

UNIOESTE – Universidade Estadual do Oeste do Paraná

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

Colegiado de Ciência da Computação

Curso de Bacharelado em Ciência da Computação

**Desenvolvimento de uma Máquina de Comitê Estática
para a tarefa de Classificação na Ferramenta YADMT**

Tiago Alexandre Schulz Sippert

CASCABEL

2012

TIAGO ALEXANDRE SCHULZ SIPPERT

**DESENVOLVIMENTO DE UMA MÁQUINA DE COMITÊ ESTÁTICA
PARA A TAREFA DE CLASSIFICAÇÃO NA FERRAMENTA YADMT**

Monografia apresentada como requisito parcial
para obtenção do grau de Bacharel em Ciência
da Computação, do Centro de Ciências Exatas
e Tecnológicas da Universidade Estadual do
Oeste do Paraná - Campus de Cascavel

Orientador: Prof. Dr. Clodis Boscaroli

CASCADEL

2012

TIAGO ALEXANDRE SCHULZ SIPPERT

**DESENVOLVIMENTO DE UMA MÁQUINA DE COMITÊ ESTÁTICA
PARA A TAREFA DE CLASSIFICAÇÃO NA FERRAMENTA YADMT**

Monografia apresentada como requisito parcial para obtenção do Título de *Bacharel em Ciência da Computação*, pela Universidade Estadual do Oeste do Paraná, Campus de Cascavel, aprovada pela Comissão formada pelos professores:

Prof. Dr. Clodis Boscaroli (Orientador)
Colegiado de Ciência da Computação,
UNIOESTE

Prof. Dr. Jerry Adriani Johann
Centro de Ciências Exatas e Tecnológicas,
UNIOESTE

Prof. Esp. Gustavo Rezende Krüger
Colegiado de Tecnologia em Análise e
Desenvolvimento de Sistemas, UNIVEL

Cascavel, 30 de outubro de 2012.

DEDICATÓRIA

A meus familiares, meus professores e a todos que acreditaram que hoje eu estaria comemorando o final desta luta, e início de muitas outras.

AGRADECIMENTOS

Agradeço primeiramente a Deus pela dádiva da vida, por cada novo dia que me dá; ao meu pai e minha mãe, pela minha criação e pelos exemplos de vida que são; ao meu irmão, pelo companheirismo e força; aos meus amigos, pelos incentivos dados durante toda essa jornada.

Agradeço também aos meus professores, que me ensinaram muito, e dentre eles em especial ao meu orientador, que esteve sempre disposto a me guiar na jornada acadêmica. Ainda agradeço aos colegas de turma e de trabalho, com os quais pude aprender e compartilhar conhecimento.

Agradeço a todos que de uma ou outra forma tornaram este trabalho possível.

Lista de Figuras

2.1	Exemplo de Máquina de Comitê Estática	8
2.2	Exemplo de Máquina de Comitê Dinâmica	11
3.1	Tela da ferramenta YADMT – Etapa de classificação	17
3.2	Tela da ferramenta YADMT – Configuração da Máquina de Comitê	18
3.4	Tela da ferramenta YADMT – Resultados da execução da Máquina de Comitê	19

Lista de Tabelas

4.1	Bases de dados escolhidas para comparação das Técnicas	20
4.2	Aplicações da Análise de Imagens	22
4.3	Resultados dos Testes para a Base de Dados <i>Iris</i>	25
4.4	Resultados dos Testes para a Base de Dados <i>Acute Inflammation</i>	26
4.5	Resultados dos Testes para a Base de Dados <i>Teaching Assistant Evaluation</i>	27
4.6	Resultados dos Testes para a Base de Dados <i>Ecoli</i>	28
4.7	Resultados dos Testes para a Base de Dados <i>Abalone</i>	29
4.8	Resultados dos Testes para a Base de Dados <i>Spambase</i>	30
4.9	Resultados dos Testes para a Base de Dados <i>LIBRAS movement</i>	31
4.10	Resultados dos Testes para a Base de Dados <i>Statlog (Image Segmentation)</i>	32
4.11	Resultados dos Testes para a Base de Dados <i>Statlog (Landsat Satelite)</i>	33
4.12	Resultados dos Testes para a Base de Dados <i>Wine Quality</i>	34
4.13	Resumo dos Resultados dos Testes	35

Lista de Algoritmos

3.1	Processo de classificação por Votação Simples	15
3.2	Processo de classificação por Votação Ponderada	15

Lista de Abreviaturas e Siglas

DM	<i>Data Mining</i>
FN	<i>False Negative</i>
FP	<i>False Positive</i>
GIA	Grupo de Pesquisa em Inteligência Aplicada
KDD	<i>Knowledge Discovery in Databases</i>
K-NN	<i>K-Nearest Neighbors</i>
LVQ	<i>Learning Vector Quantization</i>
MLP	<i>Multilayer Perceptron</i>
P	Precisão
RBF	<i>Radial Basis Function</i>
RNA	Redes Neurais Artificiais
S	Taxa de Renovação
SGBD	Sistema Gerenciador de Banco de Dados
TA	Taxa de Acerto
TN	<i>True Negative</i>
TP	<i>True Positive</i>
UCI	<i>University of California</i>
YADMT	<i>Yet Another Data Mining Tool</i>
WEKA	<i>Waikato Environment for Knowledge Analysis</i>

Sumário

Lista de Figuras	vi
Lista de Tabelas	vii
Lista de Algoritmos	viii
Lista de Abreviaturas e Siglas	ix
Sumário	x
Resumo	xiii
1 Introdução	1
1.1 Justificativas	2
1.2 Objetivos	3
1.3 Metodologia	3
1.4 Organização do Texto	4
2 Máquinas de Comitê	5
2.1 Motivações de Uso	6
2.2 Estruturas Estáticas	7
2.2.1 Média do Conjunto	9
2.2.2 Impulso	9
2.2.3 Votação Majoritária	10
2.3 Estruturas Dinâmicas	10
2.3.1 Mistura de Especialistas	11
2.3.2 Mistura de Hierárquica de Especialistas	12
2.4 Trabalhos Correlatos	12

3	Módulo de Máquina de Comitê Proposto	14
3.1	Algoritmos	14
3.1.1	Votação Majoritária Simples	14
3.1.2	Votação Majoritária Ponderada	15
3.2	Integração na YADMT	16
4	Avaliação experimental	20
4.1	Bases de Dados	20
4.2	Metodologia de Avaliação	21
4.3	Análise dos Resultados	23
4.3.1	Experimento com a Base de Dados <i>Iris</i>	24
4.3.2	Experimento com a Base de Dados <i>Acute Inflammations</i>	25
4.3.3	Experimento com a Base de Dados <i>Teaching Assistant Evaluation</i>	26
4.3.4	Experimento com a Base de Dados <i>Ecoli</i>	27
4.3.5	Experimento com a Base de Dados <i>Abalone</i>	28
4.3.6	Experimento com a Base de Dados <i>Spambase</i>	29
4.3.7	Experimento com a Base de Dados <i>LIBRAS movement</i>	30
4.3.8	Experimento com a Base de Dados <i>Statlog (Image Segmentation)</i>	31
4.3.9	Experimento com a Base de Dados <i>Statlog (Landsat Satelite)</i>	32
4.3.10	Experimento com a Base de Dados <i>Wine Quality</i>	33
4.4	Considerações Finais	34
5	Conclusões	36
A	Descrição das Bases de Dados	38
A.1	Base de Dados <i>Iris</i>	38
A.2	Base de Dados <i>Acute Inflammations</i>	38
A.3	Base de Dados <i>Teaching Assistant Evaluation</i>	39
A.4	Base de Dados <i>Ecoli</i>	40
A.5	Base de Dados <i>Abalone</i>	40

A.6	Base de Dados <i>Spambase</i>	42
A.7	Base de Dados <i>LIBRAS movement</i>	42
A.8	Base de Dados <i>Statlog (Image Segmentation)</i>	43
A.9	Base de Dados <i>Statlog (Landsat Satellite)</i>	44
A.10	Base de Dados <i>Wine Quality</i>	45
	Referências Bibliográficas	46

Resumo

Na tomada de decisão, a análise de dados vem se tornando uma necessidade, dada a constante evolução tecnológica e o aumento no volume de dados, que dificulta a extração de informações e de conhecimento úteis a partir de tais dados. No processo de mineração de dados, a partir de técnicas de classificação, é possível realizar a extração de conhecimento em grandes quantidades de dados. Apesar de existirem diversas técnicas para a classificação de dados, nem todas são adequadas para certos conjuntos de dados, assim como também não existe uma técnica que trate de maneira adequada todos os conjuntos de dados. Para a resolução deste problema, pode-se usar uma máquina de comitê, uma técnica de aprendizagem supervisionada baseada na abordagem da divisão e conquista descrita como um estimador universal. A combinação de vários especialistas, neste caso, classificadores, constitui uma máquina de comitê. Neste trabalho foi implementado um módulo de máquina de comitê para a ferramenta YADMT (*Yet Another Data Mining Tool*), que já possui vários algoritmos de classificação. A abordagem para combinação dos resultados dos classificadores escolhida foi votação majoritária. Foi feita uma avaliação experimental dos resultados em uma seleção de bases de dados para classificação do repositório de aprendizado da UCI (*University of California – Irvine Machine Learning Repository*). Como métricas de verificação para os testes experimentais foram utilizadas acuidade, precisão e taxa de renovação e a matriz de confusão. A avaliação demonstrou que o uso de máquina de comitê otimiza os resultados de classificação em comparação ao uso de somente um especialista, e torna a classificação mais confiável por se tratar de um comitê de vários especialistas.

Palavras-chave: Mineração de Dados, Redes Neurais Artificiais, Máquinas de Comitê.

Capítulo 1

Introdução

Na tomada de decisão, a análise de dados vem se tornando uma necessidade, pois a constante evolução tecnológica e a informatização dos mais variados processos organizacionais ocasionaram um aumento no volume de dados, dificultando a extração de conhecimento a partir de tais dados.

Este processo de extração de conhecimento é chamado de Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Database – KDD*), definido por (Fayad, *et al.*, 1996) como um processo não trivial, de várias etapas, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de bases de dados. Estas etapas são: seleção de dados, pré-processamento e limpeza, transformação, mineração de dados e interpretação.

A principal etapa de KDD é a Mineração de dados (*Data Mining – DM*), onde é feita a procura de padrões consistentes em grandes quantidades de dados por meio de algoritmos de aprendizagem para a extração de conhecimento (Fayad, Piatetsky-Shapiro e Smyth, 1996).

Aprendizagem de máquina é a área que trata do projeto e criação de modelos computacionais capazes de aprender e/ou gerarem conhecimento e habilidades encontradas no ser humano (Magalhães, 2007). Os algoritmos de aprendizagem podem ser classificados em dois modelos conforme a sua interação com o ambiente: aprendizagem supervisionada, e aprendizagem não supervisionada (Haykin, 2001).

Na aprendizagem supervisionada os algoritmos requerem um treinamento com dados já rotulados, ou seja, cada tupla de atributos de entrada está associada com sua saída esperada. Já na aprendizagem não supervisionada a saída para cada tupla de atributos de entrada é desconhecida, assim a aprendizagem ocorre por observação e descoberta, por meio da detecção de padrões pelas regularidades dos dados de entrada (Russel e Norvig, 2004).

Uma das tarefas de aprendizagem de máquina para a mineração de dados utilizada em larga escala é a de classificação, que consiste no processo de encontrar, através de aprendizado de máquina, um modelo ou função que descreva diferentes classes de dados. Este modelo encontrado é chamado de classificador, o qual prediz rótulos de classes de valores discretos e não ordenados (Han e Kamber, 2005).

Vários métodos de classificação foram propostos por pesquisadores nas áreas de aprendizado de máquina, reconhecimento de padrões e estatística. A classificação de dados têm inúmeras aplicações reais, a exemplo de detecção de *fraude*, marketing dirigido, previsão de desempenho, produção industrial, diagnóstico médico (Han, Kamber e Pei, 2012).

Apesar de existirem diversas técnicas para a classificação de dados, nem todas são adequadas para certos conjuntos de dados, assim como também, não existe uma técnica que trate de maneira adequada todos os conjuntos de dados (Bonifácio, 2010). Uma técnica com um alto desempenho para um domínio ou conjunto de dados específico, não necessariamente irá ter o mesmo desempenho para qualquer conjunto de dados.

A máquina de comitê é uma das técnicas de aprendizagem supervisionada baseada na abordagem da divisão e conquista. Nessa abordagem uma tarefa computacional complexa é dividida em tarefas computacionais simples, para que a partir da combinação da solução de cada tarefa simples seja resolvida a tarefa complexa.

Nessa aprendizagem supervisionada a tarefa de aprendizagem é distribuída em um número de especialistas. A combinação destes especialistas constitui uma máquina de comitê (Haykin, 2001). Um especialista é um algoritmo projetado para emular a especialização humana de um domínio específico (Flores, 2003). No âmbito deste trabalho, o especialista é um algoritmo classificador de dados.

1.1 Justificativas

A YADMT (*Yet Another Data Mining Tool*) é uma ferramenta de Descoberta de Conhecimento em Banco de Dados (KDD), alternativa a outras já existentes na área, em desenvolvimento na Unioeste, no âmbito do GIA – Grupo de Pesquisa em Inteligência Aplicada, no campus de Cascavel. Dentre suas principais características destacam-se que é uma ferramenta livre, que facilita evoluções, e não está fixada a um SGBD (Sistema Gerenciador de Banco de Dados) ou a um formato de entrada em específico, conforme afirmado em (Benfatti, *et al.*, 2010).

Mesmo em fase de desenvolvimento, já foram implementados vários métodos de classificação. No trabalho de Benfatti (2010) foram implementados os algoritmos K-NN (K-Nearest Neighbors), C4.5 (Gerador de Árvore de Decisão) e *Naive Bayes* (Classificador Bayesiano), e no trabalho de Bonifácio (2010) foram implementadas as RNA (Redes Neurais Artificiais) MLP (*Multilayer Perceptron*), RBF (*Radial Basis Function*) e LVQ (*Learning Vector Quantization*). A ferramenta contém ainda outros algoritmos de classificação, e como está em fase de desenvolvimento, novos algoritmos – inclusive para diferentes tarefas de mineração de dados, estão sendo constantemente implementados.

Em uma ferramenta como essa se justifica a inserção de um Módulo de Máquina de Comitê como mais uma alternativa ao usuário, e principalmente, como uma forma de extrair melhores resultados por meio da combinação dos diversos algoritmos de classificação. É esperado, como hipótese inicial, concluir que o uso de máquina de comitê, a partir da combinação de vários algoritmos otimiza os resultados alcançados em relação aos resultados dos algoritmos aplicados de forma individual

1.2 Objetivos

Este trabalho se propôs a realizar a implementação de um módulo de máquina de comitê de classificação na YADMT, de forma genérica, ou seja, um módulo em que o usuário da ferramenta possa criar uma máquina de comitê dinamicamente, escolhendo entre os métodos estáticos (votação majoritária e votação majoritária ponderada), bem como os classificadores e a quantidade destes que formarão esta máquina.

1.3 Metodologia

A metodologia adotada para o desenvolvimento da pesquisa incluiu uma revisão da literatura para compreensão em profundidade acerca do tema objeto de estudo, e a elaboração de um projeto para o módulo de máquina de comitê da ferramenta YADMT, bem como a implementação das técnicas de votação majoritária simples e ponderada na linguagem de programação Java, e a integração do módulo proposto na ferramenta.

Para a avaliação dos algoritmos e execução dos testes foi utilizada uma seleção de 10 bases de dados para classificação do repositório de aprendizado da UCI (*University of California* –

Irvine Machine Learning Repository) (Frank e Asuncion, 2010). Para fins comparativos foram utilizadas as mesmas bases de dados utilizadas por Benfatti (2010) e Bonifácio (2010).

Como métricas de verificação para os testes foram utilizadas acuidade (acurácia), precisão e taxa de renovação (*recall*), e alguns métodos de avaliação como a matriz de confusão.

1.4 Organização do texto

Além dessa Introdução, no Capítulo 2 é feita uma revisão bibliográfica sobre máquinas de comitê, suas categorizações, motivações de uso e citados alguns trabalhos correlatos, motivacionais à pesquisa.

No Capítulo 3 é apresentado o projeto e implementação dos algoritmos de votação de uma máquina de comitê estática, bem como a incorporação destes como um módulo da ferramenta YADMT.

O Capítulo 4 traz os métodos de avaliação e o domínio de aplicação usados para a avaliação experimental dos algoritmos, assim como os resultados alcançados.

O Capítulo 5, por sua vez, compreende as considerações finais e possíveis trabalhos futuros, e o Apêndice A contém a descrição detalhada das bases de dados utilizadas nos experimentos.

Capítulo 2

Máquinas de Comitê

Um sistema especialista, também denominado máquina de aprendizagem, é um algoritmo projetado para emular a especialização humana de um domínio específico. Este sistema irá possuir uma base de conhecimento formada por fatos, regras e heurísticas sobre o domínio, e devem ser capazes de fornecer sugestões e conselhos, além de adquirir novos conhecimentos (Flores, 2003).

O objetivo de um sistema especialista é bastante restrito, pois são concebidos de forma a reproduzir o comportamento dos especialistas humanos na resolução de problemas do mundo real, porém, o domínio dos problemas é altamente restrito (Barr e Feigenbaum, 1981).

A Máquina de Comitê é uma das técnicas de aprendizagem supervisionada baseada na abordagem da divisão e conquista, onde uma tarefa computacional complexa é resolvida pela divisão em diversas tarefas computacionais simples e combinação de seus resultados. Na aprendizagem supervisionada a simplificação consiste em dividir estas tarefas em um número de sistemas especialistas, a fim de combinar o resultado de cada um para obter uma resposta teoricamente superior à alcançada individualmente. A combinação destes especialistas constitui uma máquina de comitê, conforme visto em (Haykin, 2001).

A ideia geral de combinar estimadores para a melhora do desempenho final de um sistema de classificação, regressão ou predição já vem de mais tempo e já foi discutida por Ablow e Kaylor (1965). A intensificação da pesquisa na área só veio mais tarde, com o advento da evolução computacional e por meio de trabalhos importantes que retomaram o assunto na área, como Hansen e Salamon (1990) e Perrone e Cooper (1993), que obtiveram melhoras significativas no desempenho de RNA a partir do uso de comitê.

O interesse na área teve um grande impulso devido ao crescimento da disponibilidade de computadores cada vez mais velozes e, mais recentemente, aos clusters de computadores de

baixo custo, permitindo a implementação e experimentação de máquinas diversificadas (Magalhães, 2009). Este fato é considerado, pois utilizar vários especialistas para a resolução de um problema por meio da combinação dos resultados traz a necessidade de um processamento computacional maior.

A formação de um comitê permite a combinação de vários algoritmos, parâmetros, características, fontes de informação e combinações de dados. As vantagens desta abordagem estão em permitir explorar informações complementares, redundância e simultaneidade de avaliação. Além disso, é possível expressar preferências sobre um determinado especialista com base em suas tendências e custo de execução.

As máquinas de comitê são aproximadores universais que, conforme Haykin (2001), podem ser divididos em duas grandes categorias: estruturas estáticas e estruturas dinâmicas.

As estruturas estáticas são aquelas em que a saída obtida consiste somente da composição das várias respostas provenientes dos especialistas utilizados no comitê, não utilizando diretamente os dados de entrada. Já as estruturas dinâmicas são todas aquelas em que além da composição das respostas dos vários especialistas, são utilizados de forma direta os dados de entrada para a escolha dos pesos de cada especialista. Justamente por a forma de combinação dos resultados dos especialistas ser diferente dependendo do sinal de entrada, essas estruturas são chamadas de dinâmicas.

2.1 Motivações de Uso

É uma prática comum na área de redes neurais artificiais fazer o treinamento de várias redes candidatas diferentes e escolher a melhor, mantendo somente esta e descartando as demais. As desvantagens deste processo é primeiramente o desperdício de esforço para o treinamento das redes descartadas, e em segundo lugar, o problema de que a melhor rede escolhida para o conjunto de dados do treinamento pode não ser a melhor para um novo conjunto de dados (Bishop, 1995).

O uso de uma máquina de comitê provê vantagens na medida em que os problemas já apresentados são resolvidos utilizando-se de todas as redes treinadas, e é justificado por melhorar os resultados dos algoritmos de redes neurais artificiais, principalmente.

Algumas razões para o uso de máquinas de comitês são apontadas por (Polikar, 2006):

- **Razões estatísticas:** redução do risco de escolha de um especialista com baixo desempenho. O desempenho do comitê pode ou não superar o desempenho do

melhor especialista, mas certamente reduz o risco global de fazer a má escolha de um especialista;

- **Grandes volumes de dados:** o treinamento de um especialista com uma quantidade vasta de dados não é prático, porém, particionar os dados em subconjuntos menores e usar vários especialistas para uma solução tem se provado uma abordagem mais eficiente.
- **Poucos dados:** a partir de técnicas de reamostragem é possível elaborar subconjuntos aleatórios e sobrepostos para treinar cada um dos especialistas, criando um comitê.
- **Divisão e conquista:** quando um problema é muito complexo para um especialista, o comitê facilita a solução a partir da divisão em problemas menores e mais simples. Um exemplo na classificação de dados é quando as regiões de dados que representam cada classe se tornam muito complexas: a solução vem da combinação de vários classificadores que dividem as regiões em formas mais simples.
- **Fusão de dados:** é o uso de conjuntos de dados obtidos de origens diversas, onde as características não são homogêneas. Os resultados de exames de sangue, ressonâncias magnéticas e eletroencefalogramas, por exemplo, podem ser combinadas em um comitê para realizar o diagnóstico de uma desordem neurológica.

Existem várias outras aplicações para máquinas de comitê que podem ser vantajosas justamente por poderem ser aplicadas em uma diversidade de situações. Entretanto, cada uma dessas situações deve ser analisada de forma aprofundada, a fim de verificar sua aplicabilidade.

2.2 Estruturas Estáticas

As máquinas de comitê estáticas, também conhecidas como ensembles, são aquelas onde os especialistas são treinados para resolver uma determinada tarefa de forma isolada e, possivelmente independente. Cada especialista representa uma solução ao problema, podendo ser obtida por meios distintos e independentes entre si (Villanueva, 2006).

Nesse tipo de estrutura a forma de combinação das respostas provenientes de cada especialista é feita por meio de um mecanismo que não utiliza diretamente os dados de entrada do problema abordado (Haykin, 2001).

Uma representação gráfica da combinação de especialistas em uma estrutura estática pode ser vista na Figura 2.1. Os dados de entrada são repassados aos diversos especialistas escolhidos, que geram uma saída com o resultado alcançado individualmente. Os resultados de cada especialista são combinados por meio de um comitê que gera o resultado final.

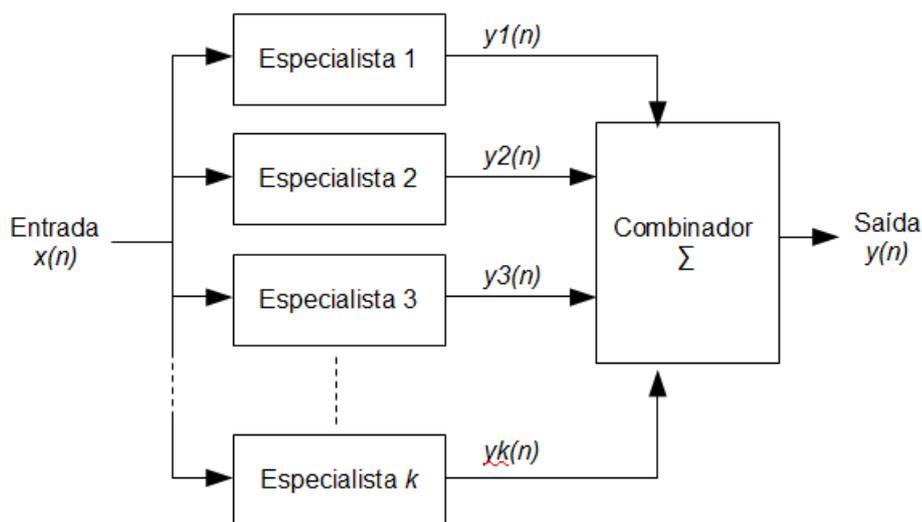


Figura 2.1: Exemplo de Máquina de Comitê Estática

Fonte: Adaptada de Kamo e Dagli (2009)

As máquinas de estruturas estáticas são construídas em dois passos: geração dos vários componentes e a combinação da saída proposta pelos componentes (Lima, 2004). Quanto à geração dos componentes Haykin (2001) inclui como principais abordagens: *ensemble averaging* (média do conjunto) e *boosting* (impulso).

Quanto à combinação das saídas, as abordagens predominantes são, segundo Lima (2004), voto majoritário simples ou ponderado para problemas de classificação e média simples ou ponderada para problemas de regressão.

A seguir são descritas as principais abordagens de geração de componentes: *ensemble averaging* e *boosting*, e a abordagem predominante na combinação das saídas para os problemas de classificação, a votação majoritária.

2.2.1 Média do Conjunto

A *ensemble averaging* ou média do conjunto é uma das técnicas mais simples para geração de um comitê. Os especialistas compartilham a mesma entrada de dados para treinamento e combinam as saídas de cada um linearmente para produzir uma nova saída (Haykin, 2001). A combinação dos especialistas é feita por métodos de votação simples no caso de classificação ou média simples no caso de predição. Este tipo de comitê pode ser visto na Figura 2.1.

O processo de treinamento para esta técnica consiste em escolher o conjunto de dados para o treinamento e treinar cada um dos especialistas com o mesmo conjunto. Para a classificação basta fazer a escolha de um método de combinação, executar cada especialista e combinar as saídas de cada especialista para obter o resultado final.

Uma técnica derivada da média do conjunto é o *bagging* (*bootstrap aggregating*) também conhecido como ensacamento. Enquanto na média do conjunto os dados de treinamento são os mesmos para cada especialista, no ensacamento os dados são levemente diferentes para cada um dos especialistas (Breiman, 1996). Os conjuntos de treinamento para cada especialista são uma amostra do conjunto de treinamento do comitê.

2.2.2 Impulso

Boosting ou impulso são as máquinas de comitê em que algoritmos de baixa precisão (menos de metade de acertos) são combinados de forma a convergir em uma saída com precisão elevada (Haykin, 2001). Isto acontece por existirem subconjuntos de dados com distribuições diferentes, fazendo com que cada algoritmo de baixa precisão se especialize em partes distintas do conjunto de dados.

Conforme Witten, Frank e Hall (2012), os algoritmos de *bagging* e *boosting* adotam as mesmas fórmulas para a combinação dos especialistas, votação ou média. A diferença está na maneira que é usada: no ensacamento ou mesmo na média do conjunto, a votação/média é feita de forma simples, enquanto que no impulso a votação/média é feita de forma ponderada.

Idealmente cada especialista em um comitê deveria complementar outro, sendo que cada um trataria a partição de dados em que é bom. Como isso não acontece sempre, os pesos para cada um dos especialistas é calculado de forma a dar mais influência no resultado final para os especialistas com maior chance de sucesso.

Esta técnica, segundo Haykin (2001), pode ser implementada de três maneiras distintas:

- **Filtragem:** envolve a filtragem dos exemplos de treinamento pelas diferentes versões do algoritmo de aprendizagem fraca, requerendo grandes conjuntos de dados para treinamento;
- **Subamostragem (*Adaboosting*):** utiliza conjunto de treinamento de tamanho fixo que são reamostrados durante o treinamento de acordo com uma distribuição de probabilidade;
- **Re-ponderação:** utiliza conjunto de treinamento fixo no qual os exemplos são ponderados por um algoritmo de aprendizado fraco.

2.2.3 Votação Majoritária

O método de votação majoritária consiste em escolher o resultado mais votado dentre os especialistas utilizados no comitê (Lima, 2004).

Existem dois tipos de votação que podem ser realizados: votação majoritária simples e votação majoritária ponderada. Na votação simples cada saída de um especialista afeta igualmente a saída global, e na votação ponderada, a saída de cada especialista tem um peso diferente sobre a saída global.

De acordo com Polikar (2006) existem ainda três variações da votação majoritária (que podem ser aplicadas tanto à votação simples quanto ponderada):

- **Votação Unânime:** onde todos os especialistas concordam no resultado;
- **Maioria Simples:** onde a metade mais um dos especialistas concordam no resultado;
- **Pluralidade ou Majoritária:** onde o resultado escolhido será simplesmente o com maior número de votos.

O método de votação ponderada é utilizado quando é evidente que certos especialistas são melhores qualificados que outros para a tarefa. Desta forma, a chance de aumento do desempenho da máquina de comitê é significativamente maior (Polikar, 2006).

2.3 Estruturas Dinâmicas

As máquinas de comitê dinâmicas se utilizam dos dados de entrada do problema abordado, integrando-os às saídas dos especialistas. Na Figura 2.2 é representada uma máquina de comitê dinâmica, onde a saída de cada especialista tem seu peso alterado por meio de uma

rede de propagação (*gating network*) que distribui os pesos considerando os dados de entrada (Haykin, 2001).

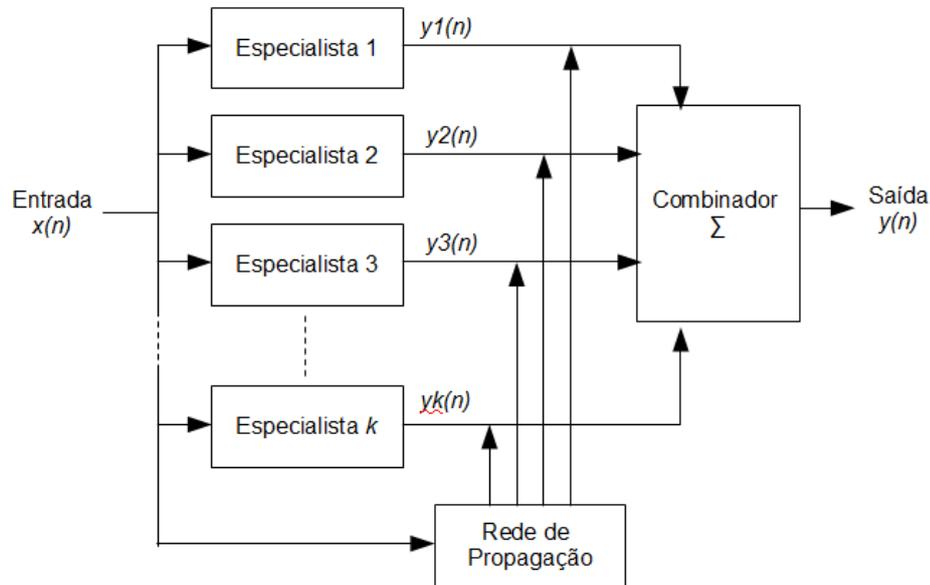


Figura 2.2: Exemplo de Máquina de Comitê Dinâmica

Fonte: Adaptada de Kamo e Dagli (2009)

A rede de propagação determina, por meio da distribuição dos pesos, qual especialista ficará responsável por cada uma das diferentes regiões no espaço de dados de entrada (Ho-Man, 2003). O comitê então combina as saídas dos especialistas com seus pesos já considerados pela rede de propagação.

A implementação da rede de propagação tem importância fundamental na saída resultante da máquina de comitê, sendo que ela deve ter como princípio a atribuição dos pesos para cada especialista considerando o conjunto de dados, de forma a atribuir maiores pesos aos especialistas que irão resolver melhor o problema a partir dos dados de entrada.

A necessidade de uma rede de propagação se dá pelo fato de que diferentes especialistas trabalham melhor em diferentes regiões do espaço. Ainda, as estruturas dinâmicas podem ser subdivididas em: mistura de especialistas e mistura hierárquica de especialistas (Haykin, 2001).

2.3.1 Mistura de Especialistas

Na Mistura de Especialistas (ME) as saídas de cada um dos especialistas são combinadas de forma não linear por meio de uma rede de propagação única, como na Figura 2.2.

Um comitê de mistura hierárquica de especialistas consiste em k módulos supervisionados chamados de especialistas, e uma unidade de integração chamada *gating network* que assume a função de mediador entre os especialistas (Haykin, 2001).

2.3.2 Mistura Hierárquica de Especialistas

A Mistura Hierárquica de Especialistas (MHE) é uma extensão natural da ME e consiste em combinar de forma também não linear as saídas de cada especialista do comitê, porém, fazendo isso por meio de várias redes de propagação agrupadas de forma hierárquica, ou seja, um comitê com mistura hierárquica de especialistas tem duas ou mais camadas de *gating networks* (Tresp, 2001).

A arquitetura da MHE difere da ME da seguinte forma: o espaço de dados de entrada é dividido em um conjunto de subespaços aninhados, com a informação sendo combinada e distribuída entre os especialistas sobre o controle das diversas redes de propagação organizadas de forma hierárquica (Haykin, 2001).

2.3 Trabalhos Correlatos

As máquinas de comitê vêm sendo usadas em diversas áreas, como sistemas de tomada de decisão, classificação de dados (por exemplo, em reconhecimento de faces (Tang, Lyu e King, 2003), (Zaho, Huang e Sun, 2004) e Libras (Madeo, *et al.*, 2010), predição (como em Bolsa de valores (Macedo, 2010), (Kamo e Dagli, 2009)), dentre outras.

Como já descrito, a combinação de estimadores para a melhora do desempenho final de um sistema de classificação, regressão ou predição já vem de longa data. Em 1965 Ablow e Kaylor (1965) escreveram um artigo discutindo sobre uma solução de um comitê para o reconhecimento de padrões, utilizando a lógica de maioria.

Mais tarde, Takyiama (1978) descreve um método geral para treinamento de máquinas de comitê com uma lógica arbitrária. O algoritmo proposto é um procedimento de correção de erros baseado em um método gradiente que procura o valor mínimo de uma função de critério racional. Após testes concluíram que o método proposto para máquinas de comitê foi efetivo.

A intensificação da pesquisa na área só veio mais tarde, no início da década de 90, por meio de trabalhos importantes de vários autores diferentes, cada um propondo uma forma diferente de combinar os resultados dos especialistas a fim de obter melhores resultados, geralmente de forma específica para um algoritmo.

Perrone e Cooper (1993) propõem uma abordagem teórica de ensemble com redes neurais híbridas para melhorar significativamente estimativas de regressão. Esta abordagem já possuía as vantagens de utilizar todas as redes de uma população, utilizar todo o conjunto de dados sem sobreposição, e estar preparada para o uso em computação paralela. Os resultados obtidos por esses autores em uma avaliação experimental foram de uma melhora significativa na tarefa de reconhecimento de caracteres.

Guo e Luh (2004) publicaram um artigo sobre predição de preço de equilíbrio de mercado utilizando um comitê de redes neurais. A representação dos dados de entrada e saída foi mais bem utilizada por um comitê do que uma rede neural artificial aplicada de forma individual. Eles propõem que os pesos de cada especialista sejam as probabilidades do especialista de capturar o verdadeiro relacionamento entre dados de entrada e saída. Os resultados são melhores quando comparados com especialistas individuais e também melhores em comparação com máquinas de comitê que utilizam algoritmos de média de conjunto.

Yang (2005) descreve uma máquina de comitê híbrida para aprendizado incremental onde propõe modificações para o algoritmo incremental de ensemble. As modificações mostraram também aumento no desempenho de classificação.

Raafat, Tolba e Shaddad (2009) discutem sobre máquinas de comitê para reconhecimento facial de gênero utilizando-se de redes neurais artificiais LVQ em comitê de votação majoritária para resolver os problemas de falta de acurácia e instabilidade nos classificadores. Os resultados de desempenho relatados são de 99% de acertos, comparados aos 93% de acertos utilizando o melhor classificador do comitê individualmente.

Diversos outros trabalhos na área de extração de conhecimento e tomadas de decisão fazem o uso de máquinas de comitê, tanto para classificação quanto regressão e predição. Cada um utilizando de diferentes terminologias e taxonomias para definir a área de comitê, assim como também com as mais variadas aplicações no mundo real.

Capítulo 3

Módulo de Máquina de Comitê Proposto

Este capítulo aborda a fase de construção da máquina de comitê proposta para o processo de classificação de dados e a sua integração como um módulo da ferramenta YADMT. Os algoritmos utilizados para a combinação dos especialistas são os de votação majoritária simples e votação majoritária ponderada, que são descritos abaixo.

3.1 Algoritmos

Dentre os algoritmos de uma máquina de comitê estática para combinação das saídas dos especialistas estão as abordagens de votação (classificação) e média (predição). Como o módulo proposto é para classificação de dados, foram implementados os algoritmos de votação majoritária simples e votação majoritária ponderada.

3.1.1 Votação Majoritária Simples

Na formação de máquina de comitê para classificação de dados do tipo votação majoritária, é necessário um número ímpar de classificadores para cumprir o papel de “eleitores” da votação. Para que os resultados tenham diferença no sentido de aumentar a taxa de acertos em comparação aos resultados individuais dos classificadores, também é necessário que o número de classificadores seja maior do que um. Considerando as restrições, o número mínimo de classificadores para esta técnica é de três (Polikar, 2006).

O processo de criação da máquina de votação consiste na escolha dos classificadores e a configuração dos mesmos de acordo com as características de cada um. Após a criação da máquina, faz-se a escolha do método de divisão dos dados em dados de treinamento e dados de teste.

Para a execução do comitê, cada classificador é treinado separadamente com o mesmo conjunto de dados de treinamento escolhido. Ao executar o processo de classificação para as entradas dos dados de testes, cada algoritmo executa de forma separada e em série a fim de obter um resultado individual. A composição dos classificadores é feita de forma a fazer a contagem dos votos dos classificadores para encontrar a saída escolhida (classe) como resultado do comitê. O Algoritmo 3.1 traz um pseudocódigo desse processo.

```
defina a tupla de atributos para o teste;  
inicialize um contador de votos para cada classe;  
para cada classificador faça  
    execute o classificador sobre a tupla;  
    incremente o contador de votos da classe escolhida em 1;  
retorne a classe com o contador mais alto
```

Algoritmo 3.1: Processo de classificação por Votação Simples

A criação de uma máquina de comitê com a combinação dos especialistas sendo pelo método de votação majoritária simples constitui uma máquina que pode ser chamada de *bagging*, conforme (Witten, Frank e Hall, 2012).

3.1.2 Votação Majoritária Ponderada

A votação majoritária ponderada ocorre de forma muito semelhante à votação majoritária simples, sendo que a diferença que ocorre entre uma e outra é que na forma de votação ponderada, é necessário atribuir um peso para cada classificador ou grupo de classificadores. Um pseudocódigo desse processo pode ser visto no Algoritmo 3.2.

```
defina a tupla de atributos para o teste;  
inicialize um contador de votos para cada classe;  
para cada classificador faça  
    execute o classificador sobre a tupla;  
    incremente o contador de votos da classe escolhida em 1 * peso;  
retorne a classe com o contador mais alto
```

Algoritmo 3.2: Processo de classificação por Votação Ponderada

Com esta técnica é possível atribuir pesos maiores para classificadores com níveis de acerto mais elevados, e pesos menores para classificadores com níveis de acertos mais baixos

(Haykin, 2001). Desta forma, o voto de um classificador com peso maior poderá definir o resultado da eleição, a menos que mais de um classificador com peso menor faça escolha outra saída como resultado.

Conforme Witten, Frank e Hall (2012), a construção de um comitê utilizando-se do método de votação majoritária ponderada para a combinação dos especialistas, é a principal característica de uma máquina de comitê de impulso (*boosting*).

O ajuste dos pesos de cada classificador é fundamental para o desempenho da máquina de comitê. É necessário sempre ter em vista que os pesos devem ser escolhidos de forma a permitir uma “eleição” justa dos resultados, não atribuindo pesos demasiadamente altos ou baixos para um classificador (Ho-Man, 2003).

Desta forma, entende-se que um peso muito alto para um classificador acabará transformando a saída do comitê na saída individual do próprio classificador. Já um peso muito baixo acabará excluindo o classificador do comitê na votação, o que não é útil considerando que o comitê despendeu tempo para o treinamento de um classificador que não será utilizado efetivamente.

3.2 Integração na YADMT

Na Figura 3.1 pode-se ver a interface da ferramenta YADMT, no momento de escolha da técnica de classificação e sua configuração.

Além dos métodos de classificação a ferramenta dispõe de técnicas de particionamento de dados para treinamento e testes (*holdout*, *k-fold*, *stratified holdout* e *stratified k-fold*) e das métricas de avaliação de resultados: acuidade, precisão e taxa de renovação (*recall*).

A técnica de Máquina de Comitê deve ser escolhida da mesma forma que outra técnica de classificação. A partir dessa escolha, as propriedades do comitê devem ser determinadas a partir da janela de configurações, a exemplo do disposto na Figura 3.2. Nesta tela é escolhida a estrutura do comitê entre as opções de votação majoritária simples ou ponderada, além dos classificadores (e a quantidade) que irão compor o comitê.

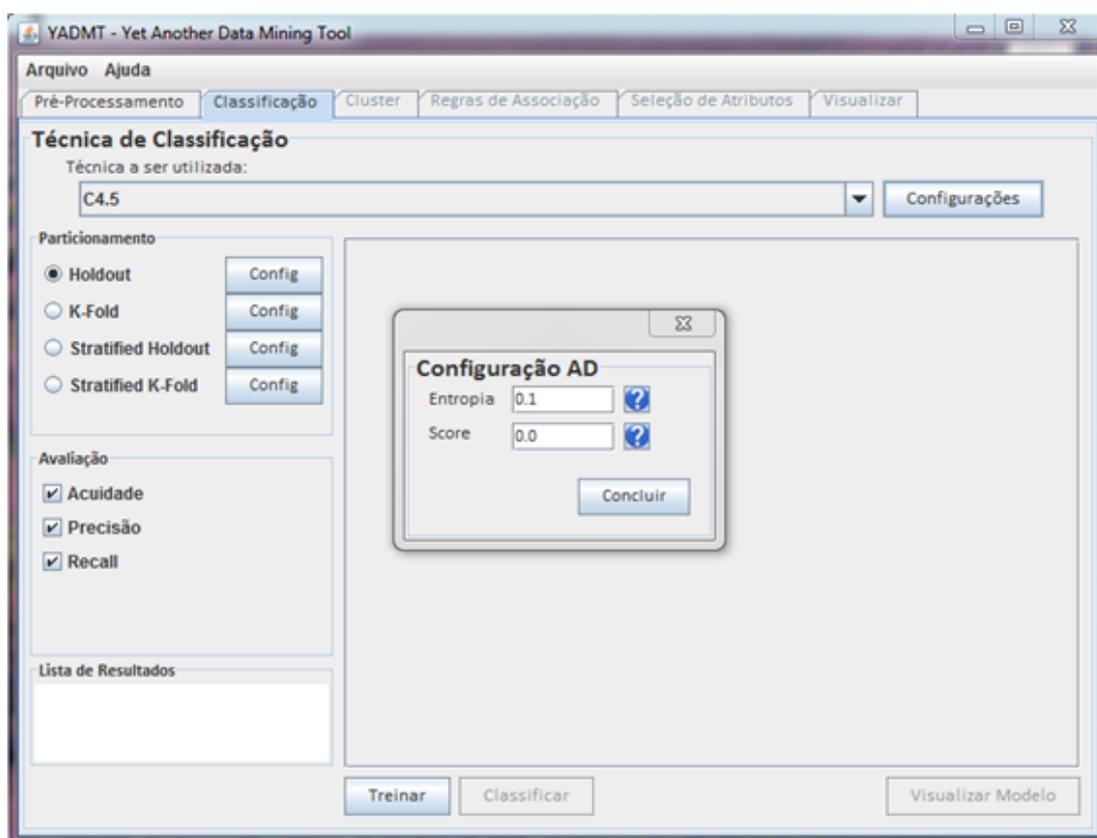


Figura 3.1: Tela da ferramenta YADMT – Etapa de classificação

Fonte: (Benfatti, *et al.*, 2011).

No exemplo da Figura 3.2, a estrutura escolhida é a de votação simples, utilizando três classificadores para a formação do comitê: as Redes Neurais Artificiais LVQ e MLP, além do algoritmo K-NN (*K-Nearest Neighbor*). O usuário pode escolher qualquer número de classificadores para a formação do comitê, sendo no mínimo de três a fim de que o resultado da votação seja diferente dos classificadores que o formam.

A ferramenta não precisou de grandes alterações em seu core para que fosse incluído o módulo de comitê de classificadores. Além da inclusão desse módulo em forma de uma biblioteca na ferramenta, foi necessária somente a inclusão da funcionalidade que permite a geração da matriz de confusão consolidada quando utilizado o método *k-fold*.

As alterações citadas foram realizadas e disponibilizadas na ferramenta e no site do projeto também desenvolvido neste trabalho de conclusão de curso, disponível no endereço eletrônico <http://www.inf.unioeste.br/~clodis/yamdt/>, que inclui uma descrição geral da ferramenta e referência a trabalhos publicados no Projeto.

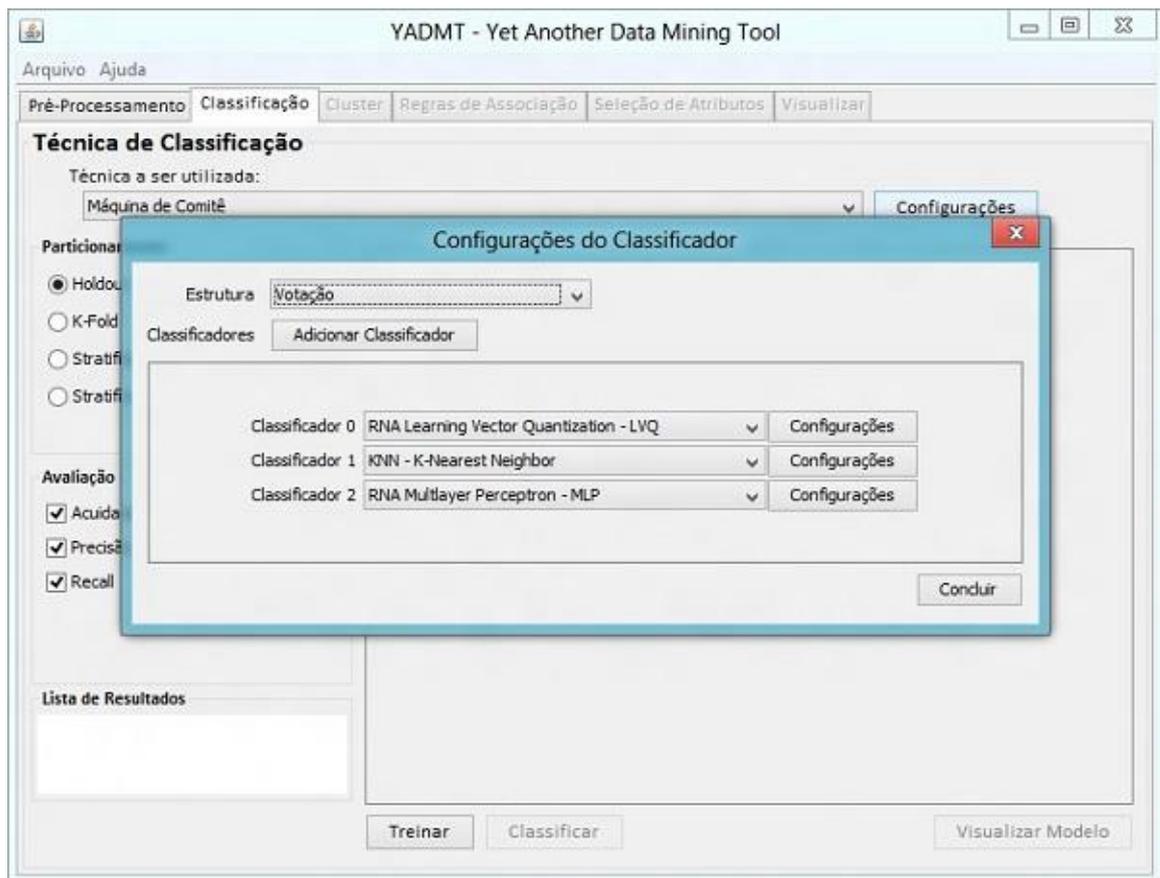


Figura 3.2 Tela da ferramenta YADMT – Configuração da Máquina de Comitê

Nas configurações vistas na Figura 3.2 pode ser escolhido o algoritmo de combinação dos classificadores além de um número ilimitado de classificadores para o comitê. Cada classificador escolhido pode ser configurado conforme suas propriedades. Um comitê pode ser formado por mais de uma instância de um mesmo classificador, desde que suas propriedades configurem resultados diferentes.

Ao executar a função “Treinar”, o algoritmo de comitê é executado e as saídas são apresentadas no painel principal em forma de texto, informando dados específicos de cada classificador, as matrizes de confusão e as métricas de avaliação geradas por classe conforme selecionadas. A Figura 3.3 traz um painel exemplo para a base de dados *Iris*.

```

===== Matriz de Confusão =====
a b c
5 0 0 a = Iris-setosa
0 4 1 b = Iris-versicolor
0 0 5 c = Iris-virginica

219ms

===== Métricas de Avaliação =====

=====
Acuracia
0,9400000000 : Iris-setosa
0,8181818182 : Iris-versicolor
0,7758620690 : Iris-virginica
0,8446812957 : Geral
=====

=====
Precisao
0,9400000000 : Iris-setosa
0,9000000000 : Iris-versicolor
0,9000000000 : Iris-virginica
0,9133333333 : Geral
=====

=====
Recall
1,0000000000 : Iris-setosa
0,9000000000 : Iris-versicolor
0,8490566038 : Iris-virginica
0,9163522013 : Geral
=====

```

Figura 3.3 Tela da ferramenta YADMT – Resultados da execução da Máquina de Comitê

Esse módulo tem por objetivo principal dar ao usuário da YADMT mais liberdade no processo de classificação de dados, permitindo utilizar comitês com qualquer classificador implementado. Além disso, objetiva prover todos os benefícios que uma máquina de comitê pode gerar, principalmente na melhoria dos resultados de classificação.

Capítulo 4

Avaliação experimental

Este capítulo traz uma avaliação experimental dos algoritmos de votação majoritária no intuito de avaliar a eficácia do uso de máquina de comitê, descrevendo as bases de dados utilizadas para a execução dos testes, assim como a metodologia de avaliação adotada.

4.1 Bases de Dados

Foram utilizadas dez das bases de dados do repositório da UCI (Frank e Asuncion, 2010) para a avaliação do módulo proposto e execução dos testes. Para fins comparativos foram utilizadas as mesmas bases de dados utilizadas por Benfatti (2010) e Bonifácio (2010), listadas na Tabela 4.1, cujas descrições são apresentadas no Apêndice A.

Tabela 4.1: Bases de dados escolhidas para comparação das Técnicas

Nome da Base de Dados	Tamanho	Atributos	Classes	Atributos Categóricos
<i>Abalone</i>	4177	8	28	1
<i>Acute Inflammations</i>	120	6	2	5
<i>Ecoli</i>	336	8	8	0
<i>Iris</i>	150	4	3	0
<i>Libras Movement</i>	360	91	15	0
<i>Spambase</i>	4601	57	2	0
<i>Statlog (Image Segmentation)</i>	2310	18	7	0
<i>Statlog (Landsat Satellite)</i>	6435	36	7	0
<i>Teaching Assistant Evaluation</i>	151	5	3	3
<i>Wine Quality</i>	4898	12	7	0

Estas bases de dados foram testadas com os algoritmos utilizados também para a formação da máquina de comitê.

4.2 Metodologia de Avaliação

Estimar a qualidade de um classificador é sem dúvida um problema de grande importância, considerando que a meta do mesmo após o treinamento é classificar com sucesso casos de dados não encontrados anteriormente (Batista, 1997). É importante a existência de mecanismos para avaliar um algoritmo, e o quão bom um algoritmo é para um problema específico (Witten e Frank, 2005).

Conforme Han e Kamber (2005), os seguintes critérios podem ser usados para a avaliação:

- **Acuidade de predição:** é a habilidade do modelo de predizer corretamente a classe para um novo dado desconhecido;
- **Velocidade:** custo computacional envolvido na geração e uso do modelo criado;
- **Robustez:** refere-se à habilidade do modelo em predizer corretamente um dado que apresente ruído ou valores ausentes;
- **Escalabilidade:** habilidade de construir modelos eficientes considerando volumes de dados grandes;
- **Interpretabilidade:** refere-se ao grau de compreensibilidade que o modelo proporciona ao usuário da aplicação.

Para uma avaliação eficaz e fiel às características dos algoritmos é necessário que as amostras utilizadas no treinamento e teste sejam escolhidas cuidadosamente (Costa, 2008).

Uma maneira de fazer isso é separar os conjuntos de teste e de treinamento em grupos mutuamente exclusivos (Batista, 1997).

Uma das técnicas para a divisão dos dados nos conjuntos de treinamento e teste é o *holdout*, onde a divisão é feita em p exemplos para treinamento (geralmente dois terços dos dados) e $p-1$ para testes. Para grandes conjuntos de dados quanto maior o conjunto de teste, maior será a precisão do classificador (Witten e Frank, 2005).

Para conjuntos de testes menores os dados podem não ser representativos o suficiente, ou ainda, uma classe pode estar presente em apenas um conjunto, prejudicando o treinamento do classificador. Para evitar estes problemas deve-se fazer a estratificação (*holdout stratification*) a fim de garantir a proporcionalidade das classes entre os conjuntos (Witten e Frank, 2005).

Outra técnica para evitar este problema é realizar a troca entre conjuntos de teste e treinamento, chamada de *k-fold-cross-validation*, no qual os dados são divididos em *k* partições (Witten e Frank, 2005) e em cada interação uma partição é usada para teste e as demais para treinamento.

Uma matriz de confusão mostra o número de classificações corretas comparando às classificações que foram pressupostas. As quatro possibilidades de uma classificação são verdadeiros positivos (classificado corretamente na classe positiva), verdadeiros negativos (classificado corretamente na classe negativa), falsos positivos (classificado erroneamente na classe positiva quando são da classe negativa) e falsos negativos (classificados erroneamente na classe negativa quando são da positiva), respectivamente TP, TN, FP, FN na Tabela 4.2.

Tabela 4.2: Aplicações da Análise de Imagens. Adaptada de (Costa, 2008)

		Classe Predita	
Classe Verdadeira		Positiva	Negativa
Positiva		TP	FN
Negativa		FP	TN

Com base na matriz de confusão podem ser calculadas algumas medidas de desempenho utilizadas na avaliação, como por exemplo: acuidade, precisão e renovação (*recall*).

A acuidade, acurácia ou taxa de acerto (TA), Equação 4.1, avalia o quão efetivo um algoritmo é, por meio da probabilidade do algoritmo fazer predições corretas (Costa, 2008).

$$TA = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \quad (4.1)$$

A precisão, conhecida também como confiança (P), Equação 4.2, é a medida que estima a probabilidade da predição positiva estar correta (Costa, 2008).

$$P = \frac{|TP|}{|TP| + |FP|} \quad (4.2)$$

A taxa de renovação (*recall*) – ou cálculo de sensibilidade (S), Equação 4.3, estima a probabilidade de um exemplo pertencente a classe positiva ser predito como positivo.

$$S = \frac{|TP|}{|TP| + |FN|} \quad (4.3)$$

4.3 Análise dos Resultados

Nesta seção são listados os resultados dos testes de avaliação da máquina de comitê proposta para as bases de dados escolhidas.

Todos os testes foram realizados com a técnica de particionamento *stratified k-fold* para a divisão dos dados em conjunto de teste e conjunto de treinamento. As avaliações foram feitas sobre as métricas de acuidade, precisão e renovação (*recall*). A técnica de combinação das saídas de cada classificador usada para a formação do comitê foi a de votação por maioria, onde para a classificação de cada tupla a classe escolhida foi a mais votada entre os classificadores de forma individual.

Para testes de todas as bases de dados foram utilizados os classificadores com as mesmas configurações para efeito de comparação. A escolha das propriedades de cada classificador foi feita empiricamente, por meio de vários testes com diferentes configurações de maneira a selecionar somente aquelas que produziam os melhores resultados e mais relevantes. Os algoritmos foram configurados da seguinte forma:

- **T1 – K-NN:** Variável K = 10, Regra de Inferência: Maioria Por Votação; Função Distância: *Manhatam*
- **T2 – C4.5:** Entropia 0.5, Score: 0;
- **T3 – Naive Bayes:** Configuração Padrão
- **T4 – LVQ:** Neurônios da camada oculta: 10; Épocas: 1000; Taxa de Aprendizado: 0.3; Taxa de Redução de Aprendizado: 0.01;
- **T5 - MLP:** Neurônios da camada oculta: 5,5; Épocas: 500; Taxa de Aprendizado: 0.5; Erro Mínimo: 0.01; Taxa *momentum*: 0.2;
- **T6 – RBF:** Neurônios da camada oculta: 5; Épocas: 500; Taxa de Aprendizado: 0.8; Erro mínimo: 0.01;
- **T7 – Comitê 1:** Comitê formado por T1, T2 e T3;
- **T8 – Comitê 2:** Comitê formado por T4, T5 e T6;
- **T9 – Comitê 3:** Comitê formado por T5, T5 e T6;
- **T10 – Comitê 4:** Comitê formado por T5, T6 e T6;

A formação dos comitês para os testes foi feita de forma a comparar os classificadores de Benfatti (2010) em T7, e os classificadores de Bonifácio (2010) de forma igualitária em T8 e

de forma ponderada em T9 (peso maior para T5) e T10 (peso maior para T6). Em todos os comitês foram usados três classificadores, sendo este o número mínimo necessário considerando a técnica de votação, porém nada impede a formação de comitês com um número maior de classificadores.

Os resultados para todas as bases de dados são demonstrados em forma de tabela, onde para cada classificador são exibidas as métricas de avaliação. O melhor resultado de cada métrica é destacado em negrito. Os resultados que contém letras iguais ao lado do seu valor significam uma igualdade de proporções, ou seja, são estatisticamente iguais ao nível de 5% de significância.

A igualdade de proporções é um teste para comparar se há diferença estatística significativa entre as proporções amostrais de duas situações diferentes a um nível de significância. Para a comparação dos resultados deste trabalho, o teste foi feito comparando todos os valores das tabelas para cada métrica de avaliação.

Como os tempos de processamento na etapa de classificação dos dados são de apenas alguns segundos e variam pouco entre todos os classificadores, eles não foram incluídos. A única exceção foi para o algoritmo C4.5 ao realizar o processo de classificação para a base de dados *Abalone*, que possui um grande número de amostras e classes, em que o tempo de execução foi superior a 2 horas.

4.3.1 Experimento com a Base de Dados *Iris*

A base de dados *Iris* é uma das mais conhecidas e utilizadas para o reconhecimento de padrões. Possui quatro atributos contínuos e um atributo categórico classificador em um total de 150 amostras. Os respectivos resultados para essa base de dados seguem na Tabela 4.3.

Os resultados quando comparados com Bonifácio (2010) que inclui somente a acuidade, foram semelhantes para os algoritmos MLP e RBF, porém, muito piores para o algoritmo LVQ. As diferenças entre os resultados obtidos neste trabalho em comparação aos outros citados anteriormente podem vir a ocorrer devido a uma configuração diferente do classificador ou ainda, do uso de uma versão diferente da ferramenta da qual usada no trabalho dos autores citados.

Comparando os resultados dos classificadores individuais K-NN, C4.5 e *Naive Bayes* com as máquinas de comitê (T7), os resultados de acurácia tendem a ser melhores que a média dos

classificadores, porém, não superiores ao melhor classificador, C4.5. A precisão foi maior no comitê T10 do que nos classificadores individuais para esta base de dados.

Tabela 4.3: Resultados dos Testes para a Base de Dados *Iris*¹

Classificador	Acuidade	Precisão	Recall
T1 – K-NN	0,1666 a	0,3333 a	0,2202 a
T2 – C4.5	0,8446 b	0,9133 b	0,9163 b
T3 – <i>Naive Bayes</i>	0,6413 c	0,7733 c	0,8389 c
T4 – LVQ	0,1349 a	0,3333 a	0,2222 a
T5 – MLP	0,9244 d	0,9600 b,d	0,9618 b,c
T6 – RBF	0,8447 b	0,9133 b	0,9129 b
T7 – Comitê 1	0,6515 c	0,7800 c	0,8388 c
T8 – Comitê 2	0,8991 b,d	0,9466 b	0,9525 b
T9 – Comitê 3	0,8447 b	0,9133 b	0,9129 b
T10 – Comitê 4	0,9486 d	0,9733 d	0,9738 c
<i>Média</i>	0,6899	0,7839	0,7750
<i>Mediana</i>	0,8443	0,9133	0,9129

Comparando os resultados das RNA com os comitês das mesmas (T8, T9, T10), pode-se notar que a acuidade e precisão também se mantêm altas, porém, não superior ao melhor classificador (MLP). O que é importante ressaltar é que quando existe a configuração de uma máquina de comitê (T9) com um classificador de baixa qualidade, o resultado tende a ficar abaixo do mais alto individual (MLP), contudo, ao utilizar o comitê de MLP, MLP e RBF (T10), todos classificadores de alta acurácia, o resultado final tanto de acurácia quanto de precisão e recall é melhor do que todos os resultados dos classificadores individuais.

4.3.2 Experimento com a Base de Dados *Acute Inflammations*

A base de dados *Acute Inflammations* possui um atributo contínuo e sete atributos categóricos em um total de 120 amostras. Os respectivos resultados para essa base de dados são apresentados na Tabela 4.4.

¹ Em todos os experimentos relatados, os resultados notados com letras iguais significam resultados estatisticamente iguais a 5% de significância. O melhor resultado de cada métrica foi destacado em negrito.

Os resultados para T1, T2 e T3 estão dentro do esperado em comparação com os resultados de Benfatti (2010). Já entre T4, T5 e T6 os resultados não foram diferentes dos encontrados em comparação a Bonifácio (2010).

Entre os comitês, T7 foi abaixo do melhor, porém, superior à média. O T8 também se sobressaiu da média. Os resultados de T9 foram reduzidos por influência de T5, e os resultados de T10 foram aumentados por influência de T6.

Tabela 4.4: Resultados dos Testes para a Base de Dados *Acute Inflammations*

Classificador	Acuidade	Precisão	Recall
T1 – K-NN	0,2688 a	0,4241 a	0,4234 a
T2 – C4.5	0,6664 b	0,8007 b	0,8027 b
T3 – <i>Naive Bayes</i>	0,5654 b,c	0,7270 b, c	0,7400 b
T4 – LVQ	0,4219 d	0,5976 d	0,6061 c
T5 – MLP	0,2541 a	0,5000 a, d	0,2541 d
T6 – RBF	0,6383 b	0,7860 b	0,8178 b
T7 – Comitê 1	0,6098 b	0,7595 b	0,7641 b
T8 – Comitê 2	0,4736 c, d	0,6544 c, d	0,6974 b, c
T9 – Comitê 3	0,2541 a	0,5000 a, d	0,2541 d
T10 – Comitê 4	0,6383 b	0,7860 b	0,8178 b
<i>Média</i>	<i>0,4790</i>	<i>0,6535</i>	<i>0,6177</i>
<i>Mediana</i>	<i>0,5195</i>	<i>0,6907</i>	<i>0,7187</i>

4.3.3 Experimento com a Base de Dados *Teaching Assistant Evaluation*

A base de dados *Teaching Assistant Evaluation* possui apenas atributos numéricos inteiros em um total de 151 amostras. Os respectivos resultados para essa base de dados são apresentados na Tabela 4.5.

Os resultados para T1, T2 e T3 estão dentro do esperado em comparação com os trabalhos anteriores. Já entre T4, T5 e T6 os níveis de acuidade, precisão e *recall* ficaram abaixo do esperando, comparando-se com os resultados de Bonifácio (2010). Entre os comitês, o

resultado de T7 foi abaixo do melhor, porém, superior à média destes. O resultado de T8 ficou abaixo dos classificadores que o compõe. Os resultados de T9 foram reduzidos por influência de T5, e os resultados de T10 foram aumentados por influências de T6.

Tabela 4.5: Resultados dos Testes para a Base de Dados *Teaching Assistant Evaluation*

Classificador	Acuidade	Precisão	Recall
T1 – K-NN	0,1621 a	0,3253 a	0,2179 a
T2 – C4.5	0,4945 b	0,6615 b	0,6617 b
T3 – <i>Naive Bayes</i>	0,4202 b	0,5924 b	0,6112 b
T4 – LVQ	0,1081 a	0,3333 a	0,1081
T5 – MLP	0,1542 a	0,3312 a	0,2241 a
T6 – RBF	0,1978 a	0,3307 a	0,3352 c
T7 – Comitê 1	0,4548 b	0,6310 b	0,6607 b
T8 – Comitê 2	0,1488 a	0,3038 a	0,2030 a
T9 – Comitê 3	0,1780 a	0,3228 a	0,3612 c
T10 – Comitê 4	0,1978 a	0,3307 a	0,335c b
<i>Média</i>	<i>0,2516</i>	<i>0,4162</i>	<i>0,3718</i>
<i>Mediana</i>	<i>0,1879</i>	<i>0,3309</i>	<i>0,3352</i>

4.3.4 Experimento com a Base de Dados *Ecoli*

A base de dados *Ecoli* possui sete atributos numéricos contínuos, com um número total de 336 amostras. A Tabela 4.6 apresenta os resultados para essa base de dados.

Em comparação com os trabalhos anteriores, os resultados para T1, T2 e T3 estão dentro do esperado. Os níveis de T4 e T6 foram bons, porém não superam T5. Entre os comitês, o resultado de T7 foi abaixo do melhor, porém superior à média. O resultado de T8 ficou abaixo dos classificadores que o compõe. Os resultados de T9 foram aumentados por influência de T5, e os resultados de T10 foram reduzidos por influências de T6. Nestes testes nenhum classificador atingiu os melhores níveis em todas as métricas. A melhor acurácia (0,5556) foi atingida pelo comitê T9, a melhor precisão (0,6653), pela rede MLP (T5) e o melhor recall

(0,6890) foi obtido pelo comitê T2. De forma geral os classificadores de comitê tiveram resultados melhores do que os classificadores individualmente.

Tabela 4.6: Resultados dos Testes para a Base de Dados *Ecoli*

Classificador	Acuidade	Precisão	Recall
T1 – K-NN	0,0531	0,1250	0,0531
T2 – C4.5	0,2646 a	0,3476 a	0,3925 a
T3 – <i>Naive Bayes</i>	0,1458 b	0,2391 b	0,2521 b
T4 – LVQ	0,3221 a	0,4027 a, c	0,4716
T5 – MLP	0,5832 c	0,6652 d	0,6467 c
T6 – RBF	0,2164 a	0,3546 a,c	0,3488 a
T7 – Comitê 1	0,1483 b	0,2288 b	0,2531 b
T8 – Comitê 2	0,4649	0,5549	0,6889 c
T9 – Comitê 3	0,5555 c	0,6310 d	0,6486 c
T10 – Comitê 4	0,2164 a	0,3546 a,c	0,3488 a
<i>Média</i>	<i>0,2970</i>	<i>0,3903</i>	<i>0,4104</i>
<i>Mediana</i>	<i>0,2405</i>	<i>0,3546</i>	<i>0,3706</i>

4.3.5 Experimento com a Base de Dados *Abalone*

A base de dados *Abalone* possui um grande número de amostras (4177), porém, um número pequeno de atributos (8), tanto numéricos quanto categóricos. Os respectivos resultados para essa base de dados são apresentados na Tabela 4.7.

Os resultados de todos classificadores executados foi péssimo, com no máximo 5% de acurácia e 10% de precisão. Logo os resultados dos comitês também não foram satisfatórios, sendo seu desempenho semelhante a dos classificadores individuais, independentemente de serem RNA ou não. É interessante ressaltar que enquanto que o tempo médio de execução dos classificadores é de apenas alguns segundos, o tempo de processamento de T2 para esta base de dados foi de 2 horas, assim como o tempo de T7 que inclui T2 na sua formação.

Tabela 4.7: Resultados dos Testes para a Base de Dados *Abalone*

Classificador	Acuidade	Precisão	Recall
T1 – K-NN	0,0114 a	0,0355 a	0,0188 a
T2 – C4.5	0,0309	0,0545 b	0,0619 b
T3 – <i>Naive Bayes</i>	0,0151 a, b	0,0430 a, c	0,0725 b
T4 – LVQ	0,0083 a	0,0655	0,0186 a
T5 – MLP	0,0594 c	0,1119	0,0964 c
T6 – RBF	0,0211d	0,0540 b	0,0364 d
T7 – Comitê 1	0,0190 b, d	0,0441 c	0,0412 d
T8 – Comitê 2	0,0222 d	0,0897 d	0,0568 b
T9 – Comitê 3	0,0549 c	0,0992 d	0,0991 c
T10 – Comitê 4	0,0211 d	0,0540 b	0,0364 d
<i>Média</i>	<i>0,0263</i>	<i>0,0651</i>	<i>0,0538</i>
<i>Mediana</i>	<i>0,0211</i>	<i>0,0542</i>	<i>0,0490</i>

4.3.6 Experimento com a Base de Dados *Spambase*

A base de dados *Spambase* possui um grande número de amostras (4601) e é composta apenas por atributos numéricos. Os respectivos resultados para essa base de dados são apresentados na Tabela 4.8.

Para os classificadores de forma individual o com melhor desempenho de acurácia e precisão foi T2 (0,5213). O comitê T7 ficou perto da média, porém longe da melhor acurácia. Como as RNA tiveram desempenho semelhante, assim também os comitês T8, T9 e T10 tiveram o seu comportamento, assumindo valores de acurácia e precisão próximos de seus classificadores.

Tabela 4.8 Resultados dos Testes para a Base de Dados *Spambase*

Classificador	Acuidade	Precisão	Recall
T1 – K-NN	0,1970	0,5000 a	0,1970
T2 – C4.5	0,5213	0,6712	0,7164 a
T3 – <i>Naive Bayes</i>	0,3626 a	0,6247 b	0,7133 a
T4 – LVQ	0,3035 b	0,5001 a	0,5172
T5 – MLP	0,3040 b	0,5008 a	0,8031 b
T6 – RBF	0,3167 b	0,5098 a	0,6867 c
T7 – Comitê 1	0,3579 a	0,6216 b	0,7136
T8 – Comitê 2	0,3035 b	0,5003 a	0,6363 c
T9 – Comitê 3	0,3040 b	0,5008 a	0,8031 b
T10 – Comitê 4	0,3167 b	0,5098 a	0,6867 c
<i>Média</i>	<i>0,3287</i>	<i>0,5439</i>	<i>0,6473</i>
<i>Mediana</i>	<i>0,3103</i>	<i>0,5053</i>	<i>0,7000</i>

4.3.7 Experimento com a Base de Dados *LIBRAS movement*

A base de dados *LIBRAS movement* é composta de 90 atributos numéricos e um atributo categórico utilizado para definir a classe, apresentando um grande número de classes, com um total de 360 amostras. Os respectivos resultados para essa base de dados são apresentados na Tabela 4.9.

Os resultados para T1, T2 e T3 estão dentro do esperado em comparação com os trabalhos anteriores, com T1 tendo um resultado péssimo (menos de 1% de acurácia). Os níveis de T4 e T6 foram ruins, em contraste com T5 que se destacou dentre todos os classificadores de forma individual.

Entre os comitês, o resultado de T7 foi melhor que cada um dos classificadores individualmente. T8 não obteve melhor resultado que o melhor classificador individual, porém superior à média. O resultado de T9 foi o melhor entre os comitês e, praticamente, acompanhou o resultado de T5, sendo melhor que T4 e T6. Já o resultado de T10, teve seu resultado reduzido, pois teve T6 em sua formação.

Tabela 4.9 Resultados dos Testes para a Base de Dados *LIBRAS movement*

Classificador	Acuidade	Precisão	Recall
T1 – K-NN	0,0095	0,0555	0,0110
T2 – C4.5	0,1585 a	0,2694 a	0,2860
T3 – <i>Naive Bayes</i>	0,1552 a	0,2166 a	0,6052 a
T4 – LVQ	0,1145 a, b	0,2416 a	0,1439 b
T5 – MLP	0,4999 c	0,6444 b	0,6539 a
T6 – RBF	0,0804 b	0,2305 a	0,1046 b
T7 – Comitê 1	0,1619 a, b	0,2305 a	0,5828 a,
T8 – Comitê 2	0,3157	0,4805	0,4995
T9 – Comitê 3	0,4944 c	0,6444 b	0,6116 a
T10 – Comitê 4	0,0804 b	0,2305 a	0,1046 b
<i>Média</i>	<i>0,2070</i>	<i>0,3243</i>	<i>0,3603</i>
<i>Mediana</i>	<i>0,1568</i>	<i>0,2360</i>	<i>0,3927</i>

4.3.8 Experimento com a Base de Dados *Statlog (Image Segmentation)*

A base de dados *Statlog (Image Segmentation)* possui um grande número de amostras com classes igualmente distribuídas e, é composta apenas por atributos tanto numéricos quanto categóricos, sendo composta por um número de 2310 amostras. Os respectivos resultados para essa base de dados são apresentados na Tabela 4.10.

Os resultados de todos os classificadores foram baixos, tendo o melhor resultado entre estes com relação à acurácia e precisão o classificador T6, e em relação ao nível de recall T3. O comitê T7 novamente obteve resultados acima da média, porém não melhores que o classificador de desempenho mais alto. Os comitês T8 e T9 obtiveram resultados ruins devido à sua formação de classificadores conterem T4. Os resultados para o comitê T10 foram impulsionados novamente pelo classificador T6, que obteve os melhores resultados.

Tabela 4.10 Resultados dos Testes para a Base de Dados *Statlog (Image Segmentation)*

Classificador	Acuidade	Precisão	Recall
T1 – K-NN	0,0204	0,1428 a	0,0204
T2 – C4.5	0,2351 a	0,3831 b	0,5687
T3 – <i>Naive Bayes</i>	0,1898	0,3090 c	0,8610
T4 – LVQ	0,0493	0,1324 a	0,0981
T5 – MLP	0,1544	0,2857 c	0,1618
T6 – RBF	0,2805 a	0,4450 c	0,4750 a
T7 – Comitê 1	0,2271 a	0,3705 b	0,6553
T8 – Comitê 2	0,0913	0,2134	0,2368
T9 – Comitê 3	0,0320	0,1536 a	0,1374
T10 – Comitê 4	0,2805 a	0,4450 c	0,4750 a
<i>Média</i>	<i>0,1560</i>	<i>0,2880</i>	<i>0,3689</i>
<i>Mediana</i>	<i>0,1721</i>	<i>0,2973</i>	<i>0,3559</i>

4.3.9 Experimento com a Base de Dados *Statlog (Landsat Satellite)*

A base de dados *Statlog (Landsat Satellite)* é composta apenas por atributos numéricos inteiros com um número grande de amostras, em um total de 6435. Os respectivos resultados para essa base de dados são apresentados na Tabela 4.11.

Os resultados para T2 e T3 foram os melhores entre os classificadores, enquanto que T1 obteve um resultado insignificante (apenas 3%). Entre as RNA T6 teve o melhor desempenho.

Entre os comitês, T7 teve um resultado acima da média, porém abaixo dos níveis do melhor classificador, apenas superando o nível de recall. Já os comitês T8 e T9 não tiveram resultados significativos, considerando que os seus classificadores individualmente não obtiveram bom desempenho. O resultado do comitê T10 foi melhor se comparado aos outros comitês de RNA, pois levou em consideração uma ponderação maior ao classificador T6.

Tabela 4.11 Resultados dos Testes para a Base de Dados *Statlog (Landsat Satellite)*

Classificador	Acuidade	Precisão	Recall
T1 – K-NN	0,0330 a	0,1666 a	0,0330 a
T2 – C4.5	0,6084	0,7393	0,7437
T3 – <i>Naive Bayes</i>	0,5159	0,6342	0,7755
T4 – LVQ	0,0384 a	0,1666 a	0,0384 a
T5 – MLP	0,0384 a	0,1666 a	0,0384 a
T6 – RBF	0,3061 b	0,4683 b	0,6107 b
T7 – Comitê 1	0,4544	0,5775	0,8085
T8 – Comitê 2	0,0384 a	0,1666 a	0,0384 a
T9 – Comitê 3	0,0384 a	0,1666 a	0,0384 a
T10 – Comitê 4	0,3061 b	0,4683 b	0,6107 b
<i>Média</i>	0,2377	0,3720	0,3735
<i>Mediana</i>	0,1722	0,3174	0,3245

4.3.10 Experimento com a Base de Dados *Wine Quality*

A base de dados *Wine Quality* possui um grande número de amostras (4898) e é composta apenas por atributos numéricos contínuos. Os respectivos resultados para essa base de dados são apresentados na Tabela 4.12.

Os resultados foram baixos para todos os classificadores individualmente, exceto T2 e T3 que se sobressaíram ligeiramente, sendo T2 o melhor em acurácia. Considerando os comitês, o resultado de T7 teve níveis de acurácia e recall melhores que os classificadores que o compõe. Já T8, T9 e T10 tiveram resultados baixos em função do desempenho dos classificadores que os compõe.

O comitê T8 formado por T4, T5 e T6 teve uma acurácia melhor do que os classificadores de forma individual. O comitê T9 teve um resultado pior impulsionado por T5 e o comitê T10 teve um resultado melhor impulsionado por T6.

Tabela 4.12 Resultados dos Testes para a Base de Dados *Wine Quality*

Classificador	Acuidade	Precisão	Recall
T1 – K-NN	0,0709 a	0,1666 a	0,0709 a
T2 – C4.5	0,2260 b	0,3173 b	0,3549 b
T3 – <i>Naive Bayes</i>	0,2051	0,3390	0,3525 b
T4 – LVQ	0,0581 c	0,1582 a	0,0829
T5 – MLP	0,0664 a, c	0,1666 a	0,0664 a
T6 – RBF	0,0892 d	0,1596 a	0,1329 c
T7 – Comitê 1	0,2327 b	0,3210 b	0,4639
T8 – Comitê 2	0,0914 d	0,1539 a	0,1386 c
T9 – Comitê 3	0,0664 a, c	0,1666 a	0,0664 a
T10 – Comitê 4	0,0892 d	0,1596 a	0,1329 c
<i>Média</i>	<i>0,1195</i>	<i>0,2108</i>	<i>0,1862</i>
<i>Mediana</i>	<i>0,0892</i>	<i>0,1666</i>	<i>0,1329</i>

4.4 Considerações Finais

Após a análise dos resultados para cada base de dados, foram selecionados os melhores resultados para a formação da Tabela 4.13, que lista as principais características de cada base de dados e os melhores classificadores para cada métrica.

A proporção dos melhores classificadores foi:

- segundo a acurácia T2(4), T5(2), T6(1), T7(1), T9(2) e T10 (2);
- segundo a precisão T2(4), T3(1), T5(2), T6(1), T9(2) e T10(2); e,
- segundo o *recall* T2(1), T3(1), T5(2), T6(1), T7(2), T8(1), T9(2) e T10(2).

De modo geral, pode-se concluir que sobre bases de dados de tamanhos pequenos: *Acute Inflammations*, *Ecoli*, *Iris*, *Libras Movement* e *Teaching Assistant Evaluation*, com um número de amostras abaixo de 400, não é possível estabelecer com certeza absoluta se os resultados são significativamente melhores ou não.

Tabela 4.13 Resumo dos Resultados dos Testes

Bases de dados	Tamanho	Atributos	Classes	Acuidade	Precisão	Recall
<i>Iris</i>	150	4	3	T10	T10	T10
<i>Acute Inflammations</i>	120	6	2	T2	T2	T6,T10
<i>Teaching Assistant Evaluation</i>	151	5	3	T2	T2	T2
<i>Ecoli</i>	336	8	8	T9	T5	T8
<i>Abalone</i>	4177	8	28	T5	T9	T9
<i>Spambase</i>	4601	57	2	T2	T2	T5,T9
<i>Libras Movement</i>	360	91	15	T5,T9	T5,T9	T5
<i>Statlog (Image Segmentation)</i>	2310	18	7	T6,T10	T6,T10	T3
<i>Statlog (Landsat Satellite)</i>	6435	36	7	T2	T2	T7
<i>Wine Quality</i>	4898	12	7	T7	T3	T7

Com os testes realizados pode-se constatar que para configurações de máquinas de comitê com pelo menos um classificador de baixa qualidade, o resultado final tende para acurácia e precisão menores. Na combinação dos resultados um classificador de baixa qualidade acaba por induzir o conjunto de classificadores ao erro, pelo fato do mesmo escolher um resultado ruim quando há um impasse entre os outros classificadores.

Contudo, quando é utilizada uma configuração com algoritmos de alta qualidade, ou seja, que tem acurácia e precisão altas, o resultado da máquina de comitê é melhor do que os resultados dos classificadores individuais.

Uma consideração sobre estas avaliações feitas sobre os classificadores é a de que a escolha dos classificadores que compõem a máquina de comitê tem importância fundamental e decisiva no resultado final. Como os testes foram realizados com a composição de votação por maioria simples, supõe-se que os resultados serão ainda melhores no momento em que se fizer a escolha de classificadores com alta acurácia e precisão, aliada a uma configuração de votação por maioria ponderada em que os pesos estejam corretamente definidos.

Capítulo 5

Conclusões

As máquinas de comitê têm mostrado resultados eficientes na classificação de dados em diferentes áreas de aplicação, e isto se deve ao fato de a máquina de comitê ser muito abrangente e por sua definição ser simples, utilizando-se da abordagem de divisão e conquista.

Por poderem ser utilizadas de forma tão universal, as definições formais e aplicações de máquinas de comitê divergem ligeiramente de autor para autor, considerando que cada um aplica a definição em uma área específica.

O uso de máquina de comitê otimiza os resultados alcançados em comparação ao uso de somente um especialista. Nesse trabalho foram obtidos, por meio de avaliação experimental a partir de várias bases de dados, e também pela comparação com avaliações de trabalhos anteriores, melhores resultados de classificação.

Uma grande contribuição da formação de um comitê é que resulta principalmente em um aumento da confiança no resultado justamente por ser uma votação que considera mais de um especialista na geração do resultado. Pode-se dizer que o resultado de um comitê é um resultado consolidado. Ainda como resultado, cita-se a integração do módulo de geração de máquinas de comitê na ferramenta YADMT.

Como trabalhos futuros propõe-se a realização de testes mais aprofundados sobre o módulo implementado, a fim de permitir uma validação mais completa da ferramenta. Sobre o módulo atual uma melhoria seria o armazenamento dos treinamentos dos classificadores para futura composição no comitê, sem ser necessário novo treinamento de todos classificadores ao adicionar um novo classificador ao comitê. Ainda sobre este módulo, poderia ser feita a execução de cada classificador em uma *thread* (processo) separada que executariam de forma paralela e a composição dos resultados se daria no final de todos os processos.

Além das melhorias nos testes e no módulo atual de comitê, é necessária também a reestruturação do módulo de classificação da ferramenta YADMT a fim de permitir a implementação de outras técnicas de composição, ou seja, métodos que requerem saídas em valores contínuos, por exemplo, média e impulso. Outra implementação útil para a ferramenta seria um módulo de máquina de comitê dinâmica.

Apêndice A

Descrição das Bases de Dados

A seguir são descritas mais detalhadamente as bases de dados utilizadas.

A.1 Base de Dados *Iris*

A base de dados *Iris* foi criada por Ronald Aylmer Fisher em 1988. Contém dados de três tipos de flores conhecidas por *Iris: Setosa, Vesicolour e Virginica*. Além do atributo de classificação possui mais quatro atributos que trazem informações de tamanho e espessura da sépala e tamanho e espessura de pétala, com todas as dimensões dadas em centímetros.

Sabe-se que classe *Setosa* é linearmente independente das outras duas, que não o são entre si. A base de dados é composta por 150 amostras e a distribuição igualitária dos dados é dada pela Tabela A.1.

Tabela A.1: Distribuição de classes da Base de Dados *Iris*

Classes	Número de Exemplos
<i>Iris Setosa</i>	50
<i>Iris Versicolour</i>	50
<i>Iris Virginica</i>	50

A.2 Base de Dados *Acute Inflammations*

A base de dados *Acute Inflammations* foi criada por Jacek Czerniak, do Instituto de Pesquisa de Sistemas da Academia Polonesa de Ciências, e está disponível na UCI desde 2009. Apresenta dados criados para testar um sistema especialista, que deveria realizar o diagnóstico presumível de doenças do sistema urinário.

Originalmente esta base de dados possuía 8 atributos, sendo:

- 1 atributo numérico, indicando a temperatura do paciente;
- 7 categóricos (ocorrência de náuseas, dor lombar, necessidade constante de urinar, dores de micção, dores de uretra, ocorrência de inflamação urinária na bexiga e nefrite de origem pelve renal), que podem assumir valores "yes" e "no", destes, os dois últimos atributos representam atributos de decisão de classe baseado nos seis primeiros atributos.

Para este trabalho eliminou-se o último atributo classe (nefrite de origem pelve renal), ficando a base de dados com um total de 7 atributos, incluso o atributo que define a classe “inflamação urinária na bexiga”.

Esta base de dados contém um total de 120 amostras com a distribuição das classes, de acordo com a Tabela A.2.

Tabela A.2: Distribuição de classes da Base de Dados *Acute Inflammations*

Classes	Número de Exemplos
<i>Yes</i>	59
<i>No</i>	61

A.3 Base de Dados *Teaching Assistant Evaluation*

A base de dados *Teaching Assistant Evaluation* foi criada por Wei-Yin Loh do Departamento de Estatística da Universidade de Wisconsin - Madison em 1997. Consiste de dados da avaliação de desempenho de assistentes de ensino (TA - *Teaching Assistant*) designados no Departamento de Estatística da Universidade de Wisconsin-Madison.

É composta por 6 atributos numéricos, o primeiro atributo indica se o TA tem ou não o idioma Inglês como seu idioma nativo, o segundo atributo indica qual o instrutor do curso, o terceiro atributo indica o curso, o quarto atributo indica se é um semestre regular ou um semestre de verão, o quinto atributo indica o tamanho da turma, e o sexto atributo, que indica a pontuação do assistente como sendo baixa, média ou alta.

Esta base de dados contém 151 amostras e as classes estão distribuídas nas três classes de acordo com a Tabela A.3.

Tabela A.3: Distribuição de classes da Base de Dados *Teaching Assistant Evaluation*

Classes	Número de Exemplos
1 (pequena)	49
2 (média)	50
3 (grande)	52

A.4 Base de Dados *Ecoli*

A base de dados *Ecoli* foi criada por Kenta Nakai do Instituto de Biologia Celular e Molecular da Universidade de Osaka em 1996. Contêm dados sobre a localização de proteínas em células. Originalmente consistia de 1 atributo categórico, 7 atributos numéricos e 1 atributo que define a classe.

Esta base de dados contém 336 amostras distribuídas como na Tabela A.4.

Tabela A.4: Distribuição de classes da Base de Dados *Ecoli*

Classes	Número de Exemplos
cp (citoplasma)	143
im (membrana interna sem sinal de seqüência)	77
pp (periplasma)	52
imU (membrana interna, com sinal de seqüência não clivável)	35
om (membrana externa)	20
omL (membrana externa de lipoproteína)	5
imL (membrana interna de lipoproteína)	2
imS (membrana interna, com sinal de seqüência clivável)	2

A.5 Base de Dados *Abalone*

A base de dados *Abalone* foi criada por Sam Waugh do Departamento de Ciência da Computação da Universidade da Tasmânia, em 1994. Consiste de informações relacionadas a um molusco chamado *haliotis*, que serve de alimento principalmente nos países asiáticos. Os dados foram inicialmente propostos no estudo de Nash, Sellers, Talbot, Cawthorn e Ford (1994), que analisava os moluscos a partir das características físicas e determinava qual seria a idade do mesmo para o consumo, e posteriormente foram retiradas as tuplas que continham valores ausentes, e consiste de:

- 1 atributo categórico (sexo);

- 6 atributos numéricos (comprimento, diâmetro, altura, peso bruto, peso do molusco descascado, peso das vísceras, peso da concha);
- 1 atributo classificador (anéis), utilizado para determinar a idade do *haliote*.

A base de dados consiste de 4177 amostras e a distribuição dos dados segue na Tabela A.5.

Tabela A.5: Distribuição de classes da Base de Dados *Abalone*

Classes	Número de Exemplos
1	1
2	1
3	15
4	57
5	115
6	259
7	391
8	568
9	689
10	634
11	487
12	267
13	203
14	126
15	103
16	67
17	58
18	42
19	32
20	26
21	14
22	6
23	9
24	2
25	1
26	2
27	1
29	1

A.6 Base de Dados *Spambase*

A base de dados *Spambase* foi criada por MarkHopkins, Erik Reeber, George Forman e Jaap Suermondt do Laboratório da Hewlett-Packard (HP) em 1999. Esta base de dados representa dados de vários e-mails que foram classificados como spam ou não-spam. Considera-se spam e-mails não solicitados que fazem propagandas de produtos e de sites, esquemas de ganho rápido de dinheiro, e-mails de correntes, pornografia, entre outros.

A base de dados é composta de:

- 48 atributos numéricos contínuos que representam a porcentagem de palavras do e-mail que correspondem a uma determinada palavra;
- 6 atributos numéricos contínuos que representam a porcentagem de caracteres do e-mail que correspondem a determinado caractere;
- 1 atributo numérico contínuo que representa o tamanho médio de letras maiúsculas em sequência;
- 1 atributo numérico inteiro que representa o tamanho da maior sequência ininterrupta de letras maiúsculas;
- 1 atributo numérico inteiro que representa a soma de todas as sequências ininterruptas de letras em maiúsculas;
- além de um atributo numérico inteiro que define a classe do objeto como sendo ou não um spam.

Esta base de dados contém 4601 amostras e as classes estão distribuídas de acordo com a Tabela A.6.

Tabela A.6: Distribuição de classes da Base de Dados *Spambase*

Classes	Número de Exemplos
1 (spam)	1813
0 (não-spam)	2788

A.7 Base de Dados *LIBRAS movement*

A base de dados *LIBRAS Movement* foi criada por Daniel Baptista Dias, Sarajane Marques Peres e Helton Hideraldo Bísaro, da Universidade de São Paulo - USP em 2009. Esta base de dados contém informações de movimentos da mão que representam o tipo de sinal LIBRAS (Língua Brasileira de Sinais). É composta de 90 atributos numéricos que representam a

posição da mão em cada instante de tempo em um vídeo de 45 *frames*, além de um atributo categórico que define a classe do objeto, neste caso, um dos 15 tipos de movimento LIBRAS.

Esta base de dados é composta por 360 amostras com a igualitária distribuição das classes de acordo com a Tabela A.7.

Tabela A.7: Distribuição de classes da Base de Dados *LIBRAS movement*

Classes	Número de Exemplos
<i>curved swing</i>	24
<i>horizontal swing</i>	24
<i>vertical swing</i>	24
<i>anti-clockwise arc</i>	24
<i>clockwise arc</i>	24
<i>Circle</i>	24
<i>horizontal straight-line</i>	24
<i>vertical straight-line</i>	24
<i>horizontal zigzag</i>	24
<i>vertical zigzag</i>	24
<i>horizontal wavy</i>	24
<i>vertical wavy</i>	24
<i>face-up curve</i>	24
<i>face-down curve</i>	24
<i>Tremble</i>	24

A.8 Base de Dados *Statlog (Image Segmentation)*

A base de dados *Statlog (Image Segmentation)* foi criada pelo Vision Group da Universidade de Massachusetts em 1990. Contém 19 atributos numéricos contínuos que representam várias informações sobre segmentos de imagem de tamanho 3x3 *pixels*, mais um atributo numérico que define a classe do objeto, cuja informação representa um tipo de imagem (tijolo, céu, folhagem, cimento, janela, caminho e grama).

Esta base de dados contém 2310 amostras distribuídas conforme a Tabela A.8.

Tabela A.8: Distribuição de classes da Base de Dados *Statlog (Image Segmentation)*

Classes	Número de Exemplos
1	330
2	330
3	330
4	330
5	330
6	330
7	330

A.9 Base de Dados *Statlog (Landsat Satellite)*

A base de dados *Statlog (Landsat Satellite)* foi criada por Ashwin Srinivasan do Departamento de Estatística e Modelagem de Dados da Universidade de Strathclyde em 1993. Contém valores multiespectrais de segmentos de tamanho 3×3 *pixels* de uma imagem de satélite e a respectiva classificação associada com o pixel central de cada imagem. É composta por 36 atributos numéricos na faixa de 0 a 255, que representam as 4 bandas espectrais dos 9 *pixels* do segmentos de imagem, e 1 atributo numérico classificador do pixel central.

Esta base de dados é composta por 6435 amostras e as classes estão distribuídas conforme a Tabela A.9.

Tabela A.9: Distribuição de classes da Base de Dados *Statlog (Landsat Satellite)*

Classes	Número de Exemplos
1 - solo vermelho (<i>red soil</i>)	1533
2 - cultura de algodão (<i>cotton crop</i>)	703
3 - solo cinza (<i>grey soil</i>)	1358
4 - solo cinza com umidade (<i>damp grey soil</i>)	626
5 - solo com vegetação seca (<i>soil with vegetation stubble</i>)	707
6 - mistura de classes (<i>mixture class</i>)	0
7 - solo cinza com muita umidade (<i>very damp grey soil</i>)	1508

A.10 Base de Dados *Wine Quality*

A base de dados *Wine Quality* foi criada por Paulo Cortez, Antonio Cerdeira, Fernando Almeida, Telmo Matos e Jose Reis da Universidade de Minho, Guimarães - Portugal em 2009.

Esta base de dados contém informações físico-químicas da variação do vinho Português de cor vermelha. É composta por 11 atributos numéricos contínuos que representam as características físico-químicas do vinho (*fixed acidity*, *volatile acidity*, *citric acid*, *residual sugar*, *chlorides*, *free sulfur dioxide*, *total sulfur dioxide*, *density*, *pH*, *sulphates* e *alcohol*) e um atributo quantitativo que representa a qualidade do vinho.

Esta base de dados apresenta 4898 amostras cuja distribuição das classes consta na Tabela A.10.

Tabela A.10: Distribuição de classes da Base de Dados *Wine Quality*

Classes	Número de Exemplos
3	20
4	163
5	1457
6	2198
7	880
8	175
9	5

Referências Bibliográficas

ABLOW, C. M.; KAYLOR, D. J. A committee solution of the pattern recognition problem. In: IEEE Transaction on Information Theory, 11(3):453–455, 1965.

BARR, A.; FEIGENBAUM, E. The Handbook of Artificial Intelligence. 3.ed. Los Altos, California: William Kaufmann Inc., 1981.

BARBOSA, B. H. G. Computação Evolucionária e Máquinas de Comitê na Identificação de Sistemas Não-Lineares. Tese de Doutorado (Doutorado em Engenharia Elétrica) - UFMG, Belo Horizonte - MG, Outubro, 2009.

BATISTA, G. E. de A. P. A. Um Ambiente de Avaliação de Algoritmos de Aprendizado de Máquina Utilizando Exemplos. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) - USP - Universidade Estadual de São Paulo, São Carlos, SP, Setembro, 1997.

BENFATTI, E. W. Um Estudo sobre a Aplicação de k-NN, C4.5 e Naive Bayes na Classificação de Dados. Monografia - UNIOESTE, Cascavel - PR, Dezembro 2010.

BENFATTI, E. W.; GIRARDELLO, A. D.; BONIFÁCIO, F. N.; BOSCARIOLI, C. Descrição da Arquitetura e Projeto da Ferramenta YADMT - Yet Another Data Mining Tool. Cascavel: UNIOESTE, Janeiro. Relatório Técnico 1, 2011.

BISHOP, C. M.; Neural Networks for Pattern Recognition. 1. ed. New York: Oxford University Press Inc., 1995.

BONIFÁCIO, F. N. Comparação entre as Redes Neurais Artificiais MLP, RBF e LVQ na Classificação de Dados. Monografia – UNIOESTE, Cascavel - PR, Dezembro 2010.

BREIMAN, L. Bagging Predictors. In: Machine Learning, p. 123-140, 1996.

COSTA, E. de P. Investigação de técnicas de classificação hierárquica para problemas de bioinformática. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) - UFCG - Universidade Federal de Campina Grande, Campina Grande, PB, Março 2008.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMYTH, P. The KDD process for extracting useful knowledge from volumes of data, *Communications of the ACM*, v.39, n.11, p. 27-34, November, 1996.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. *Advances in Knowledge Discovery and Data Mining*. Cambridge, Massachusetts and London, England: AAAI/MIT Press, 1996. ISSN 0-262-56097-6.

FLORES, C. D. Fundamentos dos Sistemas Especialistas. In: BARONE, D. A. C. *Sociedades Artificiais: a nova fronteira da inteligência nas máquinas*. Porto Alegre: Bookman, 2003.

FRANK, A.; ASUNCION, A. *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science, 2010. Disponível em <http://archive.ics.uci.edu/ml>. Último acesso 16 de julho de 2012.

GUO, J.; LUH, P. B. Improving market clearing price prediction by using a committee machine of neural networks. In *IEEE Transactions on Power Systems*, p. 1867-1876, 2004.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 1-55860-901-6.

HANSEN, L. K.; SALAMON, P. Neural Network Ensembles. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, p. 993-1001, 1990.

HAYKIN, S. S. *Redes Neurais – Princípio e Prática*. 2. ed. São Paulo: Bookman, 2001.

HO-MAN, T. *Face Recognition Committee Machine: Methodology, Experiments and A System Application*. Tese – Universidade de Hong Kong, China. Julho, 2003.

KAMO, T.; DAGLI, C. Hybrid approach to the Japanese candlestick method for financial forecasting. In: *Expert Systems with Applications*, vol 36, p. 5023-5030, 2009.

MACEDO, J. A. *Máquina de Comitê para a previsão do preço médio diário das ações da bolsa de valores do estado de São Paulo*. Monografia - UNIFAL, Alfenas - MG. Julho, 2010

MADEO, R. C. B.; PERES, S. M.; BISCARO, H. H.; DIAS, D. B.; BOSCARIOLI, C. A. A committee machine implementing the pattern recognition module for fingerspelling applications. In: Proceedings of the 2010 ACM Symposium on Applied Computing, p. 954-958, 2010.

MAGALHAES, R. M. Uma investigação sobre a utilização de processamento paralelo no projeto de máquinas de comitê. Dissertação de Mestrado - UFRN, Natal – RN, Janeiro 2007.

PERRONE, M. P.; COOPER, L. N. When networks disagree ensemble methods for hybrid neural network. In Mammone, R. J., editor, Neural Networks for Speech and Image Processing, p. 126–142. Chapman Hall, 1993.

POLIKAR, R. Ensemble Based Systems in Decision Making. In: IEEE Circuits and Systems Magazine, vol. 6, no. 3, p. 21-45, 2006.

RAAFAT, H. M.; TOLBA, A. S.; SHADDAD, E. Committee machines for facial-gender recognition. In International Journal of Hybrid Intelligent Systems archive, IOS Press Amsterdam, p. 41-51, 2009.

RUSSEL, S. J.; NORVIG, P.; Inteligência Artificial – Uma abordagem moderna. 2.ed. Rio de Janeiro: Elsevier, 2004.

TAKIYAMA, T. A general method for training the committee machine. In Pattern Recognition, p. 255-259, Pergamon Press Ltd, 1978.

TANG, H.; LYU, M. R.; KING, I. Face recognition committee machines: Dynamic vs. static structures. In: Proceedings 12th International Conference on Image Analysis and Processing, p. 121-126, 2003.

TRESP, V. Committee machines. Chapter in Handbook for Neural Network Signal Processing, Yu Hen Hu and Jenq-Neng Hwang (eds.), CRC Press, 2001.

VILLANUEVA, W. J. P. Comitê de Máquinas em Predição de Séries Temporais. Tese de Mestrado - UNICAMP, Campinas - SP, Outubro 2006.

WEI, L.; YANG, Y.; NISHIKAWA, R. M.; JIANG, Y.; A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. In IEEE Trans Med Imaging. p.371-380, 2005.

WEKA. 2012. Software. Disponível em: <<http://www.cs.waikato.ac.nz/#ml/weka/>>. Acesso em: 18/10/2012.

WITTEN, I. H.; FRANK, E. Data Mining: Practical Machine Learning Tools and Techniques. 2. ed. San Francisco, CA, USA: Morgan Kaufmann, 2005. ISBN 0-12-088407-0.

YANG, J. Hybrid committee machine for incremental learning. In: International Conference on Neural Networks and Brain, p. 391-395. 2005.

ZHAO, Z. Q.; HUANG, D. S.; SUN, B.Y; Human face recognition based on multi-features using neural networks committee. In: Pattern Recognition Letters 25, p. 1351–1358, Elsevier, 2004.

