BRUNA ELISA ZANCHETTA BUANI

APLICAÇÃO DA LÓGICA FUZZY KNN E ANÁLISES ESTATÍSTICAS
PARA SELEÇÃO DE CARACTERÍSTICAS E
CLASSIFICAÇÃO DE ABELHAS

SÃO PAULO

2010

BRUNA ELISA ZANCHETTA BUANI

APLICAÇÃO DA LÓGICA FUZZY KNN E ANÁLISES ESTATÍSTICAS
PARA SELEÇÃO DE CARACTERÍSTICAS E
CLASSIFICAÇÃO DE ABELHAS

Dissertação apresentada à Escola
Politécnica da Universidade de São
Paulo para obtenção do título de
Mestre em Engenharia

Área de concentração:
Sistemas Digitais

Orientador: Prof. Dr. André Riyuiti
Hirakawa

SÃO PAULO
2010

Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, de outubro de 2010.

Assinatura do autor _____

Assinatura do orientador _____

FICHA CATALOGRÁFICA

Buani, Bruna Elisa Zanchetta

Aplicação da lógica Fuzzy KNN e análises estatísticas para seleção de características e classificação de abelhas / B.E.Z. Buani. – ed.rev. – São Paulo, 2010.

133 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1. Lógica Fuzzy 2. Análise estatística de dados 3. Bioinformática 4. Abelhas I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II. t.

BRUNA ELISA ZANCHETTA BUANI

APLICAÇÃO DA LÓGICA FUZZY KNN E ANÁLISES ESTATÍSTICAS
PARA SELEÇÃO DE CARACTERÍSTICAS E
CLASSIFICAÇÃO DE ABELHAS

Dissertação apresentada à Escola
Politécnica da Universidade de São
Paulo para obtenção do título de
Mestre em Engenharia

SÃO PAULO

2010

RESUMO

Este trabalho propõe uma alternativa para o problema de classificação de espécies de abelhas a partir da implementação de um algoritmo com base na Morfometria Geométrica e estudo das *Formas* dos marcos anatômicos das imagens obtidas pelas asas das abelhas. O algoritmo implementado para este propósito se baseia no algoritmo dos k-Vizinhos mais Próximos (do inglês, kNN) e na Lógica Fuzzy kNN (Fuzzy k-Nearest Neighbor) aplicados à dados analisados e selecionados de pontos bidimensionais referentes às características geradas por marcos anatômicos. O estudo apresentado envolve métodos de seleção e ordenação de marcos anatômicos para a utilização no algoritmo por meio da implementação de um método matemático que utiliza o cálculo dos marcos anatômicos mais significativos (que são representados por marcos matemáticos) e a formulação da Ordem de Significância onde cada elemento representa variáveis de entrada para a Fuzzy kNN. O conhecimento envolvido neste trabalho inclui uma perspectiva sobre a seleção de características não supervisionada como agrupamentos e mineração de dados, análise de pré-processamento dos dados, abordagens estatísticas para estimação e predição, estudo da *Forma*, Análise de Procrustes e Morfometria Geométrica sobre os dados e o tópico principal que envolve uma modificação do algoritmo dos k-Vizinhos mais Próximos e a aplicação da Fuzzy kNN para o problema. Os resultados mostram que a classificação entre amostras de abelhas no seu próprio grupo apresentam acurácia de 90%, dependendo da espécie. As classificações realizadas entre as espécies de abelhas alcançaram acurácia de 97%.

Palavras-chave: Classificação. Abelhas. Lógica Fuzzy. Reconhecimento de Padrões.

ABSTRACT

This work presents a proposal to solve the bees classification problem by implementing an algorithm based on Geometrics Morphometrics and the Shape analysis of landmarks generated from bees wings images. The algorithm is based on the K-Nearest Neighbor (K-Nearest Neighbor) algorithm and Fuzzy Logic KNN applied to the analysis and selection of two-dimensional data points relating to landmarks. This work is part of the *Architecture Reference Model for Automatic identification and Taxonomic Classification System of Stingless Bee using the Wing Morphometry*. The study includes selection and ordering methods for landmarks used in the algorithm by developing a mathematical model to represent the significance order, generating the most significant mathematical landmarks as input variables for Fuzzy Logic kNN. The main objective of this work is to develop a classification system for bee species. The knowledge involved in the development of this work include an overview of feature selection, unsupervised clustering and data mining, analysis of data pre-processing, statistical approaches for estimation and prediction, study of Shape, Procrustes Analysis on data that comes from Geometric Morphometry and the modification of the k-Nearest Neighbors algorithm and the Fuzzy Logic kNN. The results show that the classification in bee samples of the same species presents a accuracy above 90%, depending on the specie in analysis. The classification done between the bees species reach accuracies of 97%.

Keywords: Classification. Bees. Fuzzy Logic. Pattern Recognition.

LISTA DE ILUSTRAÇÕES

Figura 1 - Arquitetura de Referência para o Sistema de Identificação e Classificação Taxonômica de Abelhas sem Ferrão.....	14
Figura 2 - Imagem de <i>Brassolis Astyra</i> Godart, 1824. Marcos Anatômicos utilizados nas análises das pupas: 1, vista lateral; 2, vista Frontal. Marcos Anatômicos utilizados nas análises das asas: 3, anterior; 4, posterior.....	21
Figura 3 - Exemplo de imagem da espécie <i>Friesella schrottkyi</i> (operária).....	22
Figura 4 - Apresentação dos marcos anatômicos de quatro amostras da espécie 0 (<i>Friesella schrottkyi</i>) sem alinhamento.....	22
Figura 5 - Apresentação de quatro amostras da espécie 0 (<i>Friesella schrottkyi</i>) alinhadas.....	23
Figura 6 - Aplicação da Análise de Componentes Principais nos dados da Espécie 0	28
Figura 7 - Interface principal do programa tpsUtil.....	39
Figura 8 - Interface principal do programa tpsDig2.....	40
Figura 9 - Diagrama de uma seção transversal do Espaço da Forma de Kendall...	41
Figura 10 - Interface principal do programa tpsRelw.....	42
Figura 11 - Apresentação de um conjunto e o conceito de dicotomia.....	47
Figura 12 - Exemplos de Funções de pertinência para duas espécies.....	48
Figura 13 - Modelo do processo envolvido para classificação de abelhas.....	55
Figura 14 - Imagem de uma Asa da espécie <i>Scaura Latittarsis</i>	57
Figura 15 - Imagem da Asa de uma amostra da espécie <i>Scaura Latittarsis</i> com os marcos anatômicos colocados.....	57
Figura 16 - Arquivo do conjunto de dados em formato TPS.....	58
Figura 17 - Arquivo dos pontos alinhados (align species) do software TPSRelw.....	59
Figura 18 - Entrada de dados do algoritmo FNN4Bees.....	60
Figura 19 - Marcos anatômicos mais significativos a partir dos gráficos de dispersão da espécie <i>Geotrigona_sp</i> (espécie 2).....	64
Figura 20 - Pontos (x, y) de três espécies diferentes com cinco marcos anatômicos	67
Figura 21 - Inserção de nova espécie para classificação.....	68

Figura 22 - Análise do gráfico para pesquisa dos marcos anatômicos mais significativos.....	69
Figura 23 - Modelo da implementação da Fuzzy kNN nos marcos anatômicos mais significativos.....	76
Figura 24 - Disposição no gráfico dos marcos anatômicos das Espécies 0, 1, 2 e indefinida.....	81
Figura 25 - Gráfico de dispersão dos marcos anatômicos da espécie E. chaly.....	85
Figura 26 - Gráfico de dispersão dos marcos anatômicos da espécie E. flammea...	86
Figura 27 - Gráfico de dispersão dos marcos anatômicos da espécie E. ignita.....	87
Figura 28 - Gráfico de dispersão dos marcos anatômicos da espécie E. imperialis..	87
Figura 29 - Gráfico de dispersão dos marcos anatômicos da espécie E. orellana....	88
Figura 30 - Grau de Pertinência da amostra de teste da espécie E. chaly com as demais espécies.....	98
Figura 31 - Grau de Pertinência da amostra de teste da espécie E. flammea com as demais espécies.....	100
Figura 32 - Grau de Pertinência da amostra de teste E. ignita com as demais espécies.....	102
Figura 33 - Grau de Pertinência da amostra de teste E. imperialis com as demais espécies.....	104
Figura 34 - Grau de Pertinência da amostra de teste E. orellana com as demais espécies.....	106
Figura 35 - Saída da Função Fuzzy para cada classe de espécies Euglossas.....	107
Figura 36 - Diagrama de Classes da implementação do algoritmo FNN4Bees.....	110
Figura 37 - Diagrama de Classes do FNN4Bees detalhado.....	111
Figura 38 - Diagrama de classes na manipulação de arquivos texto do algoritmo FNN4Bees.....	112
Figura 39 - Detalhamento da Classe SeleccionadorDeCaracteristicas.....	113

LISTA DE TABELAS

Tabela 1 - Análise de variância da matriz de correlação.....	28
Tabela 2 - Apresentação das variáveis e seus respectivos termos para cada componente principal.....	29
Tabela 3 - Apresentação dos marcos anatômicos mais significativos para as espécies de 0 até 18 a partir do FNN4Bees.....	65
Tabela 4 - Apresentação dos marcos anatômicos mais significativos para as espécies de 19 até 35 a partir do FNN4Bees.....	65
Tabela 5 - Definição dos marcos anatômicos mais significativos para as espécies de 0 até 18 a partir do gráfico de dispersão.....	67
Tabela 6 - Definição dos marcos anatômicos mais significativos para as espécies de 0 até 18 a partir do gráfico de dispersão.....	67
Tabela 7 - Cálculo da Variância e do desvio padrão para onze marcos anatômicos da espécie <i>Friesella schrottkyi</i>	73
Tabela 8 - Resultado das médias das distâncias obtidas entre a espécie 6 com a espécie 0 até a espécie 7, em mm, pelo algoritmo FNN4Bees.....	77
Tabela 9 - Resultado das médias das distâncias obtidas entre a espécie 6 com a espécie 8 até a espécie 15, em mm, pelo algoritmo FNN4Bees.....	77
Tabela 10 - Resultado das médias das distâncias obtidas entre a espécie 6 com a espécie 16 até a espécie 23, em mm, pelo algoritmo FNN4Bees.....	77
Tabela 11 - Resultado das médias das distâncias obtidas entre a espécie 6 com a espécie 24 até a espécie 31, em mm, pelo algoritmo FNN4Bees.....	77
Tabela 12 - Resultado das médias das distâncias obtidas entre a espécie 6 com a espécie 32 até a espécie 35, em mm, pelo algoritmo FNN4Bees.....	77
Tabela 13 - Coordenadas (x, y), em mm, dos marcos anatômicos das Espécies 0, 1, 2 e da espécie indefinida.....	83
Tabela 14 - Distâncias, em mm, entre a espécie indefinida e as espécies 0, 1, 2 dos marcos anatômicos mais significativos dadas pelo FNN4Bees.....	84
Tabela 15 - Listagem contendo os pontos, em mm, no arquivo TPS da <i>E. chaly</i>	85
Tabela 16 - Resultado, em mm, para o teste com a <i>E. chaly</i> como entrada.....	90

Tabela 17 - Média das distâncias, em mm, totais calculadas pelo algoritmo FNN4Bees entre as espécies Euglossas.....	91
Tabela 18 - Diferenças entre os marcos anatômicos, em mm, da espécie E. flammea com as demais espécies.....	92
Tabela 19 - Média das distâncias totais, em mm, calculadas pelo algoritmo FNN4Bees entre a espécie E. flammea e as espécies Euglossas.....	93
Tabela 20 - Diferenças entre os marcos anatômicos, em mm, da espécie E. ignita com as demais espécies.....	94
Tabela 21 - Média das distâncias totais, em mm, calculadas pelo algoritmo FNN4Bees entre a espécie E. ignita e as espécies Euglossas.....	94
Tabela 22 - Diferenças entre os marcos anatômicos, em mm, da espécie E. imperialis com as demais espécies.....	95
Tabela 23 - Médias das distâncias totais, em mm, calculadas pelo algoritmo FNN4Bees entre a espécie E. imperialis e as espécies Euglossas.....	95
Tabela 24 - Diferenças entre os marcos anatômicos, em mm, da espécie E. orellana com as demais espécies.....	96
Tabela 25 - Médias das distâncias totais calculadas, em mm, pelo algoritmo FNN4Bees entre a espécie E. orellana e as espécies Euglossas.....	97
Tabela 26 - Valores, em mm, de todos os marcos anatômicos padrões das espécies indefinida (N), 0, 1, 2, 3 e 4.....	97
Tabela 27 - Distâncias, em mm, entre a espécie indefinida e as outras três espécies	98
Tabela 28 - Distâncias, em mm, entre a espécie E. flammea e as demais espécies	101
Tabela 29 - Distâncias, em mm, entre a espécie E. ignita (espécie 3) e as demais espécies.....	103
Tabela 30 - Distâncias, em mm, entre a espécie E. imperialis e as demais espécies	105
Tabela 31 - Distâncias, em mm, entre a espécie E. orellana e as demais espécies	107
Tabela 32 - Resultados dos testes realizados com o algoritmo FNN4Bees.....	119
Tabela 33 - Resultado dos testes de validação cruzada das Espécies entre 0 a 8 do conjunto de dados.....	122

Tabela 34 - Resultado dos testes de validação cruzada das Espécies entre 9 a 17 do conjunto de dados.....	122
Tabela 35 - Resultado dos testes de validação cruzada das Espécies entre 18 a 26 do conjunto de dados.....	123
Tabela 36 - Resultado dos testes de validação cruzada das Espécies entre 27 a 35 do conjunto de dados.....	123

SUMÁRIO

1 INTRODUÇÃO	12
1.1 OBJETIVO	15
1.2 JUSTIFICATIVA	17
2 CLASSIFICAÇÃO DE ABELHAS	18
2.1 CLASSIFICAÇÃO DE ABELHAS POR MEIO DAS ASAS	18
2.2 MORFOMETRIA GEOMÉTRICA E O ESTUDO DA FORMA	19
2.3 ANÁLISE ESTATÍSTICA DA FORMA (AEF)	37
2.4 ALINHAMENTO E PADRONIZAÇÃO DOS DADOS	38
2.5 O ALGORITMO DOS K-VIZINHOS MAIS PRÓXIMOS (KNN)	43
3 LÓGICAS COMPUTACIONAIS	46
3.1 REVISÃO DA LÓGICA FUZZY TRADICIONAL	46
4 A PROPOSTA FUZZY KNN (FNN4BEES)	54
4.1 DESCRIÇÃO DO FNN4bEES	56
4.1.1 Seleção de Características	58
4.1.2 Extração de Características	63
4.1.3 O algoritmo kNN estendido	68
4.1.4 Análises Estatísticas para Identificação da Ordem de Significância	72
4.1.5 Aplicando a Fuzzy kNN (FkNN)	79
4.2 ESTUDO DE CASO – EUGLOSSAS	85
4.2.1 Aplicação da Fuzzy kNN para as Euglossas	98
5 AMBIENTE DE IMPLEMENTAÇÃO	109
5.1 FUNCIONAMENTO DO ALGORITMO FNN4BEES NO DIAGRAMA DE ATIVIDADES	110
5.1.1 Descrição do método aplicaExpressaoRegular	113
6 CLASSIFICAÇÃO E RESULTADOS	115
6.1 ANÁLISE DOS RESULTADOS DA FUZZY kNN	118
6.2 ANÁLISE DO ALGORITMO EM COMPARAÇÃO COM OS RESULTADOS OBTIDOS PELAS CLASSIFICAÇÕES ATUAIS	119
7 CONCLUSÕES	123
REFERÊNCIAS	

1 INTRODUÇÃO

A classificação de espécies surge da necessidade de conhecer e catalogar espécies em diferentes níveis, comparando características comuns. Das 3.187 morfo espécies reconhecidas em 46 levantamentos realizados no território nacional até o ano de 2002, apenas 26% foram seguramente identificadas, sendo que 6% foram representados por espécies novas, (SILVEIRA et al., 2002).

De acordo com Franco et al. (2008a), existem diversos esforços vinculados à preservação da biodiversidade e à sua manutenção, porém, a falta de taxonomistas, chaves de identificação e revisões dos grupos de espécies atuais dificultam o trabalho de identificação das espécies já descritas para o público em geral.

Desta maneira, faz-se necessário a implementação e a aplicação de novas ferramentas e métodos que permitam analisar a biodiversidade de outras maneiras que não por meio da taxonomia tradicional. Em abelhas, dados preliminares mostram a efetividade da técnica de Morfometria Geométrica nas asas anteriores para a discriminação das espécies estudadas, (FRANCOY et al., 2008a). Em Franco et al. (2006b), o autor realizou uma Análise Morfométrica populacional de abelhas africanizadas com o auxílio do software ABIS (Automated Bee Identification System).

O trabalho realizado por Franco et al. (2006b) teve por objetivo a comparação morfométrica entre as abelhas de todas as populações brasileiras amostradas; comparação entre as populações de *A. m. scutellata*, abelhas africanizadas coletadas no início do processo de africanização e abelhas de populações coletadas próximo ao ponto de introdução das abelhas africanas no Brasil; comparação de todas as populações, uma a uma com todas as suas ancestrais para determinar o grau de influência atual de cada uma delas.

Existem atualmente diversas metodologias utilizadas para a classificação de espécies de abelhas, sendo que duas delas podem ser destacadas. A primeira metodologia se refere à Análise de Discriminantes, derivada nos dois seguintes métodos: Análise de Discriminantes Lineares (LDA) e Análise de Discriminantes de Kernel (KDA) que utiliza funções discriminantes não lineares. A segunda

metodologia é baseada na Morfometria Geométrica que tem se mostrado muito útil em estudos de identificação de espécies de abelhas, segundo Drauschke et al. (2007). A Morfometria Geométrica utiliza uma série de métodos que, de acordo com Francoy et al. (2008a), "[...] visam descrever e representar a geometria das *Formas* estudadas por meio da Análise de Deformações Relativas dos padrões de nervura de asas de abelhas[...]". De acordo com Slice (2007), "em uma amostra de espécies homogêneas (com os espécimes da mesma espécie), o PCA (Análise de Componentes Principais) serve como um estimador de características da população". Outros métodos que podem ser citados incluem, de acordo com Rohlf (1999), "Análise Parcial dos Mínimos Quadrados, Análises Estatísticas para estimação da média e covariação da *Forma* que descrevem diferenças entre as espécies e a Análise Multivariada de Dados".

O estudo realizado primeiramente por Shröder et al. (1995) levou ao desenvolvimento de diversas ferramentas de automação do processo de identificação de asas de abelhas por meio de um sistema semi automático de obtenção de medidas, classificação e identificação de espécies, resultando no ABIS. De acordo com Francoy et al. (2008a) "[...] esse software é capaz de fornecer a detecção automática dos pontos de junção das nervuras e colocação automática dos marcos anatômicos (*landmarks*) das asas". O ABIS utiliza Análises de Discriminantes não lineares de Kernel que permitem uma melhor classificação das espécies. Porém, os estudos de Francoy et al. (2006a) mostraram que algumas espécies de abelhas são impossibilitados de serem reconhecidos por esse processo de identificação implementado no ABIS.

O estudo realizado neste trabalho propõe uma alternativa para o reconhecimento de espécies de abelhas, buscando padrões por meio da Morfometria Geométrica das asas utilizados como caracterizadores de espécies. Os métodos classificatórios utilizados envolvem o algoritmo dos k-Vizinhos mais Próximos estendido, Métodos Estatísticos e a Lógica Fuzzy (ou Lógica Difusa) **kNN** (**k- Nearest Neighbours**). A análise implementada pela Fuzzy kNN inclui a representação de um processo capaz de lidar com as variantes indesejadas que causam ruídos em qualquer processo de classificação, como a presença de erros de digitalização nas amostras dos dados do conjunto, variações genéticas que podem ocorrer no processo natural de

desenvolvimento das espécies, entre outros. No capítulo Lógicas Computacionais, descrito neste trabalho, são discutidos os algoritmos que envolvem a implementação do Sistema de Apoio a Decisão (SAD), baseado no kNN estendido e na Fuzzy kNN (FkNN) e, ainda, como foram realizadas a seleção e extração de característica das asas das abelhas, incluindo os procedimentos de Morfometria Geométrica de pré-processamento das imagens.

A Figura 1 apresenta a descrição do modelo criado por Bueno (2010) e foi extraída da proposta de tese do autor. Na Figura 1, nota-se o Subsistema de Classificação, no qual aplicam-se algumas técnicas para um novo processo de classificação a partir das asas das abelhas. Neste subsistema encontra-se o classificador, que utiliza métodos como o SVM (Support Vector Machines), RNA (Redes Neurais Artificiais), métodos estatístico compostos para realizar o processo de reclassificação das espécies de abelhas.

Figura 1 - Arquitetura de Referência para o Sistema de Identificação e Classificação Taxonômica de Abelhas sem Ferrão

1.1 OBJETIVO

O objetivo deste trabalho é propor um algoritmo para a classificação de abelhas baseado no método de identificação de marcos anatômicos por meio do algoritmo kNN e da Fuzzy kNN. A criação de um novo método para a classificação de espécies de abelhas deve melhorar a classificação atual feita apenas por métodos estatísticos e servir para a classificação de espécies que não são possíveis de classificação pelo ABIS. Para efetivar o objetivo é implementado um algoritmo estendido para a adaptação dos dados de entrada aos dados biológicos, juntamente com a agregação do conhecimento do especialista no modelo computacional implementado no SAD. Além disso é necessário conhecer como os marcos anatômicos são representativos para a classificação e como transformar os pontos obtidos pelos marcos anatômicos em informações relevantes para o processo de classificação. Para isso é realizada uma análise computacional por meio do algoritmo kNN estendido nestes dados. Isto faz com que o processo de agrupamento não supervisionado seja o mais fiel possível às análises realizadas pelos especialistas.

A metodologia aplicada neste processo leva em consideração que, tanto o método de captura das imagens, como a padronização dos dados são os mesmos realizados pelo especialista. Ainda é necessária a utilização de métodos computacionais para a extração e seleção de características relevantes das imagens para classificação.

Para alcançar o objetivo, as seguintes metas foram adotadas no algoritmo de classificação das abelhas:

1. Realizar o pré-processamento dos dados;
2. Selecionar e extrair as características dos marcos matemáticos;
3. Identificar, a partir da Morfometria Geométrica, os dados e a *Forma* por meio de análises do gráfico de dispersão das amostras;
4. Aplicar o algoritmo kNN, estendendo-o para o problema da classificação de abelhas;

5. Aplicar a Análise Estatística;
6. Determinar a Ordem de Significância dos marcos matemáticos;
7. Classificar os marcos matemáticos por meio do algoritmo kNN estendido;
8. Modelar a Fuzzy kNN;
9. Comparar os resultados obtidos pelas classificações atuais feitas pelos biólogos.

As etapas listadas acima constituem, juntas, o processo envolvido na classificação de abelhas e são detalhadas nos capítulos seguintes.

O pré-processamento de dados envolve o alinhamento e a padronização dos dados que são recebidos a partir dos especialistas. Estes dados devem conter as informações referentes aos marcos anatômicos aplicados nas imagens das asas de abelhas, assim como o nome de cada espécie e a quantidade de marcos anatômicos colocados. A seleção e extração de características acontecem logo após o processo de normalização dos marcos anatômicos das amostras do conjunto de dados. Esta etapa tem como objetivo buscar quais são as variáveis que caracterizam, da melhor maneira, as espécies do conjunto de dados por meio da distância euclidiana dos marcos anatômicos. Em seguida, aplica-se o algoritmo kNN estendido para análise computacional das distâncias entre os marcos anatômicos. A Análise Estatística utiliza esses dados, adquiridos no algoritmo kNN, para calcular o desvio padrão e a variância desses e criar a Ordem de Significância, em que somente os marcos anatômicos que melhor representam a espécie em análise são escolhidos e armazenados para o próximo cálculo formando o padrão da espécie. Após determinação da Ordem de Significância, obtêm-se uma primeira classificação a partir do algoritmo kNN estendido. Essa classificação é então melhorada pela otimização do processo iterativo da busca pelos melhores padrões das espécies e, a partir do resultado, já otimizado, aplica-se a Fuzzy kNN para uma nova classificação. O último passo realiza a comparação dos resultados obtidos entre a classificação realizada e os dados provenientes das classificações já realizadas e publicadas pelos especialistas.

1.2 JUSTIFICATIVA

Pela necessidade de preservação e manutenção da biodiversidade, muitos esforços estão sendo aplicados em pesquisas e desenvolvimento de ferramentas que possam auxiliar e melhorar o processo. Dificuldades podem ser encontradas devido à falta de taxonomistas, chaves de identificação de espécies e revisões dos grupos de espécies atuais. Os métodos implementados neste trabalho devem atender à classificação de espécies de abelhas a fim de melhorar os resultados atuais de classificação alcançados pelos especialistas. Além disso é premente a necessidade de pesquisa de um processo para efetuar a classificação de espécies de abelhas de maneira que seja inserido o conhecimento do especialista para a otimização do processo e aprimoramento dos resultados. Os métodos atuais utilizados para resolver o problema de classificação de abelhas se baseiam em Análises Estatísticas e por meio de algoritmos de análise de imagens implementados em ferramentas para identificação de contornos e *Formas*. Em sua maioria, essas ferramentas não estão preparadas para o recebimento de dados provenientes de informações biológicas, ou não são suficientes para lidar com todas as informações que esses dados representam. A justificativa de implementação de um algoritmo para classificação de espécies de abelhas atende ao requisito de conseguir lidar com esses dados, para extrair as informações que eles representam com o auxílio de métodos estatísticos no estudo da *Forma* e métodos de reconhecimento de padrões para classificação correta dessas espécies. O algoritmo proposto, denominado **FNN4Bees** (**F**uzzy **kNN** **f**or **B**ees) implementa tanto a Análise Estatística, quanto a inserção do conhecimento biológico para classificação das espécies por meio das análises computacionais inseridas no kNN e na Fuzzy kNN.

2 CLASSIFICAÇÃO DE ABELHAS

Existem atualmente diversas metodologias empregadas no estudo da estruturação genética das populações de abelhas para identificação e distinção das diferentes espécies (COLLET et al., 2006). Os métodos de classificação de abelhas, realizados pelos especialistas, incluem, de acordo com Francoy (2007) análises a partir da covariação de medidas lineares da *Forma* das imagens das abelhas, análises a partir da Morfometria Geométrica dos marcos anatômicos das imagens das Asas das abelhas, análises de DNA mitocondrial de abelhas e análises estatísticas dos dados provenientes da Morfometria Tradicional e Geométrica. O método envolvido neste trabalho é proveniente da análise da Morfometria Geométrica a partir das imagens das Asas das abelhas. Esses dados de imagens são utilizados para a classificação aplicados a uma lógica computacional. As seguintes seções detalham esse processo.

2.1 CLASSIFICAÇÃO DE ABELHAS POR MEIO DAS ASAS

De acordo com Mckelvey (1982), "[...] o estudo da classificação de espécies faz parte da Taxonomia e existem diversos métodos taxonômicos que podem ser aplicados no processo de classificação [...]". Um deles trata da Morfometria Geométrica. De acordo com Richtsmeier et al. (2002), "A Morfometria Geométrica é um campo de pesquisa que apresenta resultados satisfatórios atestados em classificação de dados biológicos [...]". A principal técnica de análise em Morfometria Geométrica é por meio do conceito de *Forma* e como representar a *Forma* a partir de marcos anatômicos. Em Sole et al. (2008), foram realizadas pesquisas sobre Morfometria Geométrica com a Análise de Procrustes, Análise de Discriminantes Lineares (LDA) e Análise de Componentes Principais (PCA) incluindo dados biológicos. A partir da compreensão dos conceitos de Morfometria Geométrica e da *Forma* nos dados biológicos, diversos métodos e algoritmos

podem ser aplicados para o processo de classificação, como a Análise de Componentes Principais e a Análise de Deformações Relativas (Relative Warps). De acordo com Richtsmeier et al. (2002), outros métodos podem ser combinados, como no caso de agregação de métodos estatísticos para maior compreensão da *Forma* e para aumentar a taxa de acerto nas classificações das espécies.

2.2 MORFOMETRIA GEOMÉTRICA E O ESTUDO DA FORMA

Segundo Slice (2007), “A Morfometria é um campo da Análise Estatística Multivariada com foco em procedimentos das coordenadas cartesianas de pontos anatômicos que podem ser utilizados para definir as medidas tradicionais[.]”. A análise Morfométrica em geral se baseia na análise da *Forma* de marcos anatômicos que são colocados em pontos estratégicos da figura ou imagem para análise classificatória ou não. Os marcos definem um espaço representativo que deve servir como caracterizador da *Forma*. Os pontos neste espaço são colocados em relação ao interesse do especialista na análise da *Forma*. Neste trabalho, a aplicação de marcos é devida à classificação realizada entre espécies de abelhas no estudo de suas *Formas*. Esses marcos devem definir um padrão de caracterização da espécie único e preciso para o processo de classificação. Quanto mais marcos anatômicos forem incluídos na análise, o número ou quantidade de distâncias mínimas necessárias para o ajuste da posição do marco anatômico relativo cresce ao ponto da classificação entre as espécies se tornar impraticável, ou seja, quanto mais marcos anatômicos são colocados para a análise da imagem da asa da abelha, em relação à sua *Forma*, mais difícil pode ficar de se conseguir maior precisão na classificação, ao ponto em que, se forem colocados muitos marcos, a classificação se torna impossível.

O algoritmo implementado neste trabalho tenta resolver esse problema criando uma Ordem de Significância, que exclui da Ordem os marcos anatômicos que, após análise no kNN, podem não ter significado matemático algum, limitando o número de vizinhos mais próximos. Esse problema pode ser visto da seguinte maneira:

supondo que uma espécie tenha os marcos colocados em posições P1, P2, P3, P4,..., P33 e outra espécie tenha a mesma quantidade de pontos colocados e que, dentre esses pontos, alguns são pontos aleatórios que marcam contornos ou pontos médios entre as junções das nervuras nas asas. Então, alguns pontos de encontro de nervuras de uma espécie podem se confundir com pontos médios ou de contorno da outra espécie. Os pontos colocados nas imagens das asas devem descrever a *Forma* dessa espécie. Alguns pontos colocados a mais podem definir a mesma *Forma* para duas ou mais espécies diferentes, dependendo da dimensionalidade dos dados. Uma análise pelo kNN visa eliminar os marcos anatômicos que perturbam a classificação, deixando apenas os marcos que melhor caracterizam a espécies, chamados de padrões da espécie.

Outros aspectos podem ser notados a respeito dos marcos anatômicos. De acordo com Slice (2007),

As coordenadas dos marcos anatômicos devem ser armazenadas com respeito ao eixo de coordenadas definido, e seus valores numéricos devem refletir a localização e orientação única de cada espécime em respeito a esses eixos.

Adicionalmente, o tamanho da espécie também estará codificado nessas coordenadas. A Figura 2 mostra diversas aplicações de colocações de marcos anatômicos em asas e as análises devem partir desses marcos em relação às suas *Formas*.

Figura 2 - Imagem de *Brassolis Astyra* Godart, 1824. Marcos Anatômicos utilizados nas análises das pupas: 1, vista lateral; 2, vista Frontal. Marcos Anatômicos utilizados nas análises das asas: 3, anterior; 4, posterior

O método atualmente mais utilizado para análise de *Formas* é o método de Análise de Procrustes (Procrustes Analysis) que é baseado na estimativa de translação, rotação e escalonamento de parâmetros dos mínimos quadrados, segundo Ross (2005).

Kendall (1989) traça um paralelo entre as *Formas* dos elementos com a autenticação estatística. A questão envolvida é indicar qual o lugar matemático natural para a *Forma* de um conjunto rotulado de k pontos não totalmente coincidentes em m dimensões. A idéia é filtrar efeitos resultantes da translação, mudanças de escala e rotação, sem considerar esses fatores. A Figura 3 mostra um exemplo de imagem de asa que é utilizado no processo de classificação. As imagens, como no exemplo abaixo, são armazenadas em diretórios específicos.

Figura 3 - Exemplo de imagem da espécie *Friesella schrottkyi* (operária)

A imagem apresentada na Figura 3 foi extraída de www.ib.usp.br/beesp/friesella_schrottkyi.htm, acessado em 01 de março de 2010.

A Figura 4 apresenta a colocação de marcos anatômicos e mostra graficamente os dados de quatro amostras da espécie *Friesella schrottkyi* isentos das técnicas de translação, mudanças de escala e rotações (técnicas de alinhamento).

Figura 4 - Apresentação dos marcos anatômicos de quatro amostras da espécie 0 (*Friesella schrottkyi*) sem alinhamento

Cada cor (ou *Forma Geométrica*), apresentada na Figura 4 representa uma amostra diferente. A Figura 5 mostra os mesmos dados alinhados (com a aplicação de translação, mudança de escala e rotações).

Figura 5 - Apresentação de quatro amostras da espécie 0 (*Friesella schrottkyi*) alinhadas

Uma das técnicas utilizadas para eliminar diferenças no tamanho é por meio do cálculo do centróide, mas, como apresentado em Kendall et al. (1999), "[...] esta não é a única maneira de se resolver o problema de alinhamento dos dados, pois, existem contextos nos quais vale considerar outras técnicas [...]". A *Forma* é discutida em Kendall et al. (1999) com a introdução do espaço da *Forma* associado com k pontos rotulados em R_m e a identificação de alguns espaços de *Forma* mais simples em que se organizam os espaços em um vetor bidimensional que é utilizado para encontrar características fracas, mas globais (homólogas) em todo o espaço de *Forma* com precisão suficiente para distinguir um espaço dos outros. A análise da reprodução desse procedimento é dada por eq.(1) e eq.(2).

- *Forma* = imagem da asa da abelha gerada pelos marcos anatômicos (1)
- Vetor[] = marcos anatômicos colocados nas junções das nervuras de cada *Forma* (2)

O estudo de *Formas* realizado por Kendall et al. (1999) transformou a maneira com que objetos similares são dispostos matematicamente. A configuração de todas as *Formas* pode ser reescalada na distância mínima (D_F) de uma *Forma* de referência, como em eq.(3) e eq.(4).

Sendo:

- Tamanho do centróide da *Forma* de referência = 1 (3)
- Tamanho do centróide de todas as outras *Formas* = $\cos(N)$, em que N = distância de Procrustes da *Forma* de referência. (4)

Isto define uma nova hiperesfera do espaço da *Forma* com raio 0.5 e tangencial ao espaço da *pré-Forma* orientado da *Forma* de referência. Isto é, o espaço da *Forma* de Kendall.

O espaço orientado da *pré-Forma* é apresentado em eq.(5):

- Localização centralizada, tamanho do centróide unitário, rotação para minimizar a distância parcial de Procrustes e mais adiante a distância de Procrustes (5)

Seguindo a linha de raciocínio sobre a *Forma*, Ross (2005) define que “A correspondência da *Forma* é um aspecto muito importante em imagens[..]”. A correspondência e análise da *Forma* têm papel essencial na validação de algoritmos criados para colocarem os marcos anatômicos em locais exatos e precisos. A análise de Procrustes é uma análise rígida sobre a *Forma* e usa o dimensionamento isomórfico, a translação e a rotação para encontrar o que melhor se encaixa entre dois ou mais espaços marcados. A Análise Ortogonal de Procrustes Generalizada (GPA) é a técnica utilizada no software *TPSRe/w¹* para a extração de características das asas das abelhas com ou sem ferrão. Essa técnica é considerada atualmente a mais utilizada para se descobrir a correspondência de uma *Forma* com outras. Para

¹ Relative warps analysis. Versão 1.46. Desenvolvido por F. James Rohlf.

Ross (2005), "*Formas e marcos anatômicos* são dois conceitos importantes envolvidos no GPA". A definição dada à *Forma*, segundo Ross (2005), deve ser conhecida *a priori* para o conhecimento de como se corresponder às *Formas*.

Forma é toda a informação geométrica que permanece quando os efeitos de rotação, dimensionamento e localização são filtrados de um objeto.

Esta tradução do que significa *Forma* segue a linha de raciocínio de Kendall (1989) sobre a *Forma*, como foi descrito anteriormente. Por esta definição existem transformações que podem ser feitas para permitir movimento e mudanças, preservando a *Forma* original. Dimensionamento isomórfico, translação e rotação são três transformações usadas no alinhamento de *Formas*. Essas transformações de preservação da *Forma* são chamadas, de acordo com Ross (2005) de “transformações de similaridade Euclidiana”. A definição de marcos anatômicos seria, de acordo com Ross (2005), “Um conjunto finito de pontos em uma superfície que descreve a *Forma* de maneira exata e correta”, ou seja, os marcos anatômicos podem ser definidos como um conjunto finito de pontos sobre a superfície da *Forma* que a representam de maneira inequívoca. Existem três tipos de marcos de acordo com Ross (2005):

- **Marcos anatômicos:** os quais o especialista (por exemplo, um biólogo) assina pontos que representarão um objeto biológico ou objetos de qualquer espécie.
- **Marcos matemáticos:** pontos assinados para um propósito matemático (como exemplo para uma curvatura).
- **Pseudo – Marcos:** pontos localizados entre outros dois tipos de pontos (marcos anatômicos) ou pontos em torno do contorno da *Forma*.

A descrição do algoritmo GPA é apresentada pela seguinte maneira.

1. Selecione uma *Forma* para ser a média aproximada da *Forma*;
2. Alinhe as *Formas* à média aproximada:

- a. Calcule o **centróide** de cada *Forma* (ou de um conjunto de marcos anatômicos);
 - b. Alinhe todos os centróides das *Formas* à origem;
 - c. Normalize cada *Forma* no tamanho do centróide;
 - d. Rotacione cada *Forma* para alinhar com a média aproximada da *Forma*.
3. Calcule a nova média aproximada das *Formas* alinhadas;
 4. Se a média aproximada nos passos 2 e 3 forem diferentes, retorne para o passo 2, senão, foi encontrado a verdadeira média da *Forma* do conjunto.

O cálculo do centróide de uma superfície limitada por curvas definidas por equações algébricas é normalmente determinado recorrendo aos integrais da seguinte maneira, como mostram as equações eq.(6), eq.(7) e eq.(8).

$$(Ax_G = \int x_c dA) \quad (6)$$

$$(Ay_G = \int y_c dA) \quad (7)$$

$$(Az_G = \int z_c dA) \quad (8)$$

Para as *Formas* para as quais a área ainda não for conhecida, pode ser calculada a partir dos integrais com $dA = dx dy$, indicado na eq.(1). Para uma área pequena e coordenadas do centróide do elemento da superfície, dA , x_c , y_c e z_c , são expressas em função das coordenadas de um ponto localizado sobre a curva que limita a área em questão. De acordo com Mousseau (1991),

Existe uma necessidade de se medir algumas estruturas e relacionar as variações em *Formas* ou tamanhos à taxa de crescimento ou idade para hipóteses evolucionárias relacionadas às variações geográficas e adaptação ou para questionar a sistemática.

Em alguns casos, medidas lineares simples representam os complexos eventos de desenvolvimento que frequentemente descrevem a variação associada com a

hipótese evolucionária. Neste caso, medição e análise são tarefas relativamente triviais. Mas na maioria dos casos, medidas simples podem ser deficientes e, a aplicação da análise morfométrica multivariada pode ajudar no processo.

Em Mousseau (1991) é descrito que: “Marcos anatômicos são pontos nos quais as explicações *epigenéticas* são determinadas”, em que epigenéticas, de acordo com Bird (2007), é um termo utilizado na biologia para se referir a características de organismos unicelulares e multicelulares, como as modificações de cromatina e DNA, que são estáveis ao longo de diversas divisões celulares, mas que não envolvem mudanças na sequência de DNA do organismo.

Marcos anatômicos seriam pontos de uma explicação do processo biológico envolvido no desenvolvimento da estrutura da espécie. Os melhores marcos anatômicos são pontos individualmente reconhecidos nas fronteiras entre as regiões de histologia distintas. De acordo com Ross (2005), “[...] o ingrediente chave da abordagem de Bookstein (1989) é a proposição de que 'o uso apropriado de marcos irá ligar três partes separadas da área da Morfometria: a geometria dos dados, a matemática da deformação (mudanças na *Forma* e tamanho) e as explicações biológicas'[...]”.

O formalismo em torno dos marcos anatômicos pode ser utilizado para delinear as explicações como desenvolvimento, evolução e taxonomia para simplesmente definir localizações no espaço. No artigo de Adams et al. (2004) é feita uma análise histórica dos avanços em dez anos que a Morfometria Geométrica trouxe para as pesquisas envolvendo dados biológicos.

De acordo com os autores, a partir de Bookstein (1998) a comparação de características anatômicas de organismos tem sido o elemento central da biologia por séculos. A classificação taxonômica dos organismos, e o entendimento da vida biológica, foram ambos historicamente baseados nas descrições morfológicas.

A partir disso, biólogos começaram a transição de um campo descritivo para uma ciência quantitativa, e a análise da morfologia passou por uma revolução quantitativa. Na metade do século XX, a descrição quantitativa da morfologia foi combinada com Análises Estatísticas para descrever padrões das variações das *Formas*, e o moderno campo da Morfometria começou. De acordo com Bookstein (1989), “a Morfometria é o estudo das variações das *Formas* e suas covariâncias

com outras variáveis”. Tradicionalmente, Morfometria era o campo da aplicação de Análises Estatísticas Multivariadas em conjuntos de variáveis quantitativas como: comprimento, tamanho e altura, mas nos anos 80 ocorreu uma mudança neste paradigma no modo como as estruturas morfológicas eram quantificadas e como os dados eram analisados. Esta variância foi denominada de Morfometria Geométrica, (ADAMS et al., 2004).

A análise em Morfometria Geométrica, geralmente, é acompanhada de Análises Estatísticas sobre os dados gerados a partir dos métodos provenientes da Morfometria Geométrica. Essas Análises Estatísticas tipicamente incluem a Análise de Componentes Principais, Análise de Fatores, Análises de Variáveis Canônicas (CVA) e Análise de Funções Discriminantes.

A Figura 6 mostra a aplicação da Análise de Componentes Principais em quatro amostras da espécie *E. chaly*.

Figura 6 - Aplicação da Análise de Componentes Principais nos dados da espécie *E. chaly*.

Tabela 1 - Análise de variância da matriz de correlação

Eigenvalue	109,32	36,40	6,99	0,92	0,06	0,02	0,01	0,00
Proportion	0,711	0,237	0,045	0,006	0,000	0,000	0,000	0,000
Cumulative	0,711	0,948	0,993	0,999	1,000	1,000	1,000	1,000

Tabela 2 - Apresentação das variáveis e seus respectivos termos para cada componente principal

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
C1	0,453	-0,173	0,464	-0,129	-0,709	-0,128	-0,115	0,030
C2	0,091	0,459	0,162	0,192	-0,185	0,537	0,620	-0,104
C3	0,557	-0,074	-0,284	0,731	0,086	-0,211	0,008	-0,131
C4	0,083	0,485	0,130	0,050	0,102	-0,381	0,067	0,761
C5	0,474	-0,187	0,467	-0,248	0,658	0,123	0,109	-0,033
C6	0,088	0,462	0,146	0,132	0,065	0,384	-0,762	-0,087
C7	0,483	0,112	-0,649	-0,514	-0,080	0,209	-0,007	0,136
C8	0,062	0,507	0,028	-0,257	0,042	-0,547	0,073	-0,604

Os eixos vertical (Eigenvalue) e horizontal (Component Number) da Figura 6 representam, respectivamente, os valores de variância e os números de componentes da espécie 0.

De acordo com a Figura 6 e das Tabelas 1 e 2, o primeiro componente principal tem variância (eigenvalue) de 109,32 e corresponde a 71,1% da variância total. Os coeficientes listados sobre PC1 mostram como calcular os componentes principais, de acordo com a eq.(9).

$$PC_i = \sum c_i, \forall i = \{1, \dots, 8\} \quad (9)$$

Pode-se notar que a interpretação sobre os componentes principais é subjetiva, entretanto, padrões triviais aparecem muito frequentemente. Por exemplo, pode-se avaliar o primeiro componente como representante das variáveis C1, C3, C5, e C7 porque o coeficiente desses termos têm o mesmo sinal e não estão muito próximos de zero.

O segundo componente principal tem variância de 36,40 e corresponde a 23,7% da variabilidade dos dados. Ele é calculado a partir dos dados originais utilizando-se os coeficientes listados abaixo de PC2. Assim segue-se com o cálculo e análise dos demais componentes.

O objetivo fundamental da Análise de Componentes Principais é a redução da dimensão dos espaços. Existem vários critérios para encontrar o número p_r de eixos a serem eliminados, balanceando a redução da dimensão do espaço com a necessidade de explicar uma proporção importante da variância total. Numa nuvem esférica, sem alongamentos preferenciais, os valores próprios (eigenvalues) resultantes de uma análise normalizada são todos correspondentes a eq.(10), (SOUZA, 2003).

$$\lambda_\rho = 1 \quad \forall \rho \in \{1, \dots, p\} \quad (10)$$

Então, pode-se escolher p_r como o número de eixos α , tal que λ_α seja maior ou igual a 1. Seja t uma percentagem da inércia total fixada previamente, normalmente 80%, então p_r é o número de eixos. Seja a curva que relaciona o número de ordem de cada eixo com o valor próprio que lhe está associado (apresentado na Figura 6). Caso a curva evidenciar uma estabilização dos valores próprios, pode-se reter apenas os eixos com números de ordem superiores àqueles que iniciam a estabilização, (SOUZA, 2003).

A homologia das distâncias lineares é de difícil acesso, porque muitas distâncias, como o comprimento máximo, não são definidas pelos pontos homólogos. Por causa dessas dificuldades, pesquisadores exploraram métodos alternativos para quantificar e analisar a morfologia. Dados que capturavam a geometria das estruturas morfológicas eram de particular interesse e métodos para analisar esses dados foram desenvolvidos. Alguns desses métodos são descritos a seguir, (SOUZA, 2003).

- **Métodos descritivos** – Foram os primeiros métodos de Morfometria Geométrica utilizados. A abordagem geralmente utilizada era digitalizar pontos ao longo de um contorno, ajustar os pontos em uma função matemática (geralmente a Análise de *Fourier*) e então comparar as curvas usando o coeficiente das funções como variáveis da *Forma* na Análise Multivariada. Os pontos neste espaço de parâmetros multivariados poderiam

ser transformados novamente no espaço físico do organismo e seus contornos poderiam ser visualizados.

- **Métodos de marcos anatômicos** – métodos de Morfometria Geométrica baseados em marcos anatômicos começaram com uma coleção de coordenadas com duas ou três dimensões de marcos biológicos definíveis. Uma análise direta dessas coordenadas como variáveis poderia ser inapropriada, pois os efeitos da variação de posição, orientação e escala entre as espécies ainda estaria presente. Portanto, depois da exclusão matemática dessas variações, as variáveis se tornam variáveis da *Forma* e assim podem ser utilizadas em amostras estatisticamente comparáveis e representações gráficas da *Forma* podem ser geradas. Diversos métodos foram propostos, cada um usando pequenas diferenças nos critérios de protocolos e otimização como o GPA (General Procrustes Analysis), que é descrito pelos itens a seguir.
 - Primeiro o centróide de cada configuração é transladado à origem, e configurações são escaladas para um ponto em comum, em um único tamanho, por meio da divisão pelo tamanho do centróide.
 - Finalmente as configurações são otimamente rotacionadas para minimizar as diferenças quadradas entre os marcos anatômicos correspondentes.

O próximo passo da Morfometria Geométrica, após a colocação dos marcos anatômicos é sobrepor as imagens. Na sobreposição, as diferenças nas *Formas* podem ser descritas pelas diferenças nas coordenadas dos marcos anatômicos correspondentes entre os objetos. Essas diferenças são usadas como os dados nas comparações multivariadas das variações das *Formas*, (KUMAR et al., 2003). Os parâmetros que descrevem estas deformações (média dos desvios parciais) podem ser utilizados como variáveis da *Forma* entre a população. Muitos procedimentos alternativos para obter informação sobre a *Forma* por meio dos dados dos marcos anatômicos têm sido propostos. Entre eles incluem-se: Análise da Matriz da Distância Euclidiana (EDMA), Análise de Dimensionamento dos Elementos Finitos

(FESA), métodos baseados nos ângulos interiores e, a partir da proposta deste trabalho, métodos computacionais de Reconhecimento de Padrões baseados no algoritmo dos kNN e Fuzzy kNN. De acordo com Adams et al. (2004),

Pode-se realizar uma análise dos dados a partir dos marcos anatômicos, uma vez que quase todos os estudos baseados na Morfometria dos marcos anatômicos analisam a *Forma* com os procedimentos do espaço da *Forma* de Kendall, distâncias de Procrustes e as aproximações da tangente do espaço.

O espaço da *Forma* de Kendall tem o melhor resultado estatístico, o menor erro médio quadrado, e apresenta mínimas restrições nos padrões das variações que podem ser detectados. As espécies alinhadas pelo GPA apresentam pontos que podem ser projetados em um espaço que é tangente ao espaço das *Formas* de Kendall, (DRYDEN e MARDIA, 1998). Uma outra maneira de realizar análise morfológica nos dados se deve à utilização da Análise de Deformações nas Placas Finas. Além delas, para se desenvolver uma análise morfológica geométrica, qualquer projeção ortogonal ao espaço tangencial é suficiente, (ADAMS et al., 2004). O último passo da Análise Morfométrica envolve uma visualização gráfica dos resultados das Análises Estatísticas. Existem outras formas de abordagens por meio dos marcos anatômicos que são descritas em Adams et al. (2004), como comparações de amostras de *Formas* usando cálculo dos tamanhos correlacionados das distâncias entre os pares de marcos anatômicos e a comparação de *Formas* usando ângulos a partir da triangulação dos marcos anatômicos. O processo de descobrimento da importância da *Forma* no sentido matemático levou ao surgimento de diversas técnicas que adotaram os pontos anatômicos para representação da *Forma*.

Um exemplo de utilização da técnica de Deformação usando placas finas (que é a técnica utilizada para padronização e alinhamento dos dados das imagens das abelhas) está descrita em Whitbeck e Guo (2006) no qual o autor mostra como as deformações das imagens têm sido de enorme importância em reconhecimento de faces e apresenta muitas aplicações e contribuições para a ciência da computação que estuda reconhecimento de imagens, imagens biológicas e segurança

doméstica. Ele apresenta o popular algoritmo desenvolvido por Alex Rosen, conhecido como *Alex-Warp*. Também apresenta uma proposta alternativa ao algoritmo de Alex utilizando interpolação entre as placas finas. O conceito básico é permitir ao usuário que ele entre com diferentes conjuntos e pontos de marcos anatômicos destinatários versus a abordagem um por vez usada no *Applet Alex Warp*. Em Ross (1964), no princípio da taxonomia numérica, que é uma vertente da taxonomia por meio de marcos anatômicos, são elucidados problemas e premissas referentes à taxonomia numérica e como ela se difere das outras metodologias. A taxonomia numérica, segundo Ross (1964), é definida como “[...] a avaliação numérica da afinidade ou similaridade entre unidades taxonômicas e a ordenação dessas unidades em taxas baseadas nas afinidades [...]”.

Em Dos Reis (2002), a Morfometria Geométrica é descrita como uma metodologia usada para estudar a variação de *Forma* da configuração de marcos anatômicos arquivados em um sistema de coordenadas cartesianas. Em Bookstein (1996), as análises em Morfometria Geométrica são antecidas pelas três operações dadas a seguir.

1. Conversão dos dados de marcos anatômicos digitalizados de cada organismo em um plano. Em seguida, a configuração de marcos anatômicos de cada organismo estudado é transladada, isto é, a configuração de marcos anatômicos é centralizada de modo que, o centróide (ponto médio ou centro de massa) de cada configuração de marcos anatômicos, coincida com a origem do sistema cartesiano.
2. Posteriormente, cada configuração de marcos anatômicos homólogos é escalonada, na qual o tamanho é definido em termos do centróide da configuração. O tamanho do centróide é definido como a distância de cada um dos marcos anatômicos ao centróide.
3. A última transformação é a remoção dos efeitos de orientação, nas quais todas as configurações são rotacionadas de maneira ótima, com a finalidade de minimizar a soma das distâncias quadráticas entre os marcos anatômicos homólogos (distância de Procrustes) de um par de espécimes. Uma

representação gráfica dos procedimentos descritos acima pode ser encontrada em Monteiro e Reis (1999).

Para Rohlf (2002), “A abordagem da Morfometria tradicional é baseada na aplicação de Análises Multivariadas nos padrões de coleções arbitrárias de medidas de distâncias, raios e ângulos [...]”. Estas variáveis representam tipicamente apenas parte da informação que pode ser obtida por meio das posições relativas dos marcos nos quais essas medidas foram tomadas. Esses métodos não levam em conta informações sobre os relacionamentos espaciais entre essas variáveis medidas.

Os métodos tradicionais também permitem apenas a visualização de relacionamentos estatísticos, tanto numericamente ou como gráficos de dispersão, não como estimativas da *Forma*. Em Rohlf (2002), “*Forma* é uma função das posições relativas de marcos morfológicos [...]”. A análise da *Forma* de contornos também é parte da Morfometria Geométrica mas não é citada no artigo.

Os marcos devem ser suficientes para capturar a *Forma* da estrutura de interesse, os pontos devem indicar a localização da mesma característica anatômica em diferentes espécimes.

As estruturas devem ser homólogas em algum sentido. Marcos são simplesmente pontos utilizados para rastrear as mudanças na *Forma* de alguma estrutura de interesse.

Em Rohlf (2002) não é assumido que as deformações parciais ou outras funções matemáticas das coordenadas desses pontos sejam também homólogas. Análises convencionais em Estatística Multivariada normalmente provêm gráficos de dispersão que permitem a visualização de padrões de variância e covariância no conjunto de dados. Métodos de Morfometria Geométrica permitem visualizações adicionais dos resultados das Análises Multivariadas.

Os pontos podem ser visualizados como *Formas* e os vetores podem ser visualizados como sequências nas mudanças das *Formas*. Isto é possível porque pontos no espaço multivariado podem ser mapeados em uma posição correspondente no espaço tangente e, a partir daí, de volta ao conjunto de coordenadas dos marcos do espaço físico de um organismo.

No artigo de Richtsmeier et al. (2002), a Morfometria é definida pelos autores como uma fusão entre a geometria e a biologia que lida com o estudo da *Forma* em espaços de duas e três dimensões. A área de pesquisa é conhecida pelo apoio às análises de *Formas* biológicas de modo a preservar a integridade física das *Formas* em duas ou três dimensões e evitar o colapso de *Formas* em uma série de medidas lineares ou angulares que não contenham informações pertencentes aos relacionamentos geométricos. Dados alguns marcos, os autores diferenciam os aspectos da *Forma* que podem ser conhecidos daqueles que não podem, e estimulam para que esses dados não sejam ignorados.

A aplicação de métodos quantitativos em dados biológicos é geralmente feita para dois propósitos diferentes que destinam a dois objetivos diferentes: geração de hipóteses e confirmação de hipóteses.

Na Morfometria, a primeira abordagem envolve a aplicação de métodos para descobrir informações nos dados, chamados de “métodos de exploração dos dados”. Nessa abordagem, métodos quantitativos são aplicados em um conjunto de dados e a análise procura por padrões nos dados que sugerem processos ou efeitos biológicos subjacentes, (RICHTSMEIER et al., 2002).

A segunda abordagem envolve testes de hipóteses mais formais. A captura da geometria por meio dos dados dos marcos tem se tornado trivial porque os marcos representam localizações exatas de *Formas* biológicas que contém significado de desenvolvimento funcional, estrutural ou evolutivo. Quando os mesmos marcos são coletados de um conjunto de objetos, esses marcos são chamados de marcos correspondentes, (RICHTSMEIER et al., 2002).

A base para esta correspondência, ainda de acordo com Richtsmeier et al. (2002), pode ser filogenética (para determinar as relações ancestrais entre espécies conhecidas), estrutural, de desenvolvimento ou biomecânica uma vez que objetos biológicos são compostos por diversos componentes estruturais cujas localizações podem ser precisamente definidas.

Porém, os marcos sozinhos não contém informações dos espaços, curvas ou superfícies entre eles. Se os dados relativos às regiões entre os marcos não forem parte dos dados coletados, então, não será possível obter informações verificáveis a respeito da *Forma* ou mudanças na *Forma* que ocorreram entre os marcos.

Os métodos de Morfometria para estudo das diferenças nas *Formas* de acordo com Richtsmeier et al. (2002) são descritos a seguir.

- Métodos de Superimposição (ou Sobreposição): envolvem a combinação dos dados dos marcos de duas maneiras diferentes no mesmo espaço de coordenadas de acordo com uma regra específica. Uma *Forma* é designada como sendo a *Forma* de referência e a outra como a *Forma* alvo. A mudança na *Forma* é determinada pelo deslocamento dos marcos da *Forma* alvo para os correspondentes marcos da *Forma* de referência. O pesquisador escolhe uma regra particular de superimposição. Todas as regras de superimposição envolvem os três passos descritos:
 - Ajustar uma das médias da *Forma* em uma orientação particular e denominá-la de objeto referência;
 - Translacionar e rotacionar as outras médias de *Formas* para que elas combinem com o objeto de referência de acordo com algum critério;
 - Estudar a magnitude e direção das diferenças entre as *Formas* de cada marco.
- Métodos de deformação: utilizam a área ou volume de uma *Forma* de referência e as deformam para corresponder com a *Forma* alvo. A principal idéia nas abordagens de deformações é:
 - Escolher uma das *Formas* como o objeto de referência;
 - Deformar o objeto alvo para que depois da deformação ele combine exatamente com o objeto referência;
 - Estudar a deformação a fim de aprender sobre as diferenças nas *Formas*.
- Métodos baseados em distância linear: comparam as distâncias lineares que conectam os pares de marcos em uma *Forma* com a distância linear correspondente em outra *Forma* e provê informação pertencente à diferença no comprimento dessas distâncias. A principal idéia dessa abordagem é:

- Para cada espécime, reescrever a coordenada da matriz dos marcos como uma matriz das distâncias lineares entre todos os pares únicos de marcos;
- Comparar cada elemento da matriz de *Formas* com o raio ou uma diferença absoluta ou, ainda, outra métrica;
- Estudar a matriz das comparações de distâncias lineares para determinar a diferença entre as *Formas* e as específicas distâncias lineares.

Todos os aspectos relacionados ao pré-processamento e análise da *Forma* podem ser descritos por meio desses métodos e a aplicação desses nesse trabalho segue o modelo algorítmico que descreve a aplicação do algoritmo kNN, a Análise Estatística da *Forma* e aplicação da Fuzzy kNN para classificação de abelhas, denominado FNN4Bees.

Dentre os métodos de Morfometria para estudo das diferenças nas *Formas*, de acordo com Richtsmeier et al. (2002), é utilizado, para o algoritmo implementado, uma variante dos métodos baseados em distância linear, implementado pelo FNN4Bees.

A seguir são apresentados os conceitos relacionados à Análise Estatística da *Forma*, alinhamento e padronização dos dados biológicos e uma breve introdução aos conceitos do algoritmo kNN para classificação.

2.3 ANÁLISE ESTATÍSTICA DA FORMA (AEF)

A Análise Estatística da *Forma* (AEF) está relacionada com características e comparações de *Formas* e objetos, de acordo com Silveira (2008). Algumas etapas devem ser realizadas com o intuito de representar a *Forma* de um objeto de uma maneira matematicamente conveniente. Uma aproximação conveniente é colocar marcos anatômicos no objeto, que, como foi descrito anteriormente, são pontos para identificar posições espaciais do objeto. “As coordenadas numéricas dos

marcos anatômicos são então usadas para representar um objeto”, (SILVEIRA, 2008). Um novo conjunto de coordenadas de um objeto, que será chamado de coordenadas de *pré-Forma*, pode ser obtido a partir das coordenadas de um objeto no espaço de marcos anatômicos. Em Cavalcanti et al. (1999) pode-se encontrar um exemplo do processo de seleção de características que acontece no procedimento para a colocação dos marcos anatômicos em que as coordenadas de todos os espécimes são alinhadas, isto é, transladadas, rotacionadas e escalonadas para combinar umas com as outras, usando o método de superimposição ortogonal generalizada dos mínimos quadrados de Procrustes. Este método fixa uma configuração sobre outra minimizando a soma das distâncias quadradas entre os marcos anatômicos homólogos.

Em Jirakanjanakit e Dujardin (2005) foi sugerido que as propriedades da métrica das asas provessem informações quantitativas precisas para a identificação das espécies, no qual foi aplicada uma análise da Morfometria Geométrica baseada em marcos anatômicos em mosquitos *Aedes aegypti*. De acordo com Jirakanjanakit e Dujardin (2005), “as asas dos insetos representam estruturas rígidas bidimensionais que são muito apropriadas para trabalhos morfométricos”.

Em Rodrigues e Santos (2004) procurou-se quantificar e discriminar, utilizando-se diversos métodos morfométricos, diferenças morfométricas entre vinte e duas pegadas de *Sauropodomorpha*, com evidências de marcas de quatro dedos. As técnicas utilizadas foram Sobreposição de marcos anatômicos e Análise de Deformações. Em Matias et al. (2001) os espécimes examinados constituíam-se de pedaços de fêmeas mortas em más condições de preservação morfológicas, portanto, era importante que a geometria da *Forma* analisada para discriminação fosse apropriada, isso foi alcançado por meio dos métodos de Morfometria Geométrica que utilizavam as coordenadas dos marcos anatômicos.

2.4 ALINHAMENTO E PADRONIZAÇÃO DOS DADOS

O alinhamento e a padronização dos dados envolvem procedimentos que compreendem a utilização de diversas ferramentas. Desses procedimentos, os que são realizados pelo especialista é, em primeiro lugar, a aquisição das amostras para análise e, em seguida, a colocação dos marcos anatômicos nessas amostras, realizada no software tpsDig2². A partir destes procedimentos, as imagens são enviadas no formato de arquivo do programa tpsDig (formato TPS) para a classificação e constituirão o conjunto de dados. A escolha do arquivo que foi utilizado para a inserção de dados de entrada no algoritmo FNN4Bees foi determinada pela análise dos dados obtidos pelos diversos métodos do aplicativo tpsRelw³, e a decisão do arquivo proveniente da utilização da técnica de espécies alinhadas foi tomada juntamente com o consentimento do especialista por melhor representar a *Forma* dos marcos anatômicos aplicados às imagens.

O primeiro passo para alinhar as espécies é realizado utilizando-se a ferramenta tpsUtil⁴. Nesse programa é possível converter diversos arquivos e outras funções utilitárias para auxiliar a manipulação dos arquivos TPS (que são os arquivos que possuem os marcos anatômicos das imagens das asas das abelhas). Ainda nesse programa os arquivos de entrada são os arquivos que contém as imagens das asas das abelhas. A entrada do sistema conterà o diretório no qual as imagens se encontram e em seguida é criado um arquivo TPS que contém a informação desse diretório para a localização dessas imagens e outras variáveis que são preenchidas pelo programa tpsDig2. A maneira como a ferramenta tpsUtil se apresenta para o desenvolvimento dessas operações é apresentada por sua interface principal na Figura 7.

2 Digitalizador de marcos anatômicos a partir de arquivos de imagens, imagens escaneadas ou vídeos. Realizado por F. James Rohlf.

3 O objetivo deste software é facilitar a Análise Estatística de marcos anatômicos na morfometria. Este software foi desenvolvido por F. James Rohlf.

4 O objetivo deste software é proporcionar a conversão de diversos arquivos e outras funções utilitárias para auxiliar a manipulação de arquivos TPS, desenvolvido por F. James Rohlf.

Figura 7 - Interface principal do programa
tpsUtil

O segundo passo no processo envolve a utilização do programa tpsDig2. O propósito desse programa é facilitar a Análise Estatística dos dados referentes aos marcos anatômicos na Morfometria transformando esse processo mais simples para a coleta e manutenção dos dados dos marcos anatômicos digitalizados. Dado um arquivo inicial, que especifica os nomes das imagens que contém as imagens dos espécimes, o programa pode ser utilizado para marcar a localização dos marcos anatômicos das imagens apresentadas. É imprescindível atender as necessidades especificadas para a colocação dos marcos anatômicos, para que eles representem a *Forma* da melhor maneira possível. Ele também permite a captura de estruturas de contorno e medidas lineares. O arquivo de saída desse programa é no formato TPS e pode ser utilizado em outros programas como o tpsRelw. O programa apresenta sua interface principal mostrada na Figura 8.

Figura 8 - Interface principal do programa
tpsDig2

O terceiro e último passo do processo envolve a utilização da ferramenta *tpsRelw*, que irá prover o definitivo alinhamento dos dados. O objetivo desse programa é possibilitar a Análise Estatística dos dados referentes aos marcos anatômicos na Morfometria. Ele é utilizado para explorar a diversidade da *Forma* dentro de sua variação global na amostra corrente. Ele fornece uma aproximação de baixa dimensão (por meio da Análise de Componentes Principais) para a aproximação do espaço tangente do espaço de *Forma*. Também apresenta uma estimativa da *Forma* de um organismo em locais diferentes neste espaço. O aspecto principal na utilização dessa ferramenta é o processo de dimensionamento para alinhar espécies. Os tamanhos das amostras podem ser dimensionados, após um alinhamento GLS (Generalized Least Squares), de duas maneiras. O diagrama da Figura 9, extraído de Slice (2001), ilustra as diferenças quando a variação da *Forma* é muito grande.

Figura 9 - Diagrama de uma secção transversal do Espaço da *Forma* de Kendall

Em Rohlf (1999) e Slice (2001), o diagrama de uma secção transversal do espaço da *Forma* de Kendall (círculo com um raio de $\frac{1}{2}$), aparece a partir do hemisfério de pré-*Formas* alinhadas com a referência (meio círculo com um raio de 1), e com o espaço tangente (linha tangente em cima). A distância de Procrustes é a distância r entre os ângulos em radianos. O ponto A, na Figura 9, representa a posição de uma *Forma* no espaço de Kendall e B é a posição correspondente no hemisfério após uma superposição GLS quando é escalado para uma unidade de tamanho do centróide. O ponto C é a projeção estereográfica do ponto A para o espaço tangente e D é a projeção ortogonal do ponto B para o espaço tangente. O tamanho do centróide é a raiz quadrada da soma dos quadrados das distâncias dos marcos anatômicos para o baricentro na ausência de alometria, que é a medida de tamanho único que está correlacionado com todas as variáveis de *Forma*. O arquivo resultante dessas transformações apresenta-se alinhado e será utilizado nas operações de classificação utilizando-se Lógicas Computacionais como a Fuzzy kNN. A interface principal do programa *tpsRelw* é mostrada na Figura 10.

Figura 10 - Interface principal do programa
tpsRelw

Como pode ser visto na Figura 10, os botões que se concentram no quadrante inferior da interface realizam as principais tarefas de Consensus, Deformações Relativas e Deformações Parciais. Após a aplicação desses métodos, o arquivo final deve ser salvo em relação ao processo que se quer destacar. O subitem para salvar destaca diversas opções: *Aligned Specimens*, *Bending energy matrix*, *Principal Warps matrix*, *centroid size*, *U matrix*, *Weight matrix*, *Just uniform Component*, *Relative warp scores matrix*, *R² matrix* e *Eigenvectors matrix*. A opção de utilizada neste trabalho é o *Aligned Specimens* (espécimes alinhados), pois esse arquivo irá conter os dados necessários para a classificação computacional com os pontos referentes aos marcos anatômicos já alinhados.

2.5 O ALGORITMO DOS K-VIZINHOS MAIS PRÓXIMOS (KNN)

O algoritmo dos k-Vizinhos mais Próximos (kNN) é utilizado normalmente na resolução de problemas de classificação de forma eficiente, mas que depende da dimensionalidade dos dados e o quanto eles são complexos. A aplicação desse algoritmo deve ser estudada para cada problema que envolva uma coleção de pontos que representam dados e um ponto a ser classificado em um espaço m-dimensional. Alguns problemas podem ser impossíveis de resolução com o

algoritmo kNN, incluindo aqueles cujo problema envolve a seguinte característica: um dado indefinido em um conjunto que contém os seus vizinhos mais próximos é instável para um dado α se a distância do ponto indefinido, em relação à maioria dos pontos no espaço, for menor que $(1 + \alpha)$ vezes a distância do ponto indefinido e seu vizinho mais próximo, (BEYER et al., 1999). Em problemas de classificação existem dois extremos de conhecimento que os especialistas podem ter. Ele pode ter um conhecimento estatístico completo da distribuição conjunta subjacente da observação de x e sua categoria verdadeira em θ ou ele pode não ter conhecimento da distribuição subjacente, exceto ao que se pode inferir de exemplos de amostras. No primeiro extremo, uma análise padrão Bayesiana irá prover um procedimento de decisão ótimo e a correspondente probabilidade de classificação do erro mínimo em R^* , (BEYER et al., 1999). No outro extremo, a decisão de classificar x em θ depende apenas do conjunto de amostras corretamente classificadas e o procedimento de classificação não fica claro. Para esses problemas, a decisão não paramétrica mais simples envolva o procedimento de regras na descrição do algoritmo kNN, que classifica x na categoria do vizinho mais próximo (COVER e HART, 1967). Alguns outros problemas podem tratar de dados simbólicos. Quando as características são numéricas, a distância normalizada Euclidiana pode ser utilizada para comparar os exemplos, como é realizado neste trabalho, porém, quando os valores de características são simbólicos, não ordenados (como a representação de conhecimento especialista em linguagem natural) os métodos do algoritmo kNN não são suficientes (COST e SALZBERG, 1993). Algumas propriedades do algoritmo kNN incluem os itens a seguir.

1. O algoritmo requer apenas um parâmetro, o número k de vizinhos mais próximos;
2. Não necessita de nenhum conhecimento prévio sobre a distribuição dos dados de treinamento;
3. Tem provado convergir na abordagem ótima Bayesiana sobre certas condições.

Em síntese, neste capítulo foi apresentado o processo de classificação de abelhas. O principal ponto abordado foi o conceito da Morfometria Geométrica e o estudo da *Forma*, que foram estudados para obter o conhecimento necessário sobre os dados que são provenientes dessa metodologia no processo de classificação proposto neste trabalho. Neste estudo, algumas técnicas foram aplicadas aos dados para alinhamento e padronização. A principal dela se refere ao cálculo do centróide, que foi aplicado a alguns dados do conjunto para exemplificação desse conceito.

O cálculo do centróide, como foi avaliado, serve para eliminar diferenças de dimensão das amostras. Também foram avaliadas algumas técnicas que são utilizadas pelos biólogos para classificação de abelhas. A principal técnica apresentada foi a Análise de Discriminantes Lineares (LDA) e o PCA.

Mais adiante, foram mostradas as interfaces dos programas utilizados pelos biólogos e, neste trabalho, para a padronização dos dados de forma automática. A partir da execução desses programas é gerado o arquivo final utilizado no processo de classificação do algoritmo kNN e da Fuzzy kNN. Por último, foram apresentados alguns conceitos do algoritmo kNN refinados em seu pseudocódigo.

Para a classificação baseada nesses procedimentos, são incluídas a Análise Estatística dos marcos anatômicos alinhados na determinação da Ordem de Significância para formar o padrão da espécie, a aplicação do algoritmo kNN estendido e, por fim, a aplicação da Fuzzy kNN sobre os dados resultantes do kNN. O cálculo da Ordem de Significância é feito por meio das funções de média, desvio padrão e variância. O seguinte item apresenta o conceito de Lógicas Computacionais e a Fuzzy kNN utilizada.

3 LÓGICAS COMPUTACIONAIS

Para o processo de classificação de abelhas diversas lógicas foram inicialmente propostas para a aplicação, entre elas as Redes Neurais Artificiais, as Máquinas de Suporte Vetorial, entre outros.

Porém, a partir de uma análise dos requisitos do problema, descobriu-se que para a apresentação dos resultados da classificação à um especialista em taxonomia de abelhas, o resultado apresentado por estas lógicas não seria trivial. A utilização do algoritmo kNN na análise entre as distâncias dos pontos e a Fuzzy kNN são algoritmos que, além de retornarem ótimos resultados no processo de classificação, como no caso do kNN, apresentam uma modelagem que pode ser apresentada de forma simples à um especialista que não seja da área de sistemas e computação. Por isso a escolha desses algoritmos.

3.1 REVISÃO DA LÓGICA FUZZY TRADICIONAL

A noção de conjuntos Fuzzy é altamente intuitiva e transparente já que ela captura o que realmente é percebido e descrito no mundo real. Pode-se encontrar diversas categorias de objetos cuja pertinência à uma dada categoria (classe ou conceito) é sempre uma questão de grau. Existem diversos exemplos nos quais pode-se encontrar elementos cujas alocações às categorias que se quer definir podem ser satisfeitos em algum grau. Pode-se dizer que a *continuidade* da transição de total pertinência e total exclusão é a principal e definitiva característica do mundo físico e de sistemas naturais, segundo Pedrycz e Gomide (2007).

De acordo com Wu et al. (2008) "[...] existem diversos modelos computacionais e abordagens que podem ser aplicados à problemas de bioinformática. Esta área de estudo pode ser agregada à ciência computacional favorecidamente [...]". Os autores acreditam que os grupos de disciplinas em expansão, relacionadas à inteligência computacional como, Redes Neurais Artificiais, Sistemas Fuzzy,

Computação Evolutiva, entre outros, ocupam um lugar central na taxonomia. Em uma análise da história da ciência, sempre existiu a necessidade de modelar e manusear a incerteza do mundo físico. E isto é particularmente verdade na biologia, em geral e, mais recentemente, na bioinformática. A estrutura mais utilizada para lidar com incerteza tem sido a teoria da probabilidade. A probabilidade é uma ferramenta muito boa que tem servido à ciência para modelagens de situações nas quais a fonte principal de incerteza é randômica, (WU et al., 2008).

Outras fontes de incerteza que podem ser consideradas incluem a incompletude dos dados extraídos das amostras, falta de expressividade ou exatidão de algumas características que são extraídas, falta de fronteiras mais definidas entre as diferentes espécies, como no caso de uma amostra pertencer a mais de uma espécie. Nessas situações, metodologias alternativas podem ser utilizadas para auxiliar as tomadas de decisões de avaliações automáticas.

A teoria dos conjuntos Fuzzy e a Lógica Fuzzy provê uma diferente maneira de se ver os problemas de modelagem de incertezas e oferece uma grande gama de ferramentas computacionais para auxiliar na tomada de decisão pela capacidade de fornecer informações complementares das informações que são provenientes da visão probabilística.

Segundo Wu et al. (2008), recentemente, houve um aumento expressivo dos aspectos Fuzzy de sistemas biológicos, que pode ser referente à mudança de paradigma para a "nova biologia". Pode-se notar ainda que muitos processos em sistemas biológicos estão intrinsecamente ligados à Teoria Fuzzy e não à processos determinísticos. Claramente, o conceito de assinar e determinar graus de pertinência ou rótulos de variáveis linguísticas da teoria dos conjuntos Fuzzy suporta a complexa análise e tomada de decisão de aplicações relacionadas à bioinformática.

Exemplificando, uma espécie de abelha, na sua classificação, pode pertencer em certos graus a uma determinada espécie. Suponha-se que, caso a maioria dos marcos anatômicos avaliados pertençam a uma determinada espécie, deve-se averiguar em que grau esta espécie analisada pertence à espécie correta e às demais espécies. Toda a teoria do conjunto Fuzzy é baseada no conceito de

funções de pertinência que, na maioria dos casos são definidas de duas maneiras, (WU et al., 2008):

1. Pela definição do senso comum que transporta algumas expressões linguísticas e, mais genericamente;
2. Pelo conhecimento do especialista diretamente, ou ainda, em minoria, podem ser derivadas de questionários, heurísticas, entre outros.

Uma vez que os subconjuntos de um universo X são definidos, as definições para o complemento do conjunto, para a união de dois conjuntos e para a intersecção de dois conjuntos têm de ser apresentadas para definitivamente gerar a "teoria do conjunto", (WU et al., 2008).

O complemento de A_c pertencente à A é definido por eq.(11).

$$A_c(x) = 1 - A(x) \quad (11)$$

Adicionalmente, seja $B : x$ pertence a $[0,1]$ como sendo outro subconjunto de X , Zadeh (1965) definiu a União e Intersecção entre os conjuntos como sendo, respectivamente apresentado em eq.(12) e eq.(13).

$$(A \cup B)(x) = \max\{A(x), B(x)\} = A(x) \cup B(x) \quad (12)$$

$$(A \cap B)(x) = \min\{A(x), B(x)\} = A(x) \cap B(x) \quad (13)$$

Suponha-se X como sendo o universo de discurso (espaço amostral) de todos os elementos. O universo pode ser contínuo ou discreto. Porém, o conceito de dicotomia (classificação que se baseia na divisão e subdivisão sucessiva em dois) não é aplicado quando se está definindo simples conceitos, (ZADEH, 1965). A Figura 11 mostra um exemplo de dicotomia.

Figura 11 - Apresentação de um conjunto e o conceito de dicotomia

Considerando o universo de discurso $X = [0, 3]$ e o conjunto de valores de S em X que são menores que o valor limite (threshold) t em X , $t = 2.0$. Considere os conjuntos apresentados em eq.(14).

$$S = \{x \in X \mid 0 \leq x \leq 2.0\} \wedge \{x \in X \mid 2.0 \leq x \leq 3.0\} \quad (14)$$

Cada conjunto é uma classe cujos membros são elementos do universo que satisfaz a mesma propriedade. Este conjunto é equivalente à lista de elementos do universo que são membros do conjunto. Dado um valor em X , o processo de dicotomia impõe uma decisão de classificação binária (está ou não está), aceitar ou rejeitar o valor como pertencente à uma dada classe.

Funções características do tipo $A : X \rightarrow \{0, 1\}$ induzem que há uma aplicação dos limites bem definidos quanto aos elementos do universo X , podendo ou não serem atribuídos ao conjunto A . A ideia fundamental dos conjuntos Fuzzy é relaxar esta requisição admitindo valores intermediários às pertinências das classes como mostra a Figura 12.

Formalmente, um conjunto Fuzzy A é descrito pela função de pertinência que mapeia os elementos de um universo X no intervalo $[0, 1]$, em Zadeh (1965).

Figura 12 - Exemplos de Funções de pertinência para duas espécies

A semântica é evidente já que o conjunto Fuzzy é expresso na base do conhecimento do espalhamento dos conceitos e seus valores típicos, (ZADEH, 1965).

O algoritmo kNN geralmente apresenta bons resultados na classificação de muitas aplicações, entretanto, ele ainda acarreta alguns problemas. Primeiramente, o desempenho do kNN depende altamente da escolha de k . Em segundo lugar, o processo de escolha dos vizinhos mais próximos não é considerado para aplicações em que as espécies se sobrepõem umas às outras, como no caso de pontos em imagens de sensoriamento remoto. Em terceiro lugar, o tratamento “por igual” dos vizinhos é conceitualmente simples.

Uma vez que a todos os vizinhos são dados pesos iguais para determinar a taxa de pertinência ao padrão de entrada, a informação provida pelas distâncias não é utilizada em sua completude no processo de classificação. Para lidar com esses problemas, algumas variações das técnicas que envolvem o kNN foram propostas, entre elas a Fuzzy kNN.

As informações disponíveis dos vizinhos apresentam características que podem rotular o padrão, tanto de maneira consistente ou de maneira conflitante. Consistente se o padrão definir de maneira ótima o elemento em análise e

conflitante se o padrão definido não for suficiente para a classificação, apresentando resultados errados.

O algoritmo kNN convencional (denominado CRISP) é um algoritmo simples que pode ser utilizado para assinar padrões de classificações desconhecidas da maioria das classes dos seus k vizinhos mais próximos de classificações conhecidas de acordo com suas distâncias, (RHEE; HWANG, 2003).

Entretanto, cada padrão de classificação conhecido é considerado igualmente importante na assinatura de um padrão a ser classificado. Isto pode causar dificuldades nas regiões nas quais os padrões se sobrepõem. Esse problema, como será apresentado mais adiante em maiores detalhes, é tratado na classificação de abelhas do algoritmo FNN4Bees pela designação da Ordem de Significância (OS) dos marcos anatômicos seguido da análise da Fuzzy kNN.

Considerando que o agrupamento de dados é uma ferramenta importante para a mineração de dados e análise de dados exploratória, é recomendável que se apliquem diversos algoritmos de agrupamento para um dado problema e após a análise dos resultados, determinar qual o melhor algoritmo para aquele conjunto de dados. Apesar desse método nem sempre ser possível, algumas alternativas podem ser adotadas para determinar o melhor algoritmo.

Em Weng et al. (2007) é proposto um novo algoritmo de agrupamento baseado em Fuzzy kNN para gerar uma matriz de similaridade de dados para sintetizar o conjunto e depois utilizar um algoritmo de agrupamento hierárquico para gerar a partição final sem saber o número específico de espécies antecipadamente.

Para eles, análises de agrupamentos devem ser de acordo com a similaridade entre os padrões dos dados, em que, somente aqueles cujas distâncias forem menores que um certo valor limite são similares.

Alguns exemplos foram avaliados da literatura referentes às aplicações de Lógica Fuzzy em diversos problemas, como no problema de reconhecimento de escrita em Gader et al. (1995). Neste exemplo, os autores descrevem experimentos que comparam Redes Neurais Artificiais com saídas Crisps e Lógica Fuzzy.

As saídas Fuzzy foram definidas usando um algoritmo Fuzzy kNN (FkNN). Em Zhu e Basir (2005), apresenta-se uma formulação do algoritmo Fuzzy kNN para classificação de imagens remotas. A formulação combina a teoria de Dempster-

Shafer (DSET) e o algoritmo kNN. Em Gader e Keller (1994) é feito uma pesquisa sobre a teoria dos conjuntos Fuzzy para reconhecimento de escrita com a descrição de três aplicações específicas, uma envolvendo um algoritmo Fuzzy kNN para reconhecimento de palavras, uma aplicando a Lógica Fuzzy para o reconhecimento de números de ruas escritos a mão e uma Fuzzy integral para o reconhecimento de caracteres de escrita.

Em Mitra e Pal (2005), os autores descrevem o paradigma da computação baseado em conjuntos Fuzzy para modelagem de incertezas e ambiguidades geralmente encontrados na vida real. Eles buscam gerar uma contribuição no reconhecimento de padrões, processamento de imagens e inteligência de máquinas baseada em Lógica Fuzzy dos últimos 40 anos. De acordo com os autores "[...] a teoria dos conjuntos Fuzzy provê uma maneira efetiva para descrever caracteres de um sistema muito complexo ou mal definido para permitir uma análise matemática precisa". Em geral, esses sistemas inserem conhecimento abstrato ou termos simbólicos.

A abordagem Fuzzy é baseada na premissa de que os elementos chave do pensamento humano não são apenas números, mas podem ser utilizados nas tabelas dos conjuntos Fuzzy como termos simbólicos. Em outras palavras, para as classes dos objetos, a transição de membros para não membros é gradual, não abrupta. Muito da lógica por trás do raciocínio humano não é da forma da Lógica tradicional de dois valores (verdadeiro ou falso), ou mesmo de diversos valores, mas da forma da lógica com verdades do tipo Fuzzy, com conexões Fuzzy e inferência das regras da Lógica Fuzzy.

Em Mallison e Bentley (1999), os autores descrevem o uso de um sistema de programação híbrido Fuzzy-genérico para descobrir padrões em grandes bases de dados. As regras Fuzzy oferecem um método para individualizar as classes de um conjunto de dados. Elas são fáceis de compreensão, verificação e extensão. Esse fato as tornam atrativas para serem utilizadas em domínios em que existem especialistas que podem alimentar o sistema com regras do conjunto de saída.

Em Bondugula et al. (2005), os autores fazem uma introdução de uma nova abordagem para predição de estruturas de proteínas secundárias usando o algoritmo Fuzzy kNN. Da mesma maneira, Gesù e Bosco (2007) apresentam uma

combinação paralela de classificadores Fuzzy kNN (FkNN). A combinação do classificador desenvolvido por Gesù e Bosco (2007) consiste em um novo procedimento otimizado baseado no algoritmo genérico aplicado ao FkNN que difere no tipo de similaridade usada.

Em geral, os trabalhos analisados que envolvem a aplicação do algoritmo kNN apresentaram bons resultados, sendo que, esses resultados estão diretamente ligados à dimensionalidade dos dados envolvidos nos testes.

Para trabalhos nos quais a dimensionalidade dos dados era muito alta ou as classes entre os elementos do conjunto de dados se interseccionavam foi mais adequado a utilização de outros algoritmos que puderam lidar melhor com esse problema ao invés do algoritmo dos kNN.

Um dos algoritmos propostos para este problema foi o Fuzzy kNN. Na maioria dos problemas expostos nos artigos apresentados, os quais houveram a aplicação da Lógica Fuzzy, tanto tradicional quanto Fuzzy kNN, as soluções encontradas foram ótimas.

4 A PROPOSTA FUZZY KNN (FNN4BEES)

O processo para classificação de abelhas deve envolver passos que determine, da maneira esperada pelo especialista, a correta classificação entre as espécies. O FNN4Bees é uma ferramenta de apoio a decisão no processo de classificação de espécimes em suas determinadas espécies por meio da análise da Morfometria Geométrica das asas desses espécimes. A análise realizada para classificar as espécies, no KNN, é feita por meio do cálculo das distâncias euclidianas entre os marcos anatômicos das diferentes espécies.

A metodologia envolvida no processo de classificação das abelhas considera os seguintes passos.

1. Analisar a *Forma* dos marcos anatômicos considerando as informações biológicas das espécies envolvidas e a aplicação de Métodos Estatísticos sobre os dados da Morfometria Geométrica.
2. A partir dessa análise, identificar e modelar as informações significativas no conjunto de dados e na Morfometria Geométrica por meio da Análise Estatística da *Forma*.

A análise da *Forma* dos marcos anatômicos por meio da Morfometria Geométrica se deve aos proeminentes resultados obtidos na aplicação dos métodos de Morfometria Geométrica em dados biológicos, de acordo com resultados apresentados em (TOFILSKI, 2008), (HEPBURN et al., 2005) e (FRANCOY et al., 2008b), entre outros. A análise das informações significativas do resultado da análise da *Forma* das imagens das asas deve apresentar dados suficientes para a identificação e modelagem dessas informações. No processo de classificação, devem ser determinados os padrões das espécies. Esses padrões são representados por pontos em um espaço bidimensional e que são específicos para cada espécie de um grupo de espécies, diferindo, na maioria das vezes entre os espécimes dos diferentes grupos. A aplicação do algoritmo kNN, em primeira instância, visa determinar quais desses pontos são significativos para cada grupo de

espécies, determinando o padrão da espécie. Conforme a revisão de trabalhos anteriores, a aplicação do kNN é indicada para problemas que envolvam uma coleção de pontos que representam dados e a adição de um novo dado a ser classificado em um espaço m-dimensional. O objetivo do algoritmo é encontrar o dado que é o vizinho mais próximo ao dado a ser classificado. As variantes que podem inutilizar a aplicação do algoritmo kNN (presença de um dado indefinível, não classificável, presença de conjuntos instáveis, etc) podem ser resolvidas com a aplicação da Fuzzy kNN.

Algumas questões envolvidas na construção de um classificador por meio do algoritmo dos k-Vizinhos mais Próximos incluem os itens a seguir.

- ✓ Quantos vizinhos devem ser considerados? Isto é, quanto deve ser o valor de k?
- ✓ Como as distâncias são medidas?
- ✓ Como combinar a informação de mais de uma observação?

Diversos fatores precisam ser considerados para a implementação de um algoritmo que envolve dados biológicos, incluindo o conhecimento do especialista. Alguns fatores são descritos a seguir.

- Quantidade de marcos anatômicos marcados nas imagens das asas.
- Conhecimento prévio sobre quais espécies estão contidas no conjunto de dados.
- Quantidade de amostras de cada espécie.
- Quantidade de espécies.
- Valor do k.

O pseudocódigo do algoritmo **FNN4Bees** que envolve a implementação do algoritmo kNN estendido e a Fuzzy kNN é descrito a seguir.

1. Determinar os padrões no conjunto de dados;
2. Armazenar as menores distâncias de cada marco anatômico;

3. Definir os padrões pelo valor de k;
4. Para todas as espécies faça
5. Para i de 1 até k
 - a. Calcular a distância euclidiana entre pontos da espécie indefinida e os marcos anatômicos significativos dos padrões do conjunto;
6. Armazenar as menores distâncias;
7. Calcular a Ordem de Significância;
8. Calcular a classificação com a Fuzzy kNN;

O cálculo da distância euclidiana é realizado conforme a eq.(15) para pontos bidimensionais, $P = (p_x, p_y)$ e $Q = (q_x, q_y)$.

$$\sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \quad (15)$$

4.1 DESCRIÇÃO DO FNN4BEES

O processo envolvido no FNN4Bees para a classificação das abelhas é apresentado na Figura 13 e será posteriormente melhor detalhado.

Figura 13 - Modelo do processo envolvido para classificação de abelhas

A Figura 13 apresenta o modelo do processo envolvido para classificação de abelhas. Em primeira instância os dados apresentam-se como pontos não alinhados em um espaço bidimensional. O primeiro passo no processo é realizar o alinhamento e a padronização desses dados através do cálculo do centróide. Com os pontos alinhados e padronizados aplica-se o algoritmo kNN estendido para a identificação dos marcos anatômicos mais significativos que definirão os Padrões para cada espécie no conjunto de dados. Tendo os padrões das espécies, calcula-se a variância e o desvio padrão desses padrões caracterizadores em relação aos dados que constituem a base do conjunto de dados. Serão considerados pertencentes à Ordem de Significância somente os marcos que se enquadrarem na regra apresentada na eq.(16).

Obtendo-se a Ordem de Significância dos marcos anatômicos aplica-se a Fuzzy kNN para cada grupo de espécies comparando as amostras de treinamento com as amostras da base de dados somente por meio dos marcos que se encontram na Ordem. O resultado da Fuzzy kNN é o resultado final da classificação. De acordo

com a análise do modelo apresentado na Figura 13 as seguintes metas podem ser listadas:

1. Realiza-se o alinhamento e a padronização dos dados.
2. Aplica-se o algoritmo kNN estendido para obter a lista dos marcos mais significativos identificando os padrões das espécies.
3. Calcula-se a variância e o desvio padrão dos pontos representados pelos marcos mais significativos.
4. Coloca-se na Ordem de Significância os marcos mais significativos que se enquadram na regra.
5. Calcula-se a Fuzzy kNN a partir da Ordem de Significância.
6. Obtém-se a classificação por meio do resultado da Fuzzy kNN.

O processo que descreve a classificação pelo algoritmo kNN e a Fuzzy kNN envolve a seleção e extração de características das imagens das asas das abelhas que estão contidas no arquivo de formato TPS. As próximas seções detalham, por meio da aplicação de exemplos, todo o processo envolvido na classificação.

4.1.1 Seleção de Características

O processo de seleção de características deve obter, a partir das imagens das asas, os padrões necessários para cada espécie no conjunto de dados das amostras. Esse processo envolve desde a aquisição da imagem, armazenada no conjunto de dados, até a completa formatação do arquivo necessário para classificação no algoritmo FNN4Bees. Dentre os passos que são envolvidos no processo de seleção de características, alguns devem ser realizados pelos especialistas, como a colocação, nas figuras, dos marcos anatômicos.

Nesse processo, as imagens apresentadas aos biólogos especialistas encontram-se como apresentada na Figura 14.

Figura 14 - Imagem de uma Asa da espécie *Scaura latitarsis*

A Figura 14 é um exemplo de imagem de uma asa da espécie de abelha sem ferrão denominada *Scaura latitarsis*.

As imagens das diversas espécies podem variar em contraste, tamanho, dimensão, posição, entre outros fatores. Quanto melhor a qualidade dessa imagem, melhor será a classificação dela, pois, ruídos nas imagens podem provocar a deformação no momento da colocação do marco anatômico pelo especialista e confundir o padrão dessa espécie no algoritmo FNN4Bees.

A partir dessas imagens os especialistas marcam os pontos relacionados aos marcos anatômicos.

A Figura 15 mostra como ficam os pontos (em vermelho) nessas imagens após as suas identificações.

Figura 15 - Imagem da Asa de uma amostra da espécie *Scaura latitarsis* com os marcos anatômicos colocados

Após a colocação dos marcos anatômicos nas imagens das asas, o especialista utiliza o programa TPSDig para obter o arquivo no formato da Figura 16.

```

1  LM=11
2  601.00000 453.00000
3  611.00000 469.00000
4  688.00000 515.00000
5  700.00000 536.00000
6  713.00000 388.00000
7  722.00000 406.00000
8  735.00000 439.00000
9  758.00000 468.00000
10 766.00000 526.00000
11 800.00000 544.00000
12 942.00000 524.00000
13 IMAGE=Friesella schrottkyi N 275-1g.bmp
14 ID=0
15 LM=11
16 526.00000 586.00000
17 547.00000 604.00000
18 767.00000 706.00000
19 772.00000 728.00000
20 741.00000 453.00000
21 770.00000 479.00000
22 809.00000 549.00000
23 874.00000 576.00000
24 891.00000 718.00000
25 971.00000 743.00000
26 1314.00000 657.00000
27 IMAGE=Friesomellita varia N 877-1g.bmp
28 ID=25
29 LM=11
30 574.00000 977.00000
31 581.00000 1014.00000
32 940.00000 1120.00000
33 960.00000 1154.00000
34 866.00000 806.00000
35 894.00000 855.00000
36 926.00000 931.00000
37 1021.00000 971.00000
38 1074.00000 1137.00000
39 1171.00000 1176.00000
40 1526.00000 1058.00000
41 IMAGE=Geotrigona sp (RP) col 2 10d.jpg

```

Figura 16 - Arquivo do conjunto de dados em formato TPS

A Figura 16 mostra o exemplo de conjunto de dados que contém diversos pontos referentes aos marcos anatômicos de imagens da asa de abelha apresentada pela Figura 15.

O arquivo apresentado na Figura 16 contém parte das espécies listadas com seus respectivos pontos, juntamente com um ID, que é um número de identificação e o nome da imagem seguido da extensão do arquivo que contém a imagem. A representação de cada imagem de asa segue o modelo apresentado no arquivo sendo que esse modelo deve conter, sequencialmente, a quantidade de marcos anatômicos definidos na imagem (representado no arquivo por "LM=11"), a próxima

linha mostra as coordenadas x e y do primeiro marco, e segunda linha mostra as coordenadas x e y do segundo marco e assim segue até o último marco. Por último, deve conter uma linha que mostre qual o nome do arquivo de imagem no qual essas informações foram extraídas (como em "IMAGE=...") e seu respectivo formato de arquivo (no caso definido por ".jpg").

O processo de alinhamento e padronização por meio de centróides do programa *TPSRelw* transforma os pontos referentes aos marcos anatômicos no documento apresentado na Figura 17.

```
"Aligned specimens
"(from tpsRelw, ver. 1.46)
" Alignment scaling method = unit centroid size
" Alignment projection method = orthogonal
1 360L 22 0
0 1 2 3 4 5 6 7 8 9 25 26 28 29 33
37 38 39 42 43 62 63 64 74 75 76 77 78 79 80
80 86 100 101 102 103 104 106 107 108 125 126 127 129 131
132 134 135 136 138 139 141 143 144 146 147 149 153 154 155
177 178 179 180 181 182 183 184 185 186 197 198 199 200 202
-3.84172533721854E-001 -7.69735961782315E-002 -3.54564726851
-2.54238045029278E-002 -2.16089834933563E-001 1.304747836744
6.26988593399725E-001 1.33717660019486E-001
```

Figura 17 - Arquivo dos pontos alinhados (align species) do software *TPSRelw*

A Figura 17 mostra os pontos do arquivo do conjunto de dados já alinhados e padronizados. As primeiras linhas deste arquivo são ignoradas pelo algoritmo FNN4Bees, as linhas consideradas no processo correspondem aos dados representados pelos números em notação científica (apresentados a partir da linha 11 no arquivo da Figura 17). O primeiro número em notação científica apresentado no arquivo representa a coordenada X da primeira amostra da espécie 0, o segundo número é a coordenada Y correspondente, o terceiro número é a segunda coordenada X e assim por diante, as quantidades de coordenadas X e Y são de acordo com a quantidade de marcos anatômicos que foram definidos e a quantidade total de pontos é referente a quantidade de amostras da espécie 0. O algoritmo FNN4Bees utiliza expressões regulares (do tipo "[0-9.]*") que foram definidas para encontrar essas coordenadas no arquivo e em seguida armazená-las

em um vetor de pontos flutuantes para serem utilizadas posteriormente no algoritmo.

As espécies que constituem o conjunto de dados, definido pela quantidade de amostras que estão presentes no arquivo da Figura 17, são alocadas sequencialmente no vetor.

Desta maneira, um vetor deve ser criado na implementação do algoritmo FNN4Bees para armazenar a ordem na qual as espécies estão alocadas. A priori, sabe-se como as espécies estão organizadas no conjunto e a inclusão de novas amostras deve seguir essa ordem, ou seja, serem inseridas no final do vetor. A ausência de uma espécie no conjunto pode resultar, como será dito mais adiante, em conflitos pelo algoritmo kNN. Porém, na implementação da Fuzzy kNN, esses conflitos são resolvidos pela verificação dos resultados de valores dispersos para o grau de pertinência da função da Fuzzy kNN. A inclusão de novas espécies no conjunto auxilia a completude do sistema e diminui o espalhamento do grau de pertinência, uma vez que novos espécimes poderão ser reclassificados em suas respectivas espécies.

Existe um formato padrão para a entrada de dados no algoritmo FNN4Bees, em que todos os elementos devem ser descritos como pontos flutuantes separados por uma vírgula, como exemplificado na lista da Figura 18.

0.2345586985864638,0.073839472621189,-1.7327329121327329, etc...

Figura 18 - Entrada de dados do algoritmo FNN4Bees

Alguns aspectos do algoritmo FNN4Bees devem ser antes elucidados. O Algoritmo FNN4Bees está dividido em três etapas.

1. A primeira etapa se refere à busca por padrões entre os marcos anatômicos de cada espécie do banco de dados. Esse processo é denominado de Busca de padrões entre as espécies do conjunto de dados. Neste processo, é realizado uma primeira execução do kNN entre os pontos de uma mesma espécie, analisando quais desses pontos são mais próximos e excluindo aqueles que estão mais distantes;

2. A segunda etapa é considerada uma etapa classificatória, que é realizada pelo kNN. Porém, nesta etapa, a classificação não é válida, pois não está em seu estado ótimo. Nesta etapa são armazenadas as distâncias entre os vizinhos mais próximos que são utilizadas na próxima etapa;
3. A terceira e última etapa consiste em obter uma Ordem de Significância entre os marcos anatômicos mais significativos (utilizando métodos estatísticos) e aplicar a esta Ordem, a Fuzzy kNN.

Para melhor compreensão de todas essas etapas, os aspectos que envolvem esses métodos são detalhados nos próximos itens.

4.1.2 Extração de Características

O conjunto de dados utilizado para a elaboração e avaliação deste trabalho apresenta um total de trinta e seis espécies sendo descritas, respectivamente, por:

Friesella_schrottkyi, *Frieseomellita_varia*, *Geotrigona_sp*, *Lestrimellita_limao*,
Melipona_bicolor, *Melipona_flavolineata*, *Melipona_mandaçaia*,
Melipona_quadrifasciata, *Melipona_scutellaris*, *Melipona_seminigra*,
Melipona_subnitida, *Nannotrigona_testaceicornis*, *Oxytrigona_tataira*,
Paratrigona_sp, *Paratrigona_subnuda*, *Partamona_helleri*, *Plebeia_São_Sebastião*,
Plebeia_droryana, *Plebeia_flavocincta*, *Plebeia_nigriceps*, *Plebeia_sp*,
Scaptotrigona_bipunctata, *Scaptotrigona_depilis*, *Scaptotrigona_postica*,
Scaptotrigona_sp, *Scaptotrigona_tubi*, *Scaura_latitarsis*,
Schwarziana_quadripunctata, *Tetragona_clavipes*, *Tetragonisca_angustula*,
Trigona_fuscipennis, *Trigona_hyalinata*, *Trigona_hypogea*, *Trigona_recurva*,
Trigona_sp e *Trigona_spinipes*.

Listagem 1 - Lista de espécies no conjunto de dados

Todas as amostras foram cedidas pelo Professor Doutor Tiago Maurício Franco, da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (USP - EACH), incluindo as amostras das *Euglossas*, que foram utilizadas no estudo de caso. A referência ao conjunto de dados pode ser encontrada em Franco (2009). As amostras já possuem marcos anatômicos que foram colocados pelo Professor Doutor Tiago M. Franco e sua equipe de pesquisa. Existem cerca de 30 a 100 amostras para cada espécie descritas em um arquivo com extensão TPS sequencialmente alocados. Para cada imagem foram colocados onze marcos anatômicos nas amostras das trinta e seis espécies descritas acima e dezoito marcos anatômicos nas amostras das *Euglossas*.

A busca pelos padrões nos marcos anatômicos das espécies contidas no conjunto de dados é descrita pelo pseudocódigo do algoritmo FNN4Bees para extração de características.

1. Separar as espécies no arquivo TPS;
2. Calcular centróides separadamente;
3. Para cada espécie, calcular k vizinhos mais próximos utilizando uma amostra que não pertence ao conjunto, mas que seja do mesmo grupo com as demais amostras do conjunto (amostra de teste rotulada);
4. Os marcos anatômicos significativos do conjunto de dados são os que possuem menores distâncias euclidianas da amostra confrontada;
5. Realizar a análise do gráfico de dispersão dos marcos anatômicos;

O resultado da execução do algoritmo FNN4Bees para extração de características, para cada espécie do conjunto de dados, é apresentado nas Tabelas 3 e 4, na qual os índices de 0 a 35 representam as espécies de abelhas apresentadas na listagem 1, respectivamente.

Tabela 3 - Apresentação dos marcos anatômicos mais significativos para as espécies de 0 até 18 a partir do FNN4Bees

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	4	4	1	1	1		1	1	1	1	1	1	2	1	4	6	1
5	2	7		2	2				3	2	2	2	2	7	2			2
	7			6	3						4		7					4
											9							

Tabela 4 - Apresentação dos marcos anatômicos mais significativos para as espécies de 19 até 35 a partir do FNN4Bees

19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	4	1	1	4	1	1	4	1	1	1	4	1	1	1	1	4
2			2	5		2		2	2	4	9	5	4	4	4	9
			4	6		4		3	4	9	10		5	7	10	
								4					6			

Nas Tabelas 3 e 4, cada célula apresenta os marcos anatômicos mais significativos em ordem decrescente, por exemplo, o marco com índice "0" é o menos significativo para as 35 espécies analisadas. Esses valores foram calculados para cada espécie do conjunto de dados em resposta à execução do algoritmo FNN4Bees na extração de características das espécies do conjunto de dados. O número k, presente no algoritmo kNN estendido representa a quantidade de elementos com menores distâncias entre os marcos anatômicos, ou seja, a quantidade máxima de marcos anatômicos que são considerados significativos. Uma análise antecipada desse número foi realizada por meio da observação dos dados plotados em gráficos de dispersão, na etapa seguinte. Quanto maior for o valor de k, mais dados não tão significativos serão incluídos no cálculo. O contrário também se aplica, pois, a ausência de um marco significativo para a análise de uma espécie pode implicar na incorreta classificação dessa espécie no algoritmo FNN4Bees.

No caso do FNN4Bees, a extração das características do conjunto foi baseada no estabelecimento de k igual a cinco. Portanto, cinco vizinhos mais próximos deverão ser considerados. Porém, uma análise do gráfico de dispersão (Figura 19) mostra que algumas distâncias acima de 5.00 unidades tornam o marco anatômico menos

significativo. Desta forma, um valor limite (*threshold*) foi estipulado para excluir os marcos anatômicos que podem ser considerados menos significativos pelo FNN4Bees (marcos anatômicos com valor acima de 5.00). Portanto, caso a distância ultrapasse o limite esse marco anatômico não deverá ser considerado.

Figura 19 - Marcos anatômicos mais significativos a partir dos gráficos de dispersão da espécie *E. ignita* (espécie 2)

A Figura 19 mostra o gráfico de dispersão dos marcos anatômicos para dez espécimes (amostras) da Espécie 2 (*E. ignita*). A análise da significância dos marcos anatômicos deve ser realizada armazenando-se os marcos anatômicos que, pela visualização no gráfico e pelo resultado do algoritmo kNN estendido, apresentam menores distâncias entre si (estão mais próximos uns dos outros). Cada uma das cores do gráfico representa um dos dezoito marcos anatômicos diferentes da Espécie 2. Em primeira análise, todos os marcos poderão ser considerados significativos, visualmente. O mesmo processo deve acontecer com todas as espécies do conjunto de dados, ou seja, para cada ponto $P(x_a, y_a)$ da amostra rotulada e $P(x_b, y_b)$ das espécies do conjunto de dados, deve-se analisar as distâncias euclidianas entre esses pontos. O resultado deste processo está representado nas Tabelas 5 e 6.

Tabela 5 - Definição dos marcos anatômicos mais significativos para as espécies de 0 até 18 a partir do gráfico de dispersão

LMs sig.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
LM sig. 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LM sig. 2	4	1	4	2	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1
LM sig. 3	5	2	5	4	2	2	6	2	2	2	2	2	2	2	4	2	2	3	2
LM sig. 4		3	7	6	4	3	8	5	7	3	3	4	6	4	5	3	3	4	3
LM sig. 5		4	8	7	5	6	9	6	8	5	6	5	7	6	6	4	4	6	4
LM sig. 6		5	9	9	6	10		9	10	6	8	9	8	7	7	6	7	7	7
LM sig. 7		6			10					7	10	10	9	9	8	7	8	8	9
LM sig. 8		7								10					9	8	9	9	
LM sig. 9		8													10	9			
LM sig. 10		9														10			
LM sig. 11		10																	

Tabela 6 - Definição dos marcos anatômicos mais significativos para as espécies de 0 até 18 a partir do gráfico de dispersão

LMs sig.	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
LM sig. 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LM sig. 2	1	2	1	1	2	1	1	2	1	1	1	1	1	1	1	2	1
LM sig. 3	2	3	2	2	4	2	2	4	2	2	3	2	3	2	2	3	6
LM sig. 4	3	4	6	4	5	4	4	7	3	3	4	3	5	3	4	4	10
LM sig. 5	4	6	7	6	6	6	5	8	4	4	5	4	6	4	6	8	
LM sig. 6	5	7	9	7	7	7	6	9	5	7	7	5	7	5	7	9	
LM sig. 7	6	8	10	9	8	9	8	10	6	8	8	6	9	6	8	10	
LM sig. 8	7	9			9	10	9		7	9	9	7		7	9		
LM sig. 9	8						10		8	10		8		8			
LM sig. 10	9								9			9		9			
LM sig. 11												10					

As Tabelas 5 e 6 mostram as disposições dos marcos anatômicos mais significativos para todas as espécies do conjunto de dados. Esses dados são analisados em conjunto com os dados apresentados pela execução da primeira etapa do algoritmo FNN4Bees. Para isso, a referência dos marcos anatômicos para cada espécie do conjunto de dados deve ser a mesma ou estar contida no conjunto de marcos anatômicos mais significativos adquiridos até o momento. Este procedimento permite estabelecer um padrão para a espécie a partir do conjunto de dados da espécie. O padrão será anotado para cada espécie como sendo o conjunto de marcos anatômicos significativos adquiridos. A comparação entre uma amostra indefinida e as amostras do conjunto de dados será feita por meio do padrão, somente nos marcos anatômicos mais significativos. Como exemplo, para a

espécie 19 foram encontrados como marcos anatômicos significativos os que correspondem aos números: 0, 1, 2, 3, 4, 5, 6, 7, 8 e 9. Dessa forma, esses marcos anatômicos mais significativos formam o padrão da espécie 19 e são utilizados para a comparação com a espécie indefinida.

4.1.3 O algoritmo kNN estendido

O algoritmo kNN é um exemplo de algoritmo de aprendizagem baseada em instância, no qual os dados de treinamento são armazenados, para que a classificação de um novo dado possa ser encontrada simplesmente comparando-o com os dados mais similares do conjunto de treinamento, (BERGAMINI; ARAÚJO; MOTTER, 2006).

A similaridade é calculada pela distância euclidiana entre os pontos do elemento indefinido e os pontos dos elementos das diversas espécies. A espécie que possuir mais pontos semelhantes é a espécie a qual o elemento indefinido pertence.

Neste trabalho a classificação de espécies de abelhas é baseada em seus marcos matemáticos adquiridos por meio da análise dos marcos anatômicos no processo de seleção e extração de características. Para uma análise de quinze imagens contendo cinco pontos, a Figura 20 apresenta um gráfico de dispersão das coordenadas (x, y) desses pontos.

Figura 20 - Pontos (x, y) de três espécies diferentes com cinco marcos anatômicos

A Figura 20 mostra três espécies diferentes (diferenciadas na figura pelas cores preta, vermelha e verde e os respectivos símbolos: círculo, quadrado e losango), cada uma contendo cinco marcos anatômicos.

Supondo a inserção de um novo espécime com o mesmo número de marcos anatômicos para serem analisados entre as espécies do conjunto, o objetivo do algoritmo kNN estendido é classificar qual espécie pode ser definida para aquele espécime que possui os marcos anatômicos mais similares. A similaridade, no algoritmo, será medida pelas distâncias euclidianas entre os marcos. Quanto menor a distância, mais similar esse marco é em relação à espécie indeterminada.

Figura 21 - Inserção de nova espécie para classificação

A Figura 21 mostra uma nova espécie, na cor azul (com o símbolo em triângulo), com cinco marcos anatômicos colocados e plotados juntamente com a análise da distribuição dos pontos das outras espécies.

A Figura 22 apresenta o relacionamento (circundados e identificados por números) entre o conjunto de marcos mais próximos ao da nova espécie e que será atestado por meio do algoritmo kNN estendido.

Figura 22 - Análise do gráfico para pesquisa dos marcos anatômicos mais significativos

Para $k = 1$ no algoritmo FNN4Bees os novos dados $C8 * C7$ (triângulo azul) poderão ser classificados de acordo com uma observação dos dados mais próximos a eles como sendo a espécie que possui um vizinho mais próximo ao dado indefinido (o vizinho com menor distância). Neste caso, os novos dados poderão ser classificados como pertencentes à espécie verde (losango), uma vez que esta é a classificação dos pontos que estão mais perto dos novos pontos. Este procedimento funciona bem para as áreas circuladas de número 2 e 5, porém, para os demais marcos anatômicos (números 1, 3 e 4), devido à proximidade equivalente entre as espécies, não se pode definir a espécie a qual a espécie indefinida pertence. Para solucionar esse problema, utiliza-se o resultado da primeira análise (análise dos marcos anatômicos mais significativos).

Desta forma, o algoritmo FNN4Bees calcula qual dentre as espécies do conjunto é a que mais se assemelha a espécie indefinida. Para isso, segue-se, depois da análise da espécie indefinida com os padrões, à análise das menores distâncias dos k Vizinhos mais Próximos. Para cada espécie, tira-se a média das distâncias dos marcos anatômicos significativos e armazena-as para posterior análise. Esta análise posterior deve decidir qual espécie é a espécie indefinida a partir apenas do algoritmo kNN estendido.

O importante desse processo é armazenar as distâncias obtidas entre as espécies (vizinhas mais próximas) que são utilizadas na próxima etapa para análise da Fuzzy kNN. Essas distâncias devem indicar, para o elemento indeterminado, quais são as espécies que mais se assemelham a ele. A próxima análise é a respeito dessa semelhança e como implementar o conhecimento do especialista nessa análise. Quando no algoritmo kNN estendido esta semelhança é medida apenas pela distância euclidiana (considerando sempre a menor), agora, esta semelhança poderá adotar valores simbólicos, que representam mais do que a distância, representam o conhecimento do especialista. Como formular esse conhecimento e definir as semelhanças são etapas que serão analisadas a seguir.

4.1.4 Análises Estatísticas para Identificação da Ordem de Significância

A Ordem de Significância foi idealizada para que os marcos anatômicos comparados entre a espécie indefinida e a espécie do conjunto de dados sejam aqueles que melhor classificam os elementos indefinidos em relação às espécies existentes. Os marcos anatômicos mais significativos são representados como marcos matemáticos, os quais no processo de definição da Ordem de Significância dos marcos anatômicos, devem ser analisados para melhorar a classificação matemática das espécies, embora esse marcos ainda possam ser biologicamente significativos.

A discriminação da Ordem de Significância por meio do cálculo da variância aplicada sobre os dados provenientes da execução do FNN4Bees resulta nos dados apresentados na Tabela 7.

Tabela 7 - Cálculo da Variância e do desvio padrão para onze marcos anatômicos da espécie *Friesella schrottkyi*

VARIABLE	METHOD	CI FOR STDEV	CI FOR VARIANCE
LM1	STANDARD	(0,0792; 0,1283)	(0,00627; 0,01645)
LM2	STANDARD	(1,38; 2,23)	(1,90; 4,99)
LM3	STANDARD	(0,105; 0,170)	(0,0110; 0,0290)
LM4	STANDARD	(1,01; 1,63)	(1,02; 2,67)
LM5	STANDARD	(2,09; 3,39)	(4,39; 11,51)
LM6	STANDARD	(4,38; 7,10)	(19,2; 50,4)
LM7	STANDARD	(1,47; 2,38)	(2,16; 5,68)
LM8	STANDARD	(0,116; 0,188)	(0,0135; 0,0355)
LM9	STANDARD	(1,65; 2,67)	(2,72; 7,15)
LM10	STANDARD	(0,164; 0,266)	(0,0270; 0,0709)
LM11	STANDARD	(1,44; 2,34)	(2,08; 5,46)

A Tabela 7 mostra o cálculo da variância e do desvio padrão em cada um dos onze marcos anatômicos de dez amostras padronizadas contidas no conjunto de dados referente à espécie 0, denominada *Friesella schrottkyi*. A coluna *Variable* apresenta os marcos anatômicos em sequência de 1 a 11, definidos por (LM1, LM2,..., LM11), *Method* define o método de cálculo adotado para a variância, *CI for StDev* indica o resultado do cálculo do desvio padrão das coordenadas (x; y) dos marcos anatômicos e *CI for Variance* indica o resultado do cálculo da variância das coordenadas (x; y).

O método da regra estatística adotado para definição da Ordem de Significância (OS) dos marcos anatômicos é apresentado na eq.(16).

$$if \left((STDev < 1) \vee (Variance < 1) \right) then LM \in OS \quad (16)$$

Ou seja, se o valor do desvio padrão para x ou para y ou o valor da variância para x ou para y forem menor que um, então, o marco anatômico será considerado parte do vetor que armazena a Ordem de Significância. Para a espécie 0 (*Friesella schrottkyi*), o vetor possui os seguintes marcos anatômicos mais significativos armazenados: **LM1, LM3, LM8 e LM10**.

Os marcos anatômicos mais significativos na etapa de extração de características para a espécie 0 são apresentados na eq.(17).

$$LMSig(padrao) = 1, 2, 3, 5, 8, 9, 10 \quad (17)$$

Pelo cálculo da variância e desvio padrão temos como marcos anatômicos mais significativos os apresentados em eq.(18).

$$LMSig(padrao) = 1, 3, 8, 10 \quad (18)$$

Como *LMSig* (variância ou padrão, veja a eq.(18)) está contido em *LMSig* (stDev ou padrão, veja a relação da eq.(17)), então os valores retornados pelo cálculo da variância na definição da Ordem de Significância estão corretos, ou seja, esta é uma medida de avaliação que pode ser utilizada para a entrada na Fuzzy kNN, pois ela se fundamenta nos estudos realizados entre as similaridades dos marcos anatômicos por meio dos gráficos de dispersão e da execução do algoritmo FNN4Bees.

Após esse cálculo, a busca pelos marcos anatômicos que melhor definirão a espécie é continuada para eliminar os marcos anatômicos que, em comparação com outras espécies possam causar ruídos no alcance da melhor solução. Isto pode ocorrer pois, em determinadas espécies, alguns marcos anatômicos podem ser mais significativos que os marcos anatômicos que definem a espécie correta. Como solução, é feito um ajuste entre os marcos anatômicos mais significativos das espécies de maneira a se eliminar nas demais espécies (espécies que não são a espécie correta), os marcos anatômicos que estão definindo-as incorretamente. O aspecto mais interessante neste processo se refere a análise exaustiva pela busca da Ordem de Significância que deve seguir até que se alcance o resultado ótimo na busca pelos marcos anatômicos que melhor definem as espécies no processo de alinhamento. Esse processo da fase de alinhamento de marcos anatômicos é descrito por meio do pseudocódigo 1 apresentado na listagem 2.

```
//sendo N = quantidade de espécies
para i de 1 até N
    /* considere todos os marcos anatômicos
    significativos na Ordem */
    para i de 1 ate LMSig.lenght();
        ordem[] = LMSig[];
    se (espécieIndefinida != espécieVerdade)
        então calculaOrdemDeSignificancia();
```

Listagem 2 - Pseudocódigo do método buscaPadraoLandmarks

O método calculaOrdemDeSignificancia() é descrito conforme listagem 3.

```
/* sendo L a quantidade de marcos anatômicos mais
significativos */
para i de 1 até L
    /*retira da Ordem os marcos anatômicos menos
    significativos, como eles estão ordenados em ordem
    decrescente, então remove-se o primeiro elemento*/
    remover(LMSig[0]);
```

Listagem 3 - pseudocódigo do método calculaOrdemSignificancia()

Como resultado desse processo, fazem parte da Ordem de Significância somente os marcos anatômicos que se adequam a regra (que possuam cálculo do desvio padrão e/ou variância <1). A próxima análise deve ser feita caso ($\text{espécieIndefinida} \neq \text{espécieVerdade}$), ou seja, caso o resultado continue não sendo o esperado ou ainda não chegou-se ao resultado ótimo.

Para novas espécies indefinidas, em novas análises, a Ordem de Significância deverá ser refeita. O processo é executado somente na fase de treinamento para determinar quais são os marcos anatômicos que melhor classificam aquela espécie em relação às outras.

Para exemplificar a influência do processo descrito na determinação da Ordem de Significância dos marcos anatômicos, suponha que uma nova espécie (rotulada)

esteja sendo analisada entre as espécies do conjunto de dados. Sabendo-se que todas as espécies já possuem sua Ordem de Significância calculada, a próxima análise é feita para ajustar esta Ordem, caso a classificação dessa fase de ajuste nos marcos anatômicos não seja correta para alguma das espécies.

Considerando que, para a espécie 6, os marcos anatômicos mais significativos são: 0, 1, 6, 8 e 9, deve-se testar esses marcos anatômicos, por meio de treinamento, ou seja, qual a representabilidade deles como caracterizadores dessa espécie no grupo.

Para isso, executa-se o algoritmo FNN4Bees uma vez.

Caso a classificação dessa espécie de teste seja a mesma da espécie do grupo, ou seja, caso a espécie de treinamento seja a espécie 6 então esses marcos anatômicos estão caracterizando bem esta espécie e não precisam ser otimizados.

Caso a classificação não seja correta, ou seja, caso a resposta do algoritmo FNN4Bees não seja a espécie 6, então, deve-se alterar os marcos anatômicos mais significativos que estão caracterizando esta espécie, aumentando o valor limite (threshold) e excluindo um, e apenas um marco anatômico mais significativo de cada vez, o marco anatômico da Ordem, que, dentre os marcos anatômicos dessa Ordem, será o menos significativo.

Esse processo é repetido até que a classificação forneça um resultado correto da espécie no grupo, ou alcançou-se o número mínimo de marcos anatômicos que podem ser retirados da Ordem.

Neste último caso, caso a classificação não tenha sido a correta, mesmo alcançando-se o número mínimo de marcos anatômicos caracterizadores da espécie, então, a classificação dessa espécie fica incorreta pelo algoritmo kNN estendido.

Ainda na execução do algoritmo FNN4Bees para a espécie 6, têm-se as distâncias euclidianas dos marcos anatômicos mais significativos, contidos na Ordem de Significância de cada espécie, entre essas espécies e a espécie testada (espécie 6) conforme as Tabelas 8 a 12. Ao final do cálculo é feita uma média dos valores que corresponde à distância final utilizada na comparação entre as espécies pela Fuzzy kNN.

Tabela 8 - Resultado das médias das distâncias obtidas entre a espécie 6 com a espécie 0 até a espécie 7, em mm, pelo algoritmo FNN4Bees

6_0	6_1	6_2	6_3	6_4	6_5	6_6	6_7
5.01006	3.32980	1.68596	4.14120	6.25073	4.19361	1.19506	3.47811

Tabela 9 - Resultado das médias das distâncias obtidas entre a espécie 6 com a espécie 8 até a espécie 15, em mm, pelo algoritmo FNN4Bees

6_8	6_9	6_10	6_11	6_12	6_13	6_14	6_15
5.05848	7.09049	3.79035	3.87659	2.31822	3.06558	6.37632	3.93599

Tabela 10 - Resultado das médias das distâncias obtidas entre a espécie 6 com a espécie 16 até a espécie 23, em mm, pelo algoritmo FNN4Bees

6_16	6_17	6_18	6_19	6_20	6_21	6_22	6_23
3.20042	3.82335	1.47118	4.00397	4.64154	4.38568	3.17495	3.84189

Tabela 11 - Resultado das médias das distâncias obtidas entre a espécie 6 com a espécie 24 até a espécie 31, em mm, pelo algoritmo FNN4Bees

6_24	6_25	6_26	6_27	6_28	6_29	6_30	6_31
5.28588	3.85655	4.37544	2.42997	2.82239	1.70356	2.55247	3.08058

Tabela 12 - Resultado das médias das distâncias obtidas entre a espécie 6 com a espécie 32 até a espécie 35, em mm, pelo algoritmo FNN4Bees

6_32	6_33	6_34	6_35
4.38914	3.41152	2.35003	3.15799

A menor distância média dentre todas as apresentadas nas Tabelas é de: 1.19506 mm a qual corresponde à distância média da espécie 6 com a espécie do conjunto de dados 6. Portanto, a classificação já está correta e não é necessário alterar a Ordem de Significância dos marcos anatômicos mais significativos dessa espécie, esses marcos formarão, assim, o padrão da espécie 6. Como foi dito, caso a classificação não fosse a correta, e o critério de parada não tivesse ainda sido alcançado, o processo para a otimização dos marcos anatômicos continuaria.

O próximo passo no algoritmo é realizar a classificação na Fuzzy kNN utilizando as distâncias euclidianas obtidas pelo algoritmo kNN estendido. Esse processo ocorre

independentemente da classificação anterior ter sido correta ou não, contanto que tenha-se chegado aos padrões de todas as espécies.

A Fuzzy kNN é aplicada para o reconhecimento da espécie que mais se assemelha a espécie estudada. A análise é feita por meio das médias das distâncias dos marcos anatômicos mais significativos entre as espécies na implementação de uma Fuzzy kNN (FkNN). Os resultados obtidos pelo algoritmo FNN4Bees, na implementação do kNN, são utilizados para determinar as distâncias e o cálculo da média dessas formam os dados de entrada da Fuzzy kNN. Este processo pode ser descrito como ilustra a Figura 23.

Figura 23 - Modelo da implementação da Fuzzy kNN nos marcos anatômicos mais significativos

A classificação pelo algoritmo kNN estendido é feita por meio do cálculo da média entre as distâncias dos marcos anatômicos pertencentes à Ordem de Significância

de cada espécie, conforme apresentado anteriormente. Na Fuzzy kNN é realizada a comparação entre as médias das distâncias dos marcos anatômicos e, no processo de defuzzyficação, a espécie que possuir maior valor de pertinência entre os espécimes é a espécie escolhida. Lembrando-se que os marcos anatômicos presentes nesse processo são apenas aqueles descritos na Ordem de Significância. O projeto de um sistema Fuzzy tradicional consiste nos seguintes itens.

- Definir os universos de discurso das variáveis linguísticas do sistema, o erro, a variação do erro e a variação da saída do sistema (discretização ou níveis de quantização);
- Definir o número de termos primários e graus de pertinência dos conjuntos difusos que representam cada termo;
- Determinar as regras que formam o algoritmo de controle;

Para maior compreensão do universo de discurso dos elementos, é necessária uma análise de como eles se comportam na espécie.

4.1.5 Aplicando a Fuzzy kNN (FkNN)

A Fuzzy kNN (FkNN) foi criada originalmente para solucionar problemas com o algoritmo kNN e é o algoritmo kNN com a utilização de conjuntos Fuzzy como saída do algoritmo, ou seja, o algoritmo recebe como entrada de dados as médias das distâncias entre as espécies e retorna os graus de pertinência para cada espécie do conjunto de dados provenientes da função de pertinência implementada no FkNN. O processo envolvido na Fuzzy kNN no algoritmo FNN4Bees envolve as etapas descritas a seguir.

- Encontre as distâncias dos vizinhos mais próximos, assinalando esses vizinhos (definindo o valor de k);

- A seguir, para classificar um novo elemento, será dado um valor de pertinência Fuzzy para todas as espécies desses vizinhos.
- A pertinência Fuzzy é definida a partir de detalhes sobre o número de vizinhos em uma espécie e suas distâncias.
- Um novo elemento (nova amostra) irá pertencer à espécie que possuir maior pertinência a este elemento.

A base do algoritmo FkNN é assinalar graus de pertinência como funções das distâncias dos elementos (amostras) a partir dos seus vizinhos mais próximos e a pertinência desses em uma possível espécie.

Os passos descritos pelo algoritmo FkNN no processo de reconhecimento de padrões pelo kNN são descritos a seguir.

Considere $W = \{w_1, w_2, \dots, w_t\}$ como um conjunto de t dados rotulados.

- Cada objeto w_i é definido por l características:
 - $w_i = (w_{i1}, w_{i2}, \dots, w_{il})$
- A entrada é formada por y elementos não classificados (elementos indefinidos);
- Sendo k o número de vizinhos mais próximos de y ;
- Sendo E o conjunto de k vizinhos mais próximos (NN);
- $\mu_i(y)$ é o grau de pertinência de y na espécie i ;
- μ_{ij} é o grau de pertinência da i -ésima espécie do i -ésimo vetor do conjunto rotulado;

Sendo t o número de elementos que identificam a espécie (no estudo de caso esse valor é considerado o número de marcos anatômicos mais significativos). Sendo c o número de espécies do universo. Sendo W o conjunto que contém os t elementos. Cada grupo é representado por um subconjunto de elementos de W .

```

Defina k
{calculando os vizinhos mais próximos (NN)}
para i = 1 até t
    Calcule a distância de y até  $x_i$ 
    se ( $i \leq k$ )
        então adicione  $x_i$  a E
    senão se  $x_i$  for mais perto de y que qualquer outro NN
        então apague o vizinho mais distante e inclua  $x_i$  no
        conjunto E

```

Quando o algoritmo kNN classifica x_i em uma das espécies do universo de discurso, o rótulo l é determinado apenas pela frequência da representação. Ele não depende da distribuição dos k elementos do kNN. Este algoritmo é baseado, como foi dito, no algoritmo kNN e nas relações Fuzzy. Os detalhes do algoritmo FkNN são apresentados a seguir.

A equação para o cálculo de pertinência de x para a i -ésima espécie, $\mu_i(x)$, é dada em Keller, Gray e Givens (1985) como apresentado na eq.(19).

$$\mu_i(X) = \frac{\sum_{j=1}^k \mu_{ij} \left(\frac{1}{d(X, X_j)^{2/(m-1)}} \right)}{\sum_{j=1}^k \left(\frac{1}{d(X, X_j)^{2/(m-1)}} \right)} \quad (19)$$

O valor de $\mu_i(X)$ pertence à uma espécie de diferentes maneiras. Pode ser que à este valor seja dado um grau de pertinência completo (máximo) na sua própria espécie e nenhuma pertinência nas outras espécies. Pode ser baseada na distância da média entre as espécies ou nas distâncias das amostras rotuladas nas suas próprias espécies e entre as outras espécies. O espaçamento apresentado na função triangular é definido, para as variáveis a e b , que correspondem ao mínimo e máximo da função, respectivamente, em eq.(20) e eq.(21).

$$\text{Sendo } i = 1, \quad a_i = \text{distLM}_i - A * \text{distLM}_{(i-1)} \quad (20)$$

$$\text{Sendo } i = 0, \quad b_i = \text{distLM}_i + B * \text{distLM}_{(i+1)} \quad (21)$$

As variáveis A e B são pesos atribuídos à função para formatação do resultado na fase de treinamento. Esses pesos são ajustados dinamicamente, para a otimização dos valores mínimos e máximos apenas durante a fase de treinamento. Ainda, esses valores devem ser muito bem especificados para que a saída da Fuzzy kNN não seja influenciada pelos valores dos marcos anatômicos vizinhos e permitem ajustar o grau com que os marcos anatômicos das outras espécies influenciarão na espécie com menor distância.

De acordo com Souza (2003), se u e v são conjuntos Fuzzy, a função de mapeamento desses conjuntos é dada por meio de uma função chamada T-Norma, com as seguintes propriedades:

1. $T(1,a) = a$, para todo $a \in [0,1]$;
2. $T(a,b) = T(b,a)$, para todos $a,b \in [0,1]$;
3. Se $a \leq c$ e $b \leq d$, então $T(a,b) \leq T(c,d)$;
4. $T(a,T(b,c)) = T(T(a,b),c)$, para todos $a,b,c \in [0,1]$.

Com esta definição, o conceito de conjunção é generalizado.

Se considerarmos $T(a,b) = \max \{a,b\}$, esta função será uma T-Norma.

A função Fuzzy apresenta-se, na forma generalizada, para o problema de classificação de abelhas, como em eq.(22).

$$f(A(x)) = \text{função}(U_0/a_0, U_1/a_1, U_2/a_2, \dots, U_n/a_n) \quad (22)$$

Onde n é igual ao número de espécies do conjunto de dados.

A Figura 24 apresenta três espécies distintas, sendo a espécie 0, representada pelo círculo na cor preta, a espécie 1, representada pelo quadrado na cor vermelha e a

espécie 2, representada pelo losango na cor verde. Cada espécie contém amostras de teste. A amostra indefinida (representada pelo triângulo azul) deve ser agrupada em uma dessas espécies.

Figura 24 - Disposição no gráfico dos marcos anatômicos das espécies 0, 1, 2 e a amostra da espécie indefinida

Os valores das coordenadas x e y para as três espécies e a amostra da espécie indefinida são apresentados na Tabela 13, na qual N_x e N_y correspondem às coordenadas x e y da espécie indefinida, 0_x e 0_y correspondem às coordenadas x e y da espécie 0, 1_x e 1_y correspondem às coordenadas x e y da espécie 1 e 2_x e 2_y correspondem às coordenadas x e y da espécie 2, respectivamente.

Tabela 13 - Coordenadas (x, y), em mm, dos marcos anatômicos das Espécies 0, 1, 2 e da espécie indefinida

	N_x	N_y	0_x	0_y	1_x	1_y	2_x	2_y
LM0	-3,823	-7,646	-3,825	-6,622	-3,915	-4,324	-4,222	-2,575
LM1	-3,527	-2,924	-3,459	-1,841	-3,632	-1,898	-4,165	1,696
LM2	-1,255	1,065	-1,209	1,053	-6,676	1,184	-1,747	1,192
LM3	-9,014	1,684	-9,000	1,728	-6,002	1,481	5,008	1,496
LM4	-5,178	-2,682	-7,313	-2,659	-1,018	-2,224	-9,842	-2,270
LM5	-2,521	-2,151	-2,250	-2,237	-6,272	-1,874	-7,369	-1,764

LM6	1,314	-1,177	2,561	-0,999	-1,016	-0,931	-3,771	-8,759
LM7	8,102	-3,219	7,031	-3,247	7,742	-5,672	7,696	-5,723
LM8	1,046	1,389	1,096	1,306	1,003	1,346	1,275	1,226
LM9	2,049	1,920	2,250	1,841	2,081	1,683	2,388	1,563
LM10	6,240	1,330	6,300	1,137	6,703	5,243	6,177	9,097

As distâncias dos marcos anatômicos entre a espécie indefinida e as outras três espécies são dadas pela execução do algoritmo FNN4Bees, considerando apenas os padrões, para cada uma das três espécies. O resultado é apresentado na Tabela 14.

Tabela 14 - Distâncias, em mm, entre a espécie indefinida e as espécies 0, 1, 2 dos marcos anatômicos mais significativos dadas pelo FNN4Bees

LM Sig	Dist N_0	LM Sig	Dist N_1	LM Sig	Dist N_2
LM2	0,0475375	LM0	3,32278	LM5	2,53684
LM3	0,0458367	LM1	1,03077	LM7	0,281437
LM5	0,284568	LM2	5,42231	LM8	4,86266
LM8	0,0972916	LM3	3,01850		
LM9	0,215789	LM4	4,18512		
LM10	0,201994	LM5	3,76060		
		LM6	8,46039		
		LM7	2,47887		
		LM8	0,0610274		
		LM9	0,239757		
		LM10	3,93993		

Aplicando a eq.(19) para $m = 2$, e considerando apenas três marcos anatômicos mais significativos, nas espécie 0, 1 e 2 são geradas, respectivamente, as equações eq.(23), eq.(24) e eq.(25).

$$\begin{aligned}
 U_0 &= ((1/0,0475)+(1/0,0458)+(1/0,284)) / \\
 &((1/0,0475)+(1/0,0458)+(1/0,284)+(1/3,322)+(1/1,030)+(1/5,422)+(1/2,53684)+(1/0,281437)+(1/4,862 \\
 &66)) = 0.8922
 \end{aligned}
 \tag{23}$$

$$U1 = ((1/3,322) + (1/1,030) + (1/5,422)) / ((1/0,0475) + (1/0,0458) + (1/0,284) + (1/3,322) + (1/1,030) + (1/5,422) + (1/2,53684) + (1/0,281437) + (1/4,86266)) = 0.0280 \quad (24)$$

$$U2 = ((1/2,53684) + (1/0,281437) + (1/4,86266)) / ((1/0,0475) + (1/0,0458) + (1/0,284) + (1/3,322) + (1/1,030) + (1/5,422) + (1/2,53684) + (1/0,281437) + (1/4,86266)) = 0.0798 \quad (25)$$

Como pode-se notar nas eq.(23), eq.(24) e eq.(25), os valores de pertinência para N (espécie indefinida) classifica-a, com maior grau de pertinência, na espécie 0, com 0.8922 de pertinência, enquanto que a espécie indefinida se coloca na espécie 1 em apenas 0.0280 e na espécie 2 em apenas 0.0798.

4.2 ESTUDO DE CASO – *EUGLOSSAS*

A seguir é descrito um exemplo completo de estudo de caso de como é realizada a classificação de um espécime da *Euglossa* dentre cinco espécies desse grupo. O arquivo TPS das *Euglossas* é apresentado como mostra a Tabela 15.

Tabela 15 - Listagem contendo os pontos, em mm, no arquivo TPS da *E. chaly*

LM=18	
130.00000	400.00000
148.00000	457.00000
130.00000	468.00000
351.00000	662.00000
376.00000	724.00000
614.00000	422.00000
616.00000	480.00000
534.00000	544.00000
668.00000	662.00000
550.00000	687.00000
600.00000	754.00000
537.00000	789.00000
734.00000	756.00000

719.00000	681.00000
855.00000	791.00000
884.00000	562.00000
927.00000	674.00000
1071.00000	920.00000

A Tabela 15 mostra um exemplo de como os dados estão disponibilizados no arquivo TPS, sendo que, a primeira linha mostra qual a quantidade de marcos anatômicos que foram marcados na imagem (LM=18). As demais linhas mostram as coordenadas X (na coluna da esquerda) e Y (na coluna da direita) ainda não alinhadas.

O conjunto de dados, proveniente da equipe do Professor Doutor Tiago M. Franco, professor da Universidade de São Paulo (USP/EACH), é composta de:

- Vinte e nove amostras da *E. chaly*; vinte e nove amostras da *E. flammea*; vinte e oito amostras da espécie *E. ignita*; vinte e nove amostras da *E. imperialis* e vinte e oito amostras da *E. orellana*;
- Existem dezoito marcos anatômicos para cada espécie;

A Tabela 15 mostra em parte como esses dados são apresentados no arquivo TPS, porém, a Tabela 15 mostra apenas uma amostra da espécie *E. chaly*, para simplificação.

Os passos subsequentes descrevem o procedimento adotado para a classificação das espécies.

1. O primeiro passo é separar as espécies manualmente do arquivo TPS para os cálculos dos centróides;
2. Realizar o cálculo das espécies alinhadas no software TPSRelw, e gravar os resultados em arquivos distintos para cada espécie;
3. Formar o conjunto de dados, com os pontos alinhados, de todas as espécies sequencialmente;
4. Armazenar os dados do conjunto em um vetor;

5. Efetuar o cálculo dos vizinhos mais próximos entre as espécies do conjunto e a espécie fora do conjunto para determinar o padrão de cada espécie no conjunto de dados;
6. Efetuar a análise dos gráficos de dispersão para a análise das distâncias nos marcos anatômicos de cada espécie (treinamento);
7. Treinar e testar os dados para a classificação.

A Figura 25 apresenta o gráfico de dispersão das distâncias dos marcos anatômicos das vinte e nove amostras da espécie *E. chaly*.

Figura 25 - Gráfico de dispersão dos marcos anatômicos da espécie *E. chaly*

São apresentados na Figura 25 os dezoito marcos anatômicos (na cor vermelha) de cada uma das vinte e nove amostras da espécie *E. chaly*. Uma análise da Figura 25 mostra que para esta espécie todos os marcos anatômicos possuem pequena distância entre os pontos demarcados, sendo que isso implica, em primeira instância, que todos esses marcos anatômicos podem ser considerados como marcos matemáticos significativos. A Figura 26 mostra o gráfico de dispersão para análise das distâncias dos marcos anatômicos da espécie *E. flammea*.

Figura 26 - Gráfico de dispersão dos marcos anatômicos da espécie *E. flammea*

São apresentados na Figura 26 os dezoito marcos anatômicos (na cor vermelha) de cada uma das vinte e nove amostras da espécie *E. flammea*. A Figura 26 mostra que, para esta espécie, todos os marcos anatômicos possuem pequena distância entre os pontos demarcados, da mesma maneira que ocorreu na espécie 0, chamada de *E. chaly*, sendo que isto implica, em primeira instância, que todos os marcos anatômicos podem ser considerados como marcos matemáticos significativos. A Figura 27 mostra o gráfico de dispersão para análise das distâncias dos marcos anatômicos da espécie *E. ignita*.

Figura 27 - Gráfico de dispersão dos marcos anatômicos da espécie *E. ignita*

São apresentados na Figura 27 os dezoito marcos anatômicos (na cor vermelha) de cada uma das vinte e oito amostras da espécie *E. ignita*. A Figura 27 mostra que, para esta espécie, todos os marcos anatômicos possuem pequena distância entre os pontos demarcados, da mesma maneira que ocorreu na espécie 0 e 1 (chamada de *E. flammea*), sendo que isto implica, em primeira instância, que todos os marcos anatômicos podem ser considerados como marcos matemáticos mais significativos. A Figura 28 mostra o gráfico de dispersão para análise das distâncias nos marcos anatômicos da espécie *E. imperialis*.

Figura 28 - Gráfico de dispersão dos marcos anatômicos da espécie *E. imperialis*

São apresentados na Figura 28 os dezoito marcos anatômicos (na cor vermelha) de cada uma das vinte e nove amostras da espécie *E. imperialis*. A Figura 28 mostra que, para esta espécie, todos os marcos anatômicos possuem pequena distância entre os pontos demarcados, da mesma maneira que ocorreu na espécie 0, 1 e 2 (chamada de *E. ignita*), sendo que isto implica, em primeira instância, que todos esses marcos anatômicos podem ser considerados como marcos matemáticos mais significativos. A Figura 29 mostra o gráfico de dispersão para análise das distâncias dos marcos anatômicos da espécie *E. orellana*.

Figura 29 - Gráfico de dispersão dos marcos anatômicos da espécie *E. orellana*

São apresentados na Figura 29 os dezoito marcos anatômicos (na cor vermelha) de cada uma das vinte e oito amostras da espécie *E. orellana*. A Figura 29 mostra que, em primeira instância, para esta espécie, todos os marcos anatômicos possuem pequena distância entre os pontos demarcados, da mesma maneira que ocorreu na espécie 0, 1, 2 e 3 (chamada *E. imperialis*), sendo que isto implica, em primeira instância, que todos esses marcos anatômicos podem ser considerados como marcos matemáticos mais significativos.

Como especificado no item 5, nos passos descritos para a classificação das *Euglossas* pelo algoritmo FNN4Bees, o próximo passo é realizar a inserção dos dados do conjunto no algoritmo kNN juntamente com os novos dados das espécies. Após isso, deve-se executar o algoritmo para realizar a primeira etapa de classificação entre as espécies *Euglossas*. A Tabela 16 mostra o resultado do cálculo das distâncias fornecido pelo algoritmo FNN4Bees entre a espécie 0 (*E. chaly*) como treinamento e as demais espécies do conjunto de dados.

Tabela 16 - Resultado, em mm, para o teste com a *E. chaly* como entrada

Distâncias	0_(chaly_ Teste)	1_(flammea)	2_(ignita)	3_(imperialis)	4_(orellana)
LM0	0.0075884	0.0217628	0.0309340	0.0346350	0.0287667
LM1	0.0066512	0.0403725	0.0575408	0.1009031	0.0693438
LM2	0.0005981	0.0155618	0.0241298	0.0291096	0.0207468
LM3	0.0006568	0.0410812	0.0602063	0.1003596	0.0644686
LM4	0.0034546	0.0161224	0.0252173	0.0274669	0.0209825
LM5	0.0012895	0.0449931	0.0646456	0.1088910	0.0695155
LM6	0.0054421	0.0079622	0.0121967	0.0198692	0.0031696
LM7	0.0020398	0.0224926	0.0374763	0.0516235	0.0313972
LM8	0.0064391	0.0090453	0.0142597	0.0294242	0.0124073
LM9	0.0005252	0.0170642	0.0295254	0.0396725	0.0267899
LM10	0.0072669	0.0085278	0.0242095	0.0507005	0.0146923
LM11	0.0019126	1.5534415	0.0016557	0.0028529	0.0034727
LM12	0.0042399	0.0048306	0.0193079	0.0363345	0.0105205
LM13	0.0041944	0.0021228	0.0026908	0.0006663	0.0040832
LM14	0.0043236	0.0010529	0.0051008	0.0170430	0.0041577
LM15	0.0047314	0.0042369	0.0119893	0.0151672	0.0093188
LM16	0.0054454	0.0036746	0.0079470	0.0001398	0.0020166
LM17	0.0021238	0.0087810	0.0157108	0.0215309	0.0150996

A Tabela 16 mostra, para cada marco anatômico, qual foi a distância obtida entre a amostra da espécie 0 e as demais amostras do conjunto de dados. Como existem no total dezoito marcos anatômicos, o número de linhas da Tabela é dezoito e como existem cinco espécies, o número de colunas da Tabela é cinco, com o '0' representando a espécie *E. chaly* (espécie de teste, pois é a mesma da espécie comparada), o 1 representando a espécie *E. flammea*, o 2 representando a espécie *E. ignita*, o número 3 representando a espécie *E. imperialis* e o número 4 representando a espécie *E. orellana*.

As distâncias totais, calculada por meio da média das distâncias de cada marco anatômico, são dadas pelo algoritmo da Tabela 17.

Tabela 17 - Média das distâncias, em mm, totais calculadas pelo algoritmo FNN4Bees entre as espécies *Euglossas*

Espécies	0_(chaly_ Teste)	1_(flammea)	2_(ignita)	3_(imperialis)	4_(orellana)
chaly (Base)	0.6035164	35.36664789	42.5299242	53.030736155	41.71565491

Entre as distâncias apresentadas na Tabela 17, a espécie que possui menor distância com a espécie comparada (chaly_base) é a chaly de teste, ou seja, as

amostras de teste foram classificadas corretamente em seu grupo. Pode-se ainda notar que a distância entre as amostras das outras espécies e da espécie *E. chaly* da base são grandes, apresentando uma grande diferença entre essas espécies. Mesmo que a classificação tenha sido correta, uma análise dos marcos anatômicos da espécie *E. chaly* de teste mostra que os marcos anatômicos mais significativos, no algoritmo FNN4Bees são dados por: **0, 1, 2, 3, 4, 5, 7, 9 e 17**, ou seja, esses são os marcos anatômicos mais significativos para a espécie *E. chaly*. Caso a classificação tivesse de ser melhorada, o próximo cálculo de distância seria somente entre esses marcos anatômicos da espécie *E. chaly* com as demais espécies. Esses marcos anatômicos mais significativos são usados posteriormente no cálculo da Fuzzy kNN.

A próxima etapa visa comparar a espécie *E. flammea* com as demais espécies para verificar como ela se comporta no algoritmo FNN4Bees. A Tabela 18 mostra o resultado do cálculo das distâncias fornecido pelo algoritmo FNN4Bees entre a espécie 1 (*E. flammea*) como treinamento e as demais espécies do conjunto de dados.

Tabela 18 - Diferenças entre os marcos anatômicos, em mm, da espécie *E. flammea* com as demais espécies

Distâncias	0_(chaly)	1_(flammea_ Teste)	2_(ignita)	3_(imperialis)	4_(orellana)
LM0	0.0070084	0.0071659	0.0163371	0.0200382	0.0141698
LM1	0.0484531	0.0014293	0.0157389	0.0591012	0.0275420
LM2	0.0155069	0.0006530	0.0092209	0.0142007	0.0058379
LM3	0.0389165	0.0028216	0.0219467	0.0621000	0.0262089
LM4	0.0148959	0.0022281	0.0068667	0.0091163	0.0026319
LM5	0.0433103	0.0003932	0.0200457	0.0642911	0.0249156
LM6	0.0006133	0.0019067	0.0061413	0.0198692	0.0028858
LM7	0.0246395	0.0041867	0.0107969	0.0138138	0.0047178
LM8	0.0025612	0.0000499	0.0052594	0.0249441	0.0034069
LM9	0.0145636	0.0019753	0.0144365	0.0204238	0.0117009
LM10	0.0139145	0.0018802	0.0175619	0.0245835	0.0080447
LM11	0.0056052	0.0038479	0.0020368	0.0440528	0.0002198
LM12	0.0119303	0.0028596	0.0116176	0.0065455	0.0028301
LM13	0.0087268	0.0024095	0.0018415	0.0038660	0.0004491
LM14	0.0044121	0.0009644	0.0050123	0.0169545	0.0040692
LM15	0.0003989	0.0008934	0.0068589	0.0100368	0.0041884
LM16	0.0041445	0.0049754	0.0092479	0.0011610	0.0033175
LM17	0.0107049	0.0040477	0.0028819	0.0087021	0.0022708

A Tabela 18 mostra, para cada marco anatômico, qual foram as distâncias obtidas entre a amostra da espécie 1 e as demais espécies do conjunto de dados.

As médias das distâncias totais são dadas pelo algoritmo da Tabela 19.

Tabela 19 - Média das distâncias totais, em mm, calculadas pelo algoritmo FNN4Bees entre a espécie *E. flammea* e as espécies *Euglossas*

Espécies	0_(chaly)	1_(flammea_Testes)	2_(ignita)	3_(imperialis)	4_(orellana)
flammea (Base)	13.0240639	0.5334442	11.4538757	25.5932672	12.1590503

Entre as médias das distâncias apresentadas na Tabela 19, a espécie que possui a menor média de distância com a espécie comparada (flammea_base) é a flammea de teste, ou seja, as amostras de teste foram classificadas corretamente em seu grupo. Pode-se ainda notar que a distância entre as amostras das outras espécies e da espécie *E. flammea* da base, da mesma forma que aconteceu com a espécie *E. chaly*, na comparação passada, são grandes, apresentando uma grande diferença entre essas espécies.

Mesmo que a classificação tenha sido correta, uma análise dos marcos anatômicos da espécie *E. flammea* de teste mostra que os marcos anatômicos mais significativos, no algoritmo FNN4Bees são dados por: **1, 2, 3, 5, 8, 9, 10 e 14**, ou seja, esses são os marcos anatômicos mais significativos para a espécie *E. flammea*, esses marcos formam o padrão da espécie *E. flammea*.

A próxima etapa visa comparar a espécie *E. ignita* com as demais espécies para verificar como ela se comporta no algoritmo FNN4Bees. A Tabela 20 mostra o resultado do cálculo das distâncias fornecido pelo algoritmo FNN4Bees entre a espécie 2 (*E. ignita*) como treinamento e as demais espécies do conjunto de dados.

Tabela 20 - Diferenças entre os marcos anatômicos, em mm, da espécie *E. ignita* com as demais espécies

Distâncias	0_(chaly)	1_(flammea)	2_(ignita Teste)	3_(imperialis)	4_(orellana)
LM0	0.0277813	0.0136069	0.0044357	0.0007347	0.0066030
LM1	0.0661780	0.0191542	0.0019859	0.0413763	0.0098171
LM2	0.0235569	0.0073969	0.0011709	0.0061507	0.0022120
LM3	0.0612699	0.0195317	0.0004067	0.0397466	0.0038555
LM4	0.0193429	0.0066752	0.0024196	0.0046693	0.0018150
LM5	0.0637107	0.0200071	0.0003546	0.0438907	0.0045152
LM6	0.0071601	0.0046400	0.0004055	0.0072670	0.0094326
LM7	0.0390721	0.0186193	0.0036355	0.0105115	0.0097147
LM8	0.0075131	0.0049069	0.0003074	0.0154719	0.0015450
LM9	0.0300513	0.0135123	0.0010511	0.0090959	0.0037867
LM10	0.0276890	0.0118943	0.0037873	0.0302783	0.0057298
LM11	0.0027289	0.0009716	0.0008394	0.0036692	0.0002198
LM12	0.0202944	0.0112238	0.0032534	0.0202800	0.0026564
LM13	0.0055456	0.0007716	0.0013396	0.0006848	0.0055339
LM14	0.0008536	0.0045229	0.0085708	0.0205130	0.0027320
LM15	0.0003989	0.0043600	0.0033924	0.0065703	0.0076277
LM16	0.0106768	0.0015568	0.0027156	0.0053712	0.0032147
LM17	0.0125305	0.0058732	0.0010564	0.0068765	0.0004452

A Tabela 20 mostra, para cada marco anatômico, qual foi a distância obtida entre a amostra padrão da espécie 2 e as demais espécies do conjunto de dados.

As distâncias totais são dadas pelo algoritmo da Tabela 21.

Tabela 21 - Média das distâncias totais, em mm, calculadas pelo algoritmo FNN4Bees entre a espécie *E. ignita* e as espécies *Euglossas*

Espécies	0_(chaly)	1_(flammea Teste)	2_(ignita)	3_(imperialis)	4_(orellana)
flammea	22.6476523	8.6818728	0.5008225	16.3293934	5.5582331
(Base)					

Entre as médias das distâncias apresentadas na Tabela 21, a espécie que possui menor distância com a espécie comparada (ignita_base) é a ignita de teste, ou seja, as amostras de teste foram classificadas corretamente em seu grupo. Pode-se ainda notar que a distância entre as amostras das outras espécies e da espécie *E. ignita* da base, da mesma forma que aconteceu com a espécie *E. chaly* e *E. flammea*, na comparação passada, são grandes, apresentando uma grande diferença entre essas espécies.

Mesmo que a classificação tenha sido correta, uma análise dos marcos anatômicos da espécie *E. ignita* de teste mostra que os marcos anatômicos mais significativos,

no algoritmo FNN4Bees são dados por: **3, 5, 6, 7 e 8**, ou seja, esses marcos formam o padrão para a espécie *E. ignita*.

A próxima etapa visa comparar a espécie *E. imperialis* com as demais espécies para verificar como ela se comporta no algoritmo FNN4Bees. A Tabela 22 mostra o resultado do cálculo das distâncias fornecido pelo algoritmo FNN4Bees entre a espécie 3 (*E. imperialis*), considerando-se todos os marcos como o padrão da espécie, como treinamento e as demais espécies do conjunto de dados.

Tabela 22 - Diferenças entre os marcos anatômicos, em mm, da espécie *E. imperialis* com as demais espécies

Distâncias	0_(chaly)	1_(flammea)	2_(ignita)	3_(imperialis_teste)	4_(orellana)
LM0	0.0285394	0.0143650	0.0051937	0.0014927	0.0073611
LM1	0.1092167	0.0621929	0.0450245	0.0016623	0.0332215
LM2	0.0227925	0.0066325	0.0019353	0.0069151	0.0014475
LM3	0.1017379	0.0599997	0.0408747	0.0007214	0.0366124
LM4	0.0152452	0.0025775	0.0065173	0.0087669	0.0022825
LM5	0.1048146	0.0611110	0.0414585	0.0027868	0.0365886
LM6	0.0061299	0.0036098	0.0006246	0.0082972	0.0084024
LM7	0.0500398	0.0295869	0.0146032	0.0004561	0.0206824
LM8	0.0085936	0.0059874	0.0007730	0.0143914	0.0026255
LM9	0.0372007	0.0206617	0.0082005	0.0019464	0.0109361
LM10	0.0532290	0.0374343	0.0217526	0.0047383	0.0312698
LM11	0.0011234	0.0028807	0.0046918	0.0001831	0.0065088
LM12	0.0407902	0.0317196	0.0172423	0.0002158	0.0260298
LM13	0.0005709	0.0057463	0.0063143	0.0042898	0.0077067
LM14	0.0179779	0.0126012	0.0085533	0.0033888	0.0094964
LM15	0.0143728	0.0148673	0.0071148	0.0039369	0.0097854
LM16	0.0049448	0.0041752	0.0084476	0.0003607	0.0025172
LM17	0.0193537	0.0126965	0.0057667	0.0000533	0.0063779

A Tabela 22 mostra, para cada marco anatômico, qual foi a distância obtida entre a amostra padrão da espécie 3 e as demais espécies do conjunto de dados.

As médias das distâncias totais são dadas pelo algoritmo da Tabela 23.

Tabela 23 - Médias das distâncias totais, em mm, calculadas pelo algoritmo FNN4Bees entre a espécie *E. imperialis* e as espécies *Euglossas*

Espécies	0_(chaly)	1_(flammea)	2_(ignita)	3_(imperialis_teste)	4_(orellana)
flammea	35.9050042	21.9187043	13.3141893	0.3726735	14.0524675
(Base)					

Entre as médias das distâncias apresentadas na Tabela 23, a espécie que possui menor distância com a espécie comparada (*imperialis_base*) é a *imperialis* de teste, ou seja, as amostras de teste foram classificadas corretamente em seu grupo.

Os marcos anatômicos mais significativos, no algoritmo FNN4Bees, para a espécie 3 são dados por: **1, 3, 5, 7, 10, 11, 12, 16 e 17**, ou seja, esses são os marcos anatômicos mais significativos que formam o padrão da espécie *E. imperialis*.

A próxima etapa visa comparar a espécie *E. orellana* com as demais espécies para verificar como ela se comporta no algoritmo FNN4Bees. A Tabela 24 mostra o resultado do cálculo das distâncias fornecido pelo algoritmo FNN4Bees entre a espécie 4 (*E. orellana*) como treinamento e as demais espécies do conjunto de dados.

Tabela 24 - Diferenças entre os marcos anatômicos, em mm, da espécie *E. orellana* com as demais espécies

Distâncias	0_(chaly)	1_(flammea)	2_(ignita)	3_(imperialis)	4_(orellana_teste)
LM0	0.0126601	0.0015142	0.0106854	0.0143864	0.0085181
LM1	0.0702656	0.0232417	0.0060734	0.0372888	0.0057295
LM2	0.0233480	0.0071880	0.0013798	0.0063596	0.0020031
LM3	0.0620094	0.0202712	0.0011462	0.0390071	0.0031160
LM4	0.0192446	0.0065768	0.0025180	0.0047676	0.0017167
LM5	0.0650612	0.0213576	0.0017051	0.0425402	0.0017167
LM6	0.0065189	0.0039988	0.0002356	0.0079082	0.0087914
LM7	0.0302368	0.0097839	0.0051997	0.0193468	0.0008794
LM8	0.0073645	0.0047583	0.0004560	0.0156205	0.0008794
LM9	0.0241211	0.0075821	0.0048791	0.0150261	0.0021435
LM10	0.0329611	0.0171663	0.0014846	0.0250062	0.0110018
LM11	0.0034126	0.0016552	0.0001557	0.0043529	0.0019727
LM12	0.0264864	0.0174158	0.0029384	0.0140880	0.0117259
LM13	0.0032381	0.0030790	0.0036471	0.0016225	0.0050394
LM14	0.0145242	0.0091476	0.0050997	0.0068424	0.0060428
LM15	0.0042512	0.0047457	0.0030066	0.0061845	0.0003361
LM16	0.0055777	0.0035422	0.0078146	0.0002722	0.0018842
LM17	0.0131564	0.0064992	0.0004304	0.0062506	0.0001806

A Tabela 24 mostra, para cada marco anatômico, qual foi a distância obtida entre a amostra da espécie 4 e as demais amostras do conjunto de dados.

As médias das distâncias totais são dadas pelo algoritmo da Tabela 25.

Tabela 25 - Médias das distâncias totais calculadas, em mm, pelo algoritmo FNN4Bees entre a espécie *E. orellana* e as espécies *Euglossas*

Espécies	0_(chaly)	1_(flammea)	2_(ignita)	3_(imperialis)	4_(orellana_teste)
flammea	25.4172925	11.6697520	5.91364087	13.1484714	0.34341100
(Base)					

Entre as distâncias apresentadas na Tabela 25, a espécie que possui menor distância com a espécie comparada (orellana_base) é a orellana de teste, ou seja, as amostras de teste foram classificadas corretamente em seu grupo. Os marcos anatômicos mais significativos, no algoritmo FNN4Bees, para a espécie 4, são dados por: **1, 3, 5, 7, 10, 11, 12, 16 e 17**, ou seja, esses são os marcos anatômicos mais significativos para a espécie *E. orellana*.

4.2.1 Aplicação da Fuzzy kNN para as *Euglossas*

A Tabela 26 a seguir mostra todos os marcos anatômicos padrões adquiridos a partir dos marcos anatômicos mais significativos para todas as espécies, incluindo a amostra da espécie indefinida.

Tabela 26 - Valores, em mm, de todos os marcos anatômicos padrões das espécies indefinida (N), 0, 1, 2, 3 e 4

LM	N_x	N_y	0_x	0_y	1_x	1_y	2_x	2_y	3_x	3_y	4_x	4_y
LM0	-3,510	-1,833	-3,397	-1,665	-3,763	-8,870	-3,744	-1,191	-3,781	-7,581	-3,722	-1,073
LM1	-3,370	-1,389	-3,392	-1,427	-3,617	-5,425	-3,617	-7,806	-3,667	-3,790	-3,583	-7,379
LM2	-3,510	-1,303	-3,537	-1,331	-3,705	-4,032	-3,728	-6,698	-3,750	-2,274	-3,685	-6,212
LM3	-1,787	2,0922	-1,825	2,0695	-1,689	6,3049	-1,719	5,6359	-1,643	7,0506	-1,810	5,0280
LM4	-1,592	6,9266	-1,629	6,5811	-1,447	1,0777	-1,514	9,8266	-1,362	1,0841	-1,532	9,5531
LM5	2,6337	-1,662	2,0267	-1,697	-7,616	-1,700	-5,139	-1,697	-3,163	-1,652	4,3781	-1,716
LM6	2,7896	-1,209	2,2585	-1,298	-2,849	-1,319	4,3487	-1,278	-1,267	-1,258	1,3136	-1,292
LM7	-3,604	-7,108	-4,052	-7,490	-5,600	-6,451	-4,546	-6,382	-5,740	-6,065	-4,452	-6,649
LM8	6,8444	2,0923	7,8329	2,2923	8,0359	2,1995	8,1836	7,3362	7,3749	1,5160	7,5906	7,9474
LM9	-2,356	4,0416	-1,718	4,0920	-1,714	4,6920	-1,067	4,6870	-5,855	4,6245	-1,824	4,5900

LM10	1,5422	9,2658	2,0206	9,2751	4,2236	8,2843	3,5977	8,8778	4,7972	7,8846	3,6492	8,5313
LM11	-3,370	1,1995	-4,217	1,1786	-2,227	1,2683	-1,620	1,2199	-2,822	1,2206	-2,189	1,2399
LM12	1,1990	9,4218	1,2698	9,7808	1,3827	6,3049	1,4509	7,2174	1,4577	5,5344	1,3575	6,9256
LM13	1,0821	3,5738	1,1861	3,6702	1,1628	5,8652	1,2453	1,5242	1,1848	4,5485	1,1969	1,1597
LM14	2,1425	1,2150	2,0620	1,2296	2,3431	6,8181	2,3839	8,4034	2,3068	5,6860	2,2115	8,2394
LM15	2,3686	-5,705	2,3689	-5,917	2,2258	-1,121	2,1151	-9,940	2,1097	-1,213	2,2407	-9,861
LM16	2,7039	3,0279	2,6835	3,0481	2,7463	-3,299	2,6765	-1,717	2,6783	-4,624	2,7443	-1,686
LM17	3,8267	2,2209	3,7969	2,2724	4,1686	1,2829	4,0049	1,5124	4,3538	1,1144	4,1310	1,5392

As distâncias entre a espécie indefinida e as demais espécies dadas pela execução do algoritmo FNN4Bees são apresentadas na Tabela 27.

Tabela 27 - Distâncias, em mm, entre a espécie indefinida e as outras três espécies

LM Sig	Dist N_0	LM Sig	Dist N_1	LM Sig	Dist N_2	LM Sig	Dist N_3	LM Sig	Dist N_4
0	0.00758	1	0.04037	3	0.06020	1	0.10090	4	0.02269
1	0.00665	2	0.01556	5	0.06464	3	0.10035	7	0.03227
2	0.00059	3	0.04108	6	0.01219	5	0.10889	9	0.02464
3	0.00065	5	0.04499	7	0.03747	7	0.05162	15	0.00898
4	0.00345	8	0.04499	8	0.01425	10	0.05070	17	0.01528

Para o cálculo da Fuzzy kNN a quantidade de marcos anatômicos significativos tem de ser a mesma para todas as espécies, por isso, obtém-se o número mínimo de marcos anatômicos significativos, que no exemplo é 5 (cinco), e esse valor mínimo é utilizado como quantidade de marcos anatômicos significativos para todas as espécies.

Aplicando a eq.(19), para $m = 2$, nas espécies 0, 1, 2, 3 e 4, são geradas, respectivamente, as equações eq.(26), eq.(27), eq.(28), eq.(29) e eq.(30).

$$U_0 = 0.8392 \quad (26)$$

$$U_1 = 0.0348 \quad (27)$$

$$U_2 = 0.0465 \quad (28)$$

$$U_3 = 0.0150 \quad (29)$$

$$U_4 = 0.0641 \quad (30)$$

A função Fuzzy $Ch(a)$, neste caso é dada por eq.(31).

$$f(Ch(a)) = \text{função}(0.8392/a_0, 0.0348/a_1, 0.0465/a_2, 0.0150/a_3, 0.0641/a_4) \quad (31)$$

A Figura 30 mostra o gráfico de dispersão dos valores obtidos a partir da função de pertinência da espécie *E. chaly*.

Figura 30 - Grau de Pertinência da amostra de teste da espécie *E. chaly* com as demais espécies

O gráfico da função Fuzzy, neste caso, apresenta-se como na Figura 35.

A partir da análise dos dados, nota-se que a espécie que tem maior valor na execução da função de pertinência é a espécie 0 (*E. chaly*). A aplicação da Fuzzy kNN ajudou a determinar, em melhor escala, o grau a que uma amostra da espécie 0 (chamada de amostra indefinida) pertence ao seu respectivo grupo e em comparação aos demais grupos. A diferença entre os espécimes da espécie 0 e os espécimes das demais espécies é bastante significativa, de acordo com a análise do grau de pertinência. Isso significa que o espécime da espécie 0 pertence em cerca de 83% ao seu respectivo grupo, enquanto que esta informação não era possível de ser obtida apenas pelo algoritmo kNN. A partir desses dados, uma análise sobre a relação entre as espécies pode ser realizada. No exemplo dado, a amostra indefinida tinha muito pouca probabilidade de ser da espécie 3 (*E. imperialis*) o que mostra que a diferença entre a *Forma* da espécie 0 e da espécie 3 é maior do que da espécie 0 com as demais espécies.

Da mesma forma, segue o cálculo da aplicação da Fuzzy kNN para as demais espécies. Utilizando-se como amostra indefinida uma amostra da espécie *E. flammea*, tem-se as seguintes distâncias entre a amostra indefinida e a amostra padrão das demais espécies, apresentada na Tabela 28.

Tabela 28 - Distâncias, em mm, entre a espécie *E. flammea* e as demais espécies

LM Sig	Dist N_0	LM Sig	Dist N_1	LM Sig	Dist N_2	LM Sig	Dist N_3	LM Sig	Dist N_4
0	0.00700	1	0.00142	3	0.02194	1	0.05910	4	0.00263
1	0.04845	2	0.00065	5	0.02004	3	0.06210	7	0.00471
2	0.01550	3	0.00282	6	0.00614	5	0.06429	9	0.01170
3	0.03891	5	0.00039	7	0.01079	7	0.01381	15	0.00418
4	0.01489	8	0.00004	8	0.00525	10	0.02458	17	0.00227

Aplicando a eq.(19) para $m = 2$ nas espécies 0, 1, 2, 3 e 4, são geradas, respectivamente, as equações eq.(32), eq.(33), eq.(34), eq.(35) e eq.(36).

$$U_0 = 0.0099 \quad (32)$$

$$U_1 = 0.9268 \quad (33)$$

$$U_2 = 0.0166 \quad (34)$$

$$U_3 = 0.0050 \quad (35)$$

$$U_4 = 0.0417 \quad (36)$$

A função Fuzzy $FI(a)$, neste caso é dada por eq.(37)

$$f(FI(a)) = \text{função}(0.0099/a_0, 0.9268/a_1, 0.0166/a_2, 0.0050/a_3, 0.0417/a_4) \quad (37)$$

A Figura 31 mostra o grau de pertinência da amostra de teste *E. flammea* com todas as espécies *Euglossas*.

Figura 31 - Grau de Pertinência da amostra de teste da espécie *E. flammea* com as demais espécies

A partir da análise dos dados, nota-se que a espécie que tem maior valor na execução da função de pertinência é a espécie 1 (*E. flammea*), como queria-se provar.

Utilizando-se como amostra indefinida uma amostra da espécie *E. ignita*, tem-se as seguintes distâncias entre a amostra indefinida e a amostra padrão das demais espécies, apresentada na Tabela 29.

Tabela 29 - Distâncias, em mm, entre a espécie *E. ignita* (espécie 3) e as demais espécies

LM Sig	Dist N_0	LM Sig	Dist N_1	LM Sig	Dist N_2	LM Sig	Dist N_3	LM Sig	Dist N_4
0	0.02778	1	0.01915	3	0.00040	1	0.04137	4	0.00181
1	0.06617	2	0.00739	5	0.00035	3	0.03974	7	0.00971
2	0.02355	3	0.01953	6	0.00040	5	0.04389	9	0.00378
3	0.06126	5	0.02000	7	0.00363	7	0.01051	15	0.00762
4	0.01934	8	0.00490	8	0.00030	10	0.03027	17	0.00044

Aplicando a eq.(19) para $m = 2$ nas espécies 0, 1, 2, 3 e 4, são geradas, respectivamente, as equações eq.(38), eq.(39), eq.(40), eq.(41) e eq.(42).

$$U_0 = 0.0103 \quad (38)$$

$$U_1 = 0.0315 \quad (39)$$

$$U_2 = 0.7329 \quad (40)$$

$$U_3 = 0.0128 \quad (41)$$

$$U_4 = 0.2125 \quad (42)$$

A função Fuzzy $Ig(a)$, neste caso é dada por eq.(43)

$$f(Ig(a)) = \text{função}(0.0103/a_0, 0.0315/a_1, 0.7329/a_2, 0.0128/a_3, 0.2125/a_4) \quad (43)$$

A Figura 32 mostra o grau da amostra de teste *E. ignita* com as demais espécies do conjunto de dados.

Figura 32 - Grau de Pertinência da amostra de teste *E. ignita* com as demais espécies

A partir da análise dos dados, nota-se que a espécie que tem maior valor na execução da função de pertinência é a espécie 2 (*E. ignita*), como queria-se provar. Utilizando-se como amostra indefinida uma amostra da espécie *E. imperialis*, tem-se as seguintes distâncias entre a amostra indefinida e a amostra padrão das demais espécies, apresentada na Tabela 30.

Tabela 30 - Distâncias, em mm, entre a espécie *E. imperialis* e as demais espécies

LM Sig	Dist N_0	LM Sig	Dist N_1	LM Sig	Dist N_2	LM Sig	Dist N_3	LM Sig	Dist N_4
0	0.02853	1	0.06219	3	0.04087	1	0.00166	4	0.00228
1	0.10921	2	0.00663	5	0.04145	3	0.00072	7	0.02068
2	0.02279	3	0.05999	6	0.00062	5	0.00278	9	0.01093
3	0.10173	5	0.06111	7	0.01460	7	0.00045	15	0.00978
4	0.01524	8	0.00598	8	0.00077	10	0.00473	17	0.00637

Aplicando a eq.(19) para $m = 2$ nas espécies 0, 1, 2, 3 e 4, são geradas, respectivamente, as equações eq.(44), eq.(45), eq.(46), eq.(47) e eq.(48).

$$U_0 = 0.0178 \quad (44)$$

$$U_1 = 0.0400 \quad (45)$$

$$U_2 = 0.3299 \quad (46)$$

$$U_3 = 0.5211 \quad (47)$$

$$U_4 = 0.0912 \quad (48)$$

A função Fuzzy $Im(a)$, neste caso é dada por eq.(49)

$$f(Im(a)) = \text{função}(0.0178/a_0, 0.0400/a_1, 0.3299/a_2, 0.5211/a_3, 0.0912/a_4) \quad (49)$$

A Figura 33 mostra o grau de pertinência da amostra de teste *E. imperialis* e as demais espécies.

Figura 33 - Grau de Pertinência da amostra de teste *E. imperialis* com as demais espécies

A partir da análise dos dados, nota-se que a espécie que tem maior valor na execução da função de pertinência é a espécie 3 (*E. imperialis*), como queria-se provar.

Utilizando-se como amostra indefinida uma amostra da espécie *E. orellana*, tem-se as seguintes distâncias entre a amostra indefinida e a amostra padrão das demais espécies, apresentada na Tabela 31.

Tabela 31 - Distâncias, em mm, entre a espécie *E. orellana* e as demais espécies

LM Sig	Dist N_0	LM Sig	Dist N_1	LM Sig	Dist N_2	LM Sig	Dist N_3	LM Sig	Dist N_4
0	0.01266	1	0.02324	3	0.00114	1	0.03728	4	0.00171
1	0.07026	2	0.00718	5	0.00170	3	0.03900	7	0.00087
2	0.02334	3	0.02027	6	0.00023	5	0.04254	9	0.00214
3	0.06200	5	0.02135	7	0.00519	7	0.01934	15	0.00033
4	0.01924	8	0.00475	8	0.00045	10	0.02500	17	0.00018

Aplicando a eq.(19) para $m = 2$ nas espécies 0, 1, 2, 3 e 4, são geradas, respectivamente, as equações eq.(50), eq.(51), eq.(52), eq.(53) e eq.(54).

$$U_0 = 0.0103 \quad (50)$$

$$U_1 = 0.0246 \quad (51)$$

$$U_2 = 0.4140 \quad (52)$$

$$U_3 = 0.0084 \quad (53)$$

$$U_4 = 0.5427 \quad (54)$$

A função Fuzzy $O(a)$, neste caso é dada por eq(55).

$$f(O(a)) = \text{função}(0.0103/a_0, 0.0246/a_1, 0.4140/a_2, 0.0084/a_3, 0.5427/a_4) \quad (55)$$

A Figura 34 apresenta os graus de pertinência entre a amostra de teste *E. orellana* e as demais amostras padrões.

Figura 34 - Grau de Pertinência da amostra de teste *E. orellana* com as demais espécies

A partir da análise dos dados, nota-se que a espécie que tem maior valor na execução da função de pertinência é a espécie 4 (*E. orellana*), como queria-se provar. Porém, de acordo com os graus de pertinência, pode-se notar que as espécies *E. ignita* e *E. orellana* estão com graus muito próximos, respectivamente, 0.4140 e 0.5427. O algoritmo mostra uma grande proximidade entre essas duas espécies, de acordo com os marcos anatômicos estabelecidos e com a amostra

padrão adotada para as espécies. Porém, o resultado continua sendo o esperado na classificação. Atualmente, não existem estudos que possam provar que existe uma similaridade ou uma relação de parentesco maior entre essas espécies, tanto morfológicamente, quanto em respeito à outras técnicas. Portanto, essa similaridade encontrada pelo algoritmo não pode ser discutida. Trata-se de uma evidência que precisa de mais estudos para ser confirmada.

A saída da função Fuzzy encontra-se da forma apresentada pela Figura 35.

Figura 35 - Saída da Função Fuzzy para cada classe de espécies *Euglossas*

A Figura 35 mostra a função triangular da Fuzzy tradicional para cada espécie, sendo que a função indicada como mf1 representa a função Fuzzy da *E. chaly*, a função indicada na figura como mf2 corresponde a função Fuzzy da *E. ignita*, a função triangular representada por mf3 corresponde a função Fuzzy da *E. flammea*, a função mf4 representa a *E. imperialis* e a mf5 representa a *E. orellana*.

Aplicando a interseção entre conjuntos Fuzzy, por meio da T-Norma, tem-se as seguintes equações eq.(56), eq.(57), eq.(58). eq.(59) e eq.(60).

Para a espécie *E. chaly*, a equação Fuzzy é dada por eq.(56).

$$T(\text{Ch, Fl, Ig, Im, O}) = \max (\inf \text{Ch}(a), \inf \text{Fl}(a), \inf \text{Ig}(a), \inf \text{Im}(a), \inf \text{O}(a)) = \max (0.8392, 0.0348, 0.0465, 0.0150, 0.0641) = \mathbf{0.8392} \quad (56)$$

A expressão que melhor representa a espécie é a: U0

Para a espécie *E. flammea*, a equação Fuzzy é dada por eq.(57).

$$T(\text{Ch}, \text{Fl}, \text{Ig}, \text{Im}, \text{O}) = \max (\inf \text{Ch}(a), \inf \text{Fl}(a), \inf \text{Ig}(a), \text{Im}(a), \text{O}(a)) = \max (0.0099, 0.9268, 0.0166, 0.0050, 0.0417) = \mathbf{0.9268} \quad (57)$$

A expressão que melhor representa a espécie é a: U1

Para a espécie *E. ignita*, a equação Fuzzy é dada por eq.(58).

$$T(\text{Ch}, \text{Fl}, \text{Ig}, \text{Im}, \text{O}) = \max (\inf \text{Ch}(a), \inf \text{Fl}(a), \inf \text{Ig}(a), \text{Im}(a), \text{O}(a)) = \max (0.0103, 0.0315, 0.7329, 0.0128, 0.2125) = \mathbf{0.7329} \quad (58)$$

A expressão que melhor representa a espécie é a: U2

Para a espécie *E. imperialis*, a equação Fuzzy é dada por eq.(59).

$$T(\text{Ch}, \text{Fl}, \text{Ig}, \text{Im}, \text{O}) = \max (\inf \text{Ch}(a), \inf \text{Fl}(a), \inf \text{Ig}(a), \text{Im}(a), \text{O}(a)) = \max (0.0178, 0.0400, 0.03299, 0.5211, 0.0912) = \mathbf{0.5211} \quad (59)$$

A expressão que melhor representa a espécie é a: U3

Para a espécie *E. orellana*, a equação Fuzzy é dada por eq.(60).

$$T(\text{Ch}, \text{Fl}, \text{Ig}, \text{Im}, \text{O}) = \max (\inf \text{Ch}(a), \inf \text{Fl}(a), \inf \text{Ig}(a), \text{Im}(a), \text{O}(a)) = \max (0.0103, 0.0246, 0.4140, 0.0084, 0.5427) = \mathbf{0.5427} \quad (60)$$

A expressão que melhor representa a espécie é a: U4

A Figura 36 mostra a relação apresentada pela Lógica Fuzzy entre as cinco espécies, apresentada pelas regras listadas acima, onde o eixo vertical apresenta os graus de pertinência entre cada espécie e o eixo horizontal apresenta as funções U1 como 1, U2 como 2, U3 como 3, U4 como 4 e U5 como 5.

Figura 36 - Disposição das Funções Fuzzy para cada espécie no Gráfico.

Pelos valores retornados por meio da aplicação da T-Norma, pode-se notar que as espécies 0,1 e 2 estão bem classificadas em seus respectivos grupos, enquanto as espécies 3 e 4 apresentam um grau menor de pertinência em seus grupos, por meio das amostras testadas. Uma melhora nesses resultados por ser adquirida por meio de novos testes com amostras diferentes que possuam os marcos anatômicos mais parecidos com a espécie padrão, ou seja, espécies cuja *Forma* seja mais parecida entre as amostras de teste.

A relação entre as espécies Euglossas, de acordo com Ramírez et al. (2010), é apresentada na Figura 37.

Figura 37 - Relação de Parentesco entre algumas das espécies de Euglossas.

De acordo com essa relação, pode-se dizer que as relações entre as espécies *E. ignita* e *E. orellana* são pertinentes para os resultados da classificação realizados

pelo FNN4Bees. Apesar de um pouco discrepantes, os resultados indicam que as espécies *E. ignita* e *E. orellana* estão proximamente relacionadas.

Apesar de outras espécies se mostrarem molecularmente mais aparentadas, fica claro pela figura que as duas espécies são próximas.

Uma possível explicação é que as modificações nos padrões de venação de asa não seguiram exatamente o mesmo caminho da evolução molecular, o que deixou as espécies um pouco mais distantes, mas próximas o suficiente para serem relacionadas pelas análises morfométricas.

Entre os métodos utilizados por Ramírez et al. (2010) para identificação da Filogenia, padrões de diversificação e biogeografia histórica entre as euglossini destacam-se os procedimentos moleculares, métodos filogenéticos e análise de diversificação morfológica.

A partir desses resultados, apresentados pelos autores, pode-se destacar a relação entre os valores obtidos na classificação das espécies estudadas nesse trabalho e nos resultados adquiridos pela classificação da Lógica Fuzzy kNN.

5 AMBIENTE DE IMPLEMENTAÇÃO

O algoritmo FNN4Bees foi desenvolvido e implementado na linguagem de programação orientada a objetos **Java**.

A modelagem e implementação desse algoritmo foi realizada na interface de desenvolvimento (IDE) Netbeans versão 6.1.

Os resultados são apresentados nas linhas do console da IDE.

A inserção dos dados no algoritmo é realizada na forma como eles são apresentados pelos biólogos após as transformações de alinhamento e padronização. Os dados inseridos são transformados utilizando bibliotecas de manipulação de arquivos da linguagem Java. O processo de transformação dos dados auxilia no método de classificação e análise dos dados, uma vez que os softwares desenvolvidos para este propósito, como o *Weka*, *Matlab*, *Octave*, entre outros, têm seus formatos próprios para a entrada de dados, e, são diferentes dos formatos TPS e NTS, provenientes dos softwares *tpsUtil*, *tpsDig2* e *tpsRelw*.

O diagrama de classe do algoritmo FNN4Bees é apresentado na Figura 38.

Figura 38 - Diagrama de Classes da implementação do algoritmo FNN4Bees

Na Figura 38, a classe **Main** implementa os testes e chamadas aos métodos das demais classes; a classe **ManipuladorArquivos** realiza todo o pré-processamento dos arquivos que formarão o conjunto de dados para os parâmetros de entrada dos métodos implementados nas demais classes; a classe **SelecionadorDeCaracteristicas** realiza os procedimentos de seleção e extração de características apresentadas pelos dados obtidos pelo **ManipuladorArquivos** e a classe **FNN4Bees** é a classe que implementa os principais métodos que envolvem o algoritmo kNN estendido, o cálculo da Ordem de Significância dos marcos matemáticos, as Análises Estatísticas aplicadas ao resultado da Ordem de Significância e a Fuzzy kNN.

5.1 FUNCIONAMENTO DO ALGORITMO FNN4BEES NO DIAGRAMA DE ATIVIDADES

A Figura 39 apresenta um Diagrama de Classes que mostra, com maiores detalhes, os métodos e as generalizações da classe **FNN4Bees**.

Figura 39 - Diagrama de Classes do FNN4Bees detalhado

O diagrama mostra as três classes responsáveis pela classificação direta de uma espécie de abelha indefinida. As classes **KNN** e **FNN** são especializações da classe **FNN4Bees**, que serve como uma classe abstrata às duas classes. A classe KNN é responsável pelo cálculo da distância euclidiana implementado no método *calculaDistEuclidiana*.

Esse método recebe como parâmetros dois vetores que contém os marcos anatômicos da espécie padrão (que é uma das espécies do conjunto de dados) e os marcos anatômicos da espécie indefinida (que é a espécie que ainda não foi classificada em seu respectivo grupo).

O método *classificaEspecie(especieIndefinida:Array)* é executado para cada espécie do conjunto de dados, assim, todas as espécies do conjunto serão testadas no processo de classificação e a que obtiver menor distância, no retorno do método *calculaDistEuclidiana* e maior grau de pertinência no retorno do método *calculaPertinencia*, será a espécie escolhida.

A classe FNN é responsável por implementar os métodos para calcular o grau de pertinência da Fuzzy kNN e calcular o espaçamento da função Fuzzy para cada espécie, dada a sua distância média, obtida no método *calculaDistEuclidiana* da classe KNN.

O diagrama de classe da Figura 39 mostra apenas as informações mais importantes dos métodos das classes. Da mesma forma, a classe ManipuladorArquivos, apresentada na Figura 38 é detalhada no Diagrama de Classes da Figura 40.

Figura 40 - Diagrama de classes na manipulação de arquivos
texto do algoritmo FNN4Bees

A classe **ManipuladorArquivos** da Figura 40 implementa o método *arrumaArquivo* que, ao receber como parâmetro um arquivo TPS com dados do tipo *String*, rearranja esse arquivo, retirando os dados que não serão usados no processo de classificação e, deixando apenas os dados referentes aos marcos anatômicos. Esse método chama o método da classe **ExpressaoRegular** para efetuar esta operação. O método *aplicaExpressaoRegular* da classe **ExpressaoRegular** utiliza a biblioteca *Regex* do Java para implementar os métodos das classes **Pattern** e **Matcher**, utilizando a expressão regular dada pelo padrão: `"-[0-9.]*"` para obter somente os dados referentes aos marcos anatômicos no arquivo de dados.

5.1.1 Descrição do método *aplicaExpressaoRegular*

O pseudocódigo do método *aplicaExpressaoRegular* é descrito como a seguir.

```

texto = null;
padrao = "-[0-9.].*"
temp = null
Pattern p = Pattern.compile(padrao);
Matcher m = p.matcher(texto);
while (m.find()) {
    val = m.group();
}

```

No código, a variável *val* armazena os dados que combinam com o padrão, que são os pontos bidimensionais representativos dos marcos anatômicos. Esses dados são utilizados nas classes KNN e FNN, no processo de classificação das espécies.

Por último, um detalhamento na classe *SelecionadorDeCaracteristicas*, apresentada na Figura 41, mostra o principal método implementado para efetuar a busca de padrões nos marcos anatômicos, no cálculo da Ordem de Significância dos marcos anatômicos mais significativos.

SelecionadorDeCaracteristicas
- nomeDoArquivo : String - LmSignificativos : Arrays
+ buscaPadraoLandmarks(distLM_kNN : KNN) : Arrays + calculaOrdemDeSignificancia(distLM : KNN) : Arrays

Figura 41 - Detalhamento da Classe
SelecionadorDeCaracteristicas

O método *buscaPadraoLandmarks* efetua o processo de busca pelos marcos anatômicos que melhor representam a espécie. Essa busca consiste em percorrer o

vetor que contém as distâncias entre os marcos anatômicos das duas espécies em análise (uma espécie do conjunto de dados e espécie de teste) e armazenar os marcos anatômicos correspondentes às distâncias que melhor representam aquela espécie em seu respectivo grupo, ou seja, as menores distâncias. Um limite é estipulado neste método, que corresponde a quantidade de marcos anatômicos que poderão fazer parte da Ordem de Significância. O método *calculaOrdemDeSignificancia* é responsável por determinar quais marcos anatômicos melhor representam a espécie entre os grupos (espécie de teste padrão e todas as espécies do conjunto de dados) e utiliza as distâncias entre os marcos anatômicos considerando os marcos anatômicos selecionados no método *buscaPadraoLandmarks*. Esse método permite a otimização dos marcos anatômicos que representam o padrão da espécie. Nele os marcos são novamente selecionados de acordo com a classificação realizada no algoritmo kNN, portanto, a classe KNN é chamada nesta classe para que, na fase de teste ou quando a classificação não for a esperada, os marcos anatômicos da espécie de teste padrão que confundem a classificação, tornando-a incorreta, sejam eliminados passo a passo. Assim, os marcos anatômicos que sobram são os que melhor representam aquela espécie. Os pseudocódigos desses métodos podem ser encontrados nas listagem 2 e listagem 3.

6 CLASSIFICAÇÃO E RESULTADOS

Os resultados dos testes de classificação foram obtidos por meio de treinamento e testes realizados com 36 espécies de abelhas sem ferrão (*meliponini*) descritas na listagem 1 e cinco espécies do grupo das *Euglossas*, separadamente. Os treinamentos e testes foram realizados até o critério de parada ser alcançado entre cada amostra da espécie com as espécies do conjunto. Foram testadas dez amostras rotuladas para cada espécie, modificando, quando necessário, a Ordem de Significância até o final do processo de otimização.

O algoritmo FNN4Bees apresentou uma taxa de acerto acima de 97% na classificação entre as espécies. Os testes foram realizados aplicando a otimização dos marcos anatômicos no processo de busca pelos melhores marcos anatômicos caracterizadores de cada espécie.

A quantidade de marcos anatômicos utilizados para a caracterização da espécie, no processo de refinamento e otimização, foi estabelecido como sendo o número de marcos anatômicos mínimo encontrado, sendo que este mínimo não pode ser menor que três.

Esse limite foi estabelecido para que o número de marcos anatômicos na Ordem de Significância não ficasse muito baixo, pois, em certos casos, para se chegar ao resultado correto, dever-se-ia considerar apenas um marco anatômico mais significativo.

Neste caso, a espécie não pode ser corretamente classificada pelo algoritmo kNN. Esse limite de três marcos anatômicos mais significativos foi estipulado de acordo com a média do mínimo número de marcos anatômicos necessários para a classificação das espécies no processo de treinamento e testes. Entretanto, de acordo com o grau de conhecimento da espécie analisada e dos respectivos marcos anatômicos mais significativos, o algoritmo permite maior refinamento, sendo este valor passível de mudança.

A porcentagem de acerto para cada teste das amostras das espécies está descrita a seguir.

Primeiro teste: taxa de acerto foi de **52%**, nas quais as espécies que deram resultados incorretos no algoritmo FNN4Bees foram: 4 (7), 5 (7), 8 (7), 9 (7), 10 (5), 16 (6), 17 (12), 18 (12), 20 (3), 22 (2), 24 (21), 26 (3), 29 (6), 32 (6), 33 (6), 34 (6) e 35 (5), sendo que, entre parênteses estão as espécies que foram classificadas no lugar das espécies corretas. Fora dos parênteses, estão as espécies que deveriam ter sido classificadas. Lembrando que a espécie 0 corresponde a espécie *Friesella_schrottkyi*, a espécie 1 corresponde a espécie *Frieseomellita_varia*, e assim por diante. O segundo teste foi executado após correções e acertos (fase de treinamento) nos marcos matemáticos mais significativos das espécies 7 e 6. O resultado da taxa de acerto do segundo teste mostra melhoras na classificação.

Segundo teste: a taxa de acerto foi de **64%**, nas quais as espécies que deram resultado incorreto no algoritmo FNN4Bees foram: 6 (2), 7 (5), 10 (5), 12 (11), 16 (3), 18 (3), 20 (3), 22 (2), 24 (21), 26 (3), 27 (3), 32 (6) e 35 (5).

O terceiro teste foi executado após correções e acertos nos marcos matemáticos mais significativos da espécie 3.

Terceiro teste: a taxa de acerto foi de **69%**, nas quais as espécies que deram resultado incorreto no algoritmo FNN4Bees foram: 6 (2), 7 (5), 10 (5), 12 (11), 18 (17), 20 (3), 22 (2), 24 (21), 26 (3), 32 (5) e 35 (5). O quarto teste foi executado após correções e acertos nos marcos matemáticos mais significativos das espécies 5 e 11.

Quarto teste: a taxa de acerto foi de **75%**, nas quais as espécies que deram resultado incorreto no algoritmo FNN4Bees foram: 3 (20), 4 (7), 6 (2), 16 (20), 18 (12), 22 (2), 24 (21), 32 (6) e 35 (5). O quinto teste foi executado após correções e acertos nos marcos matemáticos mais significativos da espécie 20.

Quinto teste: a taxa de acerto foi de **78%**, nas quais as espécies que deram resultado incorreto no algoritmo FNN4Bees foram: 6 (12), 7 (5), 10 (5), 18 (12), 20 (16), 21 (24), 33 (32) e 35 (5). O sexto teste foi executado após correções e acertos nos marcos matemáticos mais significativos das espécies: 5, 16 e 24.

Sexto teste: a taxa de acerto foi de **88%**, sendo que as espécies que deram resultado incorreto no algoritmo FNN4Bees foram: 6(12), 18 (12), 33 (32) e 35 (5).

Sétimo teste: a taxa de acerto foi de **91%**, sendo que as espécies que deram resultado incorreto no algoritmo FNN4Bees foram: 6(12), 18 (12) e 33 (32).

Oitavo teste: a taxa de acerto foi de **94%**, sendo que as espécies que deram resultado incorreto no algoritmo FNN4Bees foram: 6(12) e 18 (12).

Nono teste: a taxa de acerto foi de **97,2%**, sendo que a espécie que deu resultado incorreto no algoritmo FNN4Bees foi: 12(18).

Mais ajustes não puderam ser efetuados pois, a medida que se tentava acertar a espécie 12, errava-se a espécie 18 e vice versa. Portanto, conclui-se que esse é o resultado ótimo da classificação e outras mudanças nos marcos anatômicos mais significativos não são mais realizadas. Os marcos matemáticos mais significativos encontrados para cada espécie nesse processo definem o padrão que foi encontrado para cada espécie.

O resultado apresentado pela Fuzzy kNN, na simulação com o conjunto de dados das *Euglossas* foi:

Primeiro teste: a taxa de acerto foi de **100%**.

A Tabela 32 mostra os resultados apresentados nos testes de 1 a 9 das espécies de abelhas sem ferrão e o teste realizado entre as *Euglossas*.

Tabela 32 - Resultados dos testes realizados com o algoritmo FNN4Bees.

nº do Teste	Espécies Incorretas	taxa de acerto
teste 1	4 (7), 5 (7), 8 (7), 9 (7), 10 (5), 16 (6), 17 (12), 18	52%
	(12), 20 (3), 22 (2), 24 (21), 26 (3), 29 (6), 32 (6),	
	33 (6), 34 (6) e 35 (5)	
teste 2	6 (2), 7 (5), 10 (5), 12 (11), 16 (3), 18 (3), 20 (3),	64%

	22 (2), 24 (21), 26 (3), 27 (3), 32 (6) e 35 (5)	
teste 3	6 (2), 7 (5), 10 (5), 12 (11), 18 (17), 20 (3), 22 (2), 24 (21), 26 (3), 32 (5) e 35 (5)	69%
teste 4	3 (20), 4 (7), 6 (2), 16 (20), 18 (12), 22 (2), 24 (21), 32 (6) e 35 (5)	75%
teste 5	6 (12), 7 (5), 10 (5), 18 (12), 20 (16), 21 (24), 33 (32) e 35 (5)	78%
teste 6	6(12), 18 (12), 33 (32) e 35 (5)	88%
teste 7	6(12), 18 (12) e 33 (32)	91%
teste 8	6(12) e 18 (12)	94%
teste 9	12(18)	97,2%
teste das <i>Euglossas</i>	Nenhuma	100%

Pelos resultados dos testes na primeira fase de classificação no algoritmo FNN4Bees, a medida que mudanças foram sendo realizadas nos marcos matemáticos mais significativos, a classificação foi aperfeiçoando o resultado da taxa de acerto. É certo que, com a inserção de novos dados no conjunto, esse processo deverá se repetir, para que se possa buscar sempre o resultado ótimo.

6.1 ANÁLISE DOS RESULTADOS DA FUZZY KNN

O primeiro teste da execução do método que implementa a Fuzzy kNN no algoritmo FNN4Bees apresentou uma melhora no resultado da execução do algoritmo (em cerca de 90% na taxa de acerto para a espécie 0). O cálculo da classificação determinado pelos graus de pertinência atribuídos a cada variável, e pela função de pertinência aplicados às amostras, representam uma extensão ao processo realizado pelo algoritmo kNN, sendo este modificado. A Fuzzy kNN permitiu avaliar com maior precisão os dados que são recebidos a partir do algoritmo implementado FNN4Bees, realocando amostras indefinidas que, eventualmente, foram classificadas erroneamente pelo FNN4Bees.

6.2 ANÁLISE DO ALGORITMO EM COMPARAÇÃO COM OS RESULTADOS OBTIDOS PELAS CLASSIFICAÇÕES ATUAIS

De acordo com Franco (2007), diversos testes foram realizados com espécies de abelhas para analisar a variabilidade genético-morfológica em populações neotropicais de *Apis mellifera*. Entre os métodos utilizados neste processo, o autor realizou pesquisas por meio da Morfometria Tradicional, Morfometria Geométrica, J-Abis e análises do DNA mitocondrial das abelhas. Os resultados concernentes à aplicação de Morfometria Geométrica relatam que,

A Análise Discriminante foi capaz de classificar 43,5% dos indivíduos dentro de seu respectivo grupo e 36,5% foram corretamente identificados nos testes de validação cruzada. Estes valores foram obtidos dividindo-se o número de classificações e identificações corretas pelo número total de indivíduos. Merecem destaque os grupos de Ribeirão Preto – 1968 com 78,5% na classificação e 85,6% na identificação, *A. m. mellifera* com 96% tanto na classificação quanto na identificação, (FRANCO, 2007).

A validação cruzada, às vezes chamada de estimativa de rotação, é uma técnica utilizada para avaliar os resultados generalizados que a análise estatística irá prover para um conjunto de dados independentes. É utilizado principalmente em ambientes nos quais o objetivo é previsão, e se quer estimar quanto exatamente um modelo preditivo irá conseguir obter de acerto na prática.

Uma rodada de validação cruzada

envolve a separação de uma amostra de dados em subconjuntos complementares, realizar a análise de um subconjunto (chamado o conjunto de treinamento), e validar a análise sobre o outro subconjunto (chamado o conjunto de validação ou conjunto de teste).

Para reduzir a variabilidade, várias rodadas de validação cruzada são realizadas utilizando partições diferentes, e os resultados da validação são calculados sobre os

Tabela 36 - Resultado da porcentagem de acerto (Espécie versus Espécie) dos testes de validação cruzada das Espécies entre 27 a 35 do conjunto de dados.

Espécie/ Espécie	27	28	29	30	31	32	33	34	35
27	1.0								
28		1.0							
29			1.0						
30				0.9994446					
31					0.9969459				
32						0.9998441			
33							0.9998410		
34								0.9991815	
35									1.0

Nas Tabelas, os valores correspondem ao grau de pertinência do espécime correlacionado a espécie ao qual ele pertence. As Tabelas mostram o resultado da defuzzyficação da Fuzzy kNN nos dados do conjunto de treinamento. Para esses dados foi executado o algoritmo completo FNN4Bees (primeira fase, de seleção e extração de características, segunda fase de implementação do algoritmo kNN e terceira fase da implementação da Fuzzy kNN). A finalidade do algoritmo é a classificação entre as espécies, e é esta análise que está sendo considerada nas discussões dos resultados finais.

7 CONCLUSÕES

Uma proposta para a solução do problema de classificação de espécies de abelhas foi estudada a partir da implementação do algoritmo FNN4Bees com a implementação da Fuzzy kNN.

O algoritmo, que tem como base a aplicação das técnicas de Morfometria Geométrica e o estudo das *Formas* dos marcos anatômicos das imagens geradas pelas asas das abelhas, proporcionou um significativo aumento da taxa de acerto dentro e fora das espécies (taxa de acerto de 90% em ambos os casos).

A partir dos resultados pode-se atestar a efetividade da técnica de Morfometria Geométrica nas asas anteriores para discriminação das espécies estudadas, uma vez que, a classificação utiliza arquivos que provêm da aplicação dessas técnicas.

O método proposto para a classificação de abelhas (FNN4Bees) pode ser considerado como um Sistema de Apoio a Decisão à classificação feita atualmente pelos biólogos especialistas, já que o sistema mostrou-se capaz de classificar boa parte das espécies do conjunto de dados.

A maior contribuição desse trabalho pode ser encontrada na descrição da Ordem de Significância e na implementação da lógica Fuzzy kNN. Para as espécies de abelhas sem ferrão, que não podem ser classificadas pelo ABIS, o sistema implementado no FNN4Bees serve como uma alternativa à classificação dessas espécies.

A implementação da Fuzzy kNN permitiu transformar os pontos obtidos pelos marcos anatômicos em informações relevantes para o processo de classificação, tornando-os representativos pelas Tabelas que agregam os valores de graus de pertinência para cada um dos marcos anatômicos mais significativos de cada espécie.

A tentativa de atender à uma classificação de abelhas com resultados melhorados ao processo atual foi alcançada com êxito na implementação do algoritmo FNN4Bees.

A partir dos resultados obtidos nos testes de classificação das abelhas do conjunto de dados e da análise de comparação desses resultados de testes com os

resultados obtidos pelos testes dos especialistas, pode-se averiguar que a implementação do algoritmo forneceu uma melhora no processo de classificação dos mesmos dados envolvidos entre os dois métodos, embora, na análise realizada pelos especialistas o objetivo principal era a classificação dentro das próprias espécies, correlacionadas.

A agregação de métodos computacionais no processo de classificação melhorou os resultados que antes provinham de métodos estatísticos, apenas.

A adequação das regras a partir do conhecimento de dados numéricos e simbólicos, que caracterizam o conhecimento do especialista, na Fuzzy kNN, permite refinar e aprimorar os dados de entrada do classificador melhorando o processo de treinamento e aprendizado do algoritmo.

Esse processo, inédito no atual processo desenvolvido pelos biólogos, otimiza o processo de alcançar o resultado esperado.

Em particular, aos resultados obtidos pela classificação da Fuzzy kNN, o principal fator de melhora no processo de classificação pode ser notado na criação da Ordem de Significância, que eliminou os marcos anatômicos que aumentavam a possibilidade do erro na classificação entre as espécies e dentro das espécies, na busca por padrões e na classificação.

A definição dos marcos anatômicos mais significativos criando os diversos padrões caracterizadores de cada espécie proporcionou uma classificação otimizada e diversos problemas que poderiam ocorrer com a utilização do algoritmo kNN e Fuzzy kNN foram eliminados, como o problema de dimensionalidade dos dados, ruídos, dados incompletos ou com informações faltantes, entre outros.

Trabalhos futuros podem ser desenvolvidos no sentido de utilizar outras lógicas computacionais na comparação dos resultados apresentados pela classificação com os algoritmos kNN e Fuzzy kNN.

Pode-se tentar aprimorar os resultados obtidos neste trabalho com a inserção de outras espécies, ou mesmo as mesmas apresentadas, mas que possuam a colocação dos marcos anatômicos nas imagens das asas efetuada da melhor maneira possível.

Outros algoritmos podem ser utilizados no lugar do kNN como o k-Means, e a distância pode ser calculada utilizando-se a distância de Mahalanobis, ao invés da

distância euclidiana entre os marcos anatômicos. Outros algoritmos, como o SVM (Support Vector Machines) e as redes neurais artificiais podem ser testados para a classificação.

REFERÊNCIAS

ADAMS, D. C. et al. **Geometric morphometrics: ten years of progress following the “revolution”**. Italian. Journal of Zoology, 2004, Vol. 71, 5 -16p.

BERGAMINI, C. M.; ARAUJO, P. V.; MOTTER, G. **Modelos de Aprendizagem de Máquina na Classificação de Caracteres Manuscritos**. Synergismus scyentifica, UTFPR, Pato Branco, 01, (1, 2, 3, 4). 1-775, 2006.

BEYER, K. et al. **When is Nearest Neighbor Meaningful?** Lecture Notes In Computer Science, archive Proceedings of the 7th International Conference on Database Theory, 1999, Vol. 1540, 217 - 235p.

BIRD, A. **Perceptions of epigenetics**. Nature 447, 24 May 2007, 396 - 398p.

BONDUGULA, R. et al. **Profiles And Fuzzy K-Nearest Neighbor Algorithm For Protein Secondary Structure Prediction**. Proceedings of APBC, The Third Asia-Pacific Bioinformatics Conference, Singapore, January, 17 - 21, 2005.

BOOKSTEIN, F. L. **Principal Warps: Thin-Plate Splines and the Decomposition of Deformations**. IEEE Transactions On Pattern Analysis And Machine Intelligence. Vol. 11. No. 6. June, 1989.

BOOKSTEIN, F. (1996). **Biometrics, Biomathematics and the Morphometric Synthesis**. *Bulletin of Mathematical Biology*, 58(2):313-365.

BOOKSTEIN, F.L. **Singularities and the Features of Deformation Grids**. In Vemuri, B., editor, *Workshop on Biomedical Image Analysis*, pages 46-55, Santa Barbara, California (USA). IEEE Computer Society Press. 1998.

BUENO, J. F. **Modelo de Arquitetura de Referência para o Sistema de Identificação Automática e Classificação Taxonômica de Abelhas sem Ferrão usando a Morfometria das Asas – Uma abordagem de Reconhecimento de Padrões**. Projeto de Pesquisa apresentado como documento para qualificação de doutorado à Escola Politécnica da Universidade de São Paulo em 17 de novembro de 2010.

CAVALCANTI, M. J. et al. **Landmarks-based Morphometric Analysis in Selected Species of Serranid Fishes (Perciformes: Teleostei)** Zoological Studies, 1999.

COLLET, T. et al. **Genetic structure of Africanized honeybee populations (*Apis mellifera* L.) from Brazil and Uruguay viewed through mitochondrial DNA COI–COII patterns.** *Heredity* (2006) 97, pp. 329–335, published online 6 September 2006.

COST, S.; SALZBERG, S. **A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features.** Machine Learning Journal, Springer Netherlands Publisher, January 31, 1993.

COVER, T. M.; HART, P. E. **Nearest Neighbor Pattern Classification.** IEEE transactions on information theory, VOL. IT-13, NO. 1, January, 1967.

DOS REIS, S. F. **Geographic Variation In Cranial Morphology In *Thrichomys apereoides* (Rodentia: Echimyidae).** I Geometric Descriptors And Patterns Of Variation In Shape Journal of Mammalogy, 2002, 333–344p.

DRAUSCHKE, M., et al. **Reliable Biometrical Analysis in Biodiversity Information Systems.** Pp. 25-36. *In:* Proceedings of the 7th International Workshop on Pattern Recognition in Information Systems, Funchal, Portugal. 2007.

DRYDEN, I. L., MARDIA, K. V. **Statistical Shape Analysis**, John Wiley, Chichester in July 1998.

FRANCOY, T. M; et al. **Patterns of wing venation corroborates populational differences in *Plebeia remota*.** *Apidologie Celle*, 2006a, v. 37, 613-614p.

FRANCOY, T. M. et al. **Morphometric differences in a single wing cell can discriminate *Apis mellifera* racial types.** *Apidologie* 37, EDP Sciences, 2006b.

FRANCOY, T. M. **Variabilidade genético-morfológica em populações Neotropicais de *A. mellifera*.** 163f. Tese (Doutorado em genética) – Curso de Pós-graduação em Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2007.

FRANCOY, T. M; et al. **A Morfometria Geométrica das asas e a identificação dos Meliponini.** VIII Encontro sobre abelhas de Ribeirão Preto, 2008, Ribeirão

Preto. Biodiversidade e uso sustentado de abelhas. Ribeirão Preto: FUNPEC editora, 2008a. 252 - 254p.

FRANCOY, M. T. et al. **Identification of Africanized honey bees through wing morphometrics: two fast and efficient procedures**. Apidologie 39 (2008) 488–494, INRA/DIB-AGIB/ EDP Sciences, 2008b.

FRANCOY, T. M. **Conjunto de dados contendo imagens de asas de abelhas e seus Marcos Anatômicos**. Disponível somente em contato direto com o autor. Adquirido em 20 de agosto de 2009.

GADER, P. et al. **Comparison of Crisp and Fuzzy Character Neural Networks in Handwritten Word Recognition**. IEEE Transactions On Fuzzy Systems, Vol. 3, No. 3, August 1995.

GADER, P. D.; KELLER, M. **Applications of Fuzzy Set Theory to Handwriting Recognition**. IEEE, 1994.

GESÙ, V.; BOSCO, G. **Combining One Class Fuzzy KNN's**. Springer-Verlag Berlin Heidelberg. 2007, 152 - 160p.

HEPBURN, H., R. et al. **Apis florea: morphometrics, classification and biogeography**. Apidologie 36 (2005) 359–376, INRA/DIB-AGIB/ EDP Sciences, 2005.

JIRAKANJANAKIT, N.; DUJARDIN, J. P. **Discrimination of Aedes Aegypti (Diptera: Culicidae) Laboratory Lines based on Wing Geometry**. Institute of Science and Technology for Research and Development, Mahidol University, Salaya, Nakhon Pathom, Thailand, 2005.

KELLER, J. M., GRAY, M. R.; GIVENS, J. A. Jr. **A fuzzy K-Nearest Neighbor Algorithm**. IEEE Trans. on SMC, Volume SMC-15, No.4, 1985.

KENDALL, D. G. **A Survey of the Statistical Theory of Shape**. *Statistical Science*, 4(2):87-120. 1989.

KENDALL, D. G. et al. **Shape and shape theory**. Wiley Series in Probability and Statistics, Book details: 318 pages, Wiley; 1 edition (October 25, 1999) ISBN-10: 0471968234, 1999.

KUMAR, D. et al. **Characterizing Shape Differences between Phantom Image Populations via Multivariate Statistical Analysis of Inverse Consistent Transformations**. Springer Berlin / Heidelberg, Volume 2717/2003, Biomedical Image Registration, pp. 367-376, November 04, 2003.

MALLISON, H.; BENTLEY, P. **Envolving Fuzzy Rules for Pattern Classification**. Computational Intelligence for Modelling, Control & Automation, M. Mohammadian (Ed.), IOS Press, 1999.

MATIAS, A. et al. **Rhodnius in Bolivia Identified by its Wings**. Memmorial Institute of Oswaldo Cruz, Rio de Janeiro, October 2001, 947 - 950p.

MCKELVEY, B. **Organizational systematics--taxonomy, evolution, classification**. University of California Press, 1982, pp. 30-31.

MITRA, S.; PAL, S. K. **Fuzzy sets in pattern recognition and machine intelligence**. Fuzzy Sets and Systems. Elsevier, 2005, 381 - 386p.

MOUSSEAU, T. A. **Landmarks in Morphometrics, or the Shape and Size of Morphometrics to Come**. Evolution, Vol. 45, N° 8. December 1991.

MONTEIRO, L.R.; REIS S. F. **Princípios de Morfometria Geométrica**. Ribeirão Preto, Holos Editora Ltda, 1999, 188p.

PEDRYCZ, W.; GOMIDE, F. **Fuzzy Systems Engineering Toward Human-Centric Computing**, Fuzzy Systems Engineering. John Wiley & Sons, 2007.

RAMÍREZ, S. R. et al. **Phylogeny, diversification patterns and historical biogeography of euglossine orchid bees (Hymenoptera: Apidae)**. Biological Journal of the Linnean Society, 2010, 100, 552–572. Received 13 October 2009; revised 5 January 2010; accepted for publication 6 January 2010.

RHEE, F. C. H.; HWANG, C. **An Interval Type-2 Fuzzy K-Nearest Neighbor**. Fuzzy Systems, vol.2, 2003. 802 - 807p.

RICHTSMEIER J. T. et al. **The promise of geometric morphometrics**. American Journal of Physical Anthropology, Vol. 119, No. S35. 2002, 63 - 91p.

RODRIGUES, L. A.; SANTOS, V. F. **Morfometria Geométrica - Aplicações em paleobiologia de Dinossauros**. Ciências da Terra, Lisboa, 2004.

ROHLF, F. J. **Shape statistics: Procrustes superimpositions and tangent spaces**. Journal of Classification 16. 1999, 197 - 223p.

ROHLF, F. J. **Geometric morphometrics in systematics**. Published as a chapter in N. Macleod and P. Forey (eds.) *Morphology, shape and phylogenetics*. Taylor&Francis: London. 2002.

ROSS, H. H. **Principles of Numerical Taxonomy**. Systematic Zoology, Vol. 13, No. 2, Jun., 1964, 106 - 108p.

ROSS, A. **Procrustes Analysis**. Department of Computer Science and Engineering University of South Carolina, 2005.

SALGADO-NETO, G.; DI MARE, R. A. **Morfometria e mortalidade em pupas e adultos de *Brassolis astyra* (Nymphalidae, Morphinae)**. Iheringia, Série Zoolgia. Vol.99, no.2, Porto Alegre, Junho, 2009.

SCHRÖDER, S., et al. **An Automated Method for the Identification of Bee Species (Hymenoptera: Apoidea)**. In International Symposium on Conserving Europe' s Bees, 1995, London. Proceedings. London: International Bee Research Association & Linnean Society, 1995. p. 6-7.

SILVEIRA, F.A., et al. **Taxonomic constraints for the conservation and sustainable use of wild pollinators: the Brazilian wild bees**. In: Pollinating bees: the conservation link between Agriculture and Nature, P.Kevan & V.L. Imperatriz-Fonseca (eds.). Ministério do Meio Ambiente, Brasília, p.41-50, 2002.

SILVEIRA, F. V. J. **Testes de Permutação e Bootstrap em Análise Estatística de Formas: Aplicações à Zoologia**. Dissertação apresentada em fev/2008, Univerdidade Federal de Pernanbuco, CCEN. Departamento de Estatística. 2008.

SLICE, D. E. **Landmarks aligned by Procrustes analysis do not lie in Kendall's shape space.** Systematic Biology 50(1). 2001, 141 - 149p.

SLICE, D. E. **A survey of the Statistical Theory of Shape.** Rev. Anthropology. 2007, 261p.

SOLE et al. **Intra-specific patterns of mitochondrial variation in natural population fragments of a localised desert dung beetle species, *Pachysoma gariepinum* (Coleoptera: Scarabaeidae).** Journal of Heredity. 99(5):464-475, 2008.

SOUZA, S. A. de O. **Alguns comentários sobre a teoria fuzzy.** Revista Exacta, vol1, 2003, 139-148 pp.

TOFILSKI, A. **Using geometric morphometrics and standard morphometry to discriminate three honeybee subspecies.** Apidologie 39 (2008), pp. 558–563, INRA/DIB-AGIB/ EDP Sciences, 2008.

WENG, F. et al. **Clustering Ensemble based on the Fuzzy KNN Algorithm.** Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. 2007

WHITBECK, M., GUO, H. **Multiple landmark warping using thin-plate splines.** In: IPCV. (2006) 256–263.

WU, D. et al. **Applications of Fuzzy Logic in Bioinformatics.** University of Missouri-Columbia, USA. Imperial College Press, Series on Advances in Bioinformatics and Computational Biology, Vol. 9, 2008.

ZADEH, L. A. **Fuzzy Sets.** Information and control, vol. 8, pp. 338-353. 1965.

ZHU, H.; BASIR, O. **An Adaptive Fuzzy Evidential Nearest Neighbor Formulation for Classifying Remote Sensing Images.** IEEE Transactions On Geoscience And Remote Sensing, Vol. 43, No. 8, August 2005.