas os valores no formato DXXX=Y, ou seja,
# aquelas regras que implicam em algum diagnostico positivo.
arff_rules_categorize <- subset(arff_rules_categorize, subset = rhs %pin% "=Y")

# Remove das regras com o consequente "HAS=Y" ou "HAS=N", uma vez que o interesse e nos outros diagnosticos
arff_rules_categorize <- subset(arff_rules_categorize, subset = ! rhs %pin% "HAS=")

# Remove das regras  com o consequente "DIABETE=Y" ou "DIABETE=N", uma vez que o interesse e nos outros diagnosticos
arff_rules_categorize <- subset(arff_rules_categorize, subset = ! rhs %pin% "DIABETE=")

# Devido a nova configuracao, filtra as regras com lift maior que 10, pois gera-se muito mais regras agora
arff_rules_categorize <- subset(arff_rules_categorize, subset=lift > 10)

# Ordena as regras filtradas pelo valor de lift
arff_rules_categorize <- sort(arff_rules_categorize, by="lift")

# Cria um arquivo txt de saida com as regras geradas considerando os atributos discretos e ordenadas pelo lift
write(arff_rules_categorize, file="c:/users/ferreira/doutorado/workspace01/rules_subset_categorize_lift_10.txt")
```