Inteligência Computacional Aplicada à Análise de Risco no Contexto do Tratado da Basiléia

Laudo Reis
lalazius@gmail.com

Roseli Aparecida Francelin Romero
rafrance@icmc.usp.br

Departamento de Ciências de Computação e Estatística
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo — Campus de São Carlos
Caixa Postal 668 - 13560-970 - São Carlos, SP.

Resumo:

No contexto de Gestão de Risco Financeiro Internacional, o Tratado de Basiléia II aperfeiçoa as regras para minimização das possibilidades de crises sistêmicas no sistema financeiro. Pela exigência da utilização de mecanismos de classificação e controle do risco de crédito e da reserva mínima de recursos financeiros de acordo com as classes de risco presentes na carteira de crédito da instituição financeira, estabelece sua política de mitigação desse risco. O tratado aperfeiçoa o requisito da reserva mínima permitindo aos bancos o desenvolvimento de sistemas internos de classificação e estimativa de risco, cuja modelagem, envolve o cálculo de probabilidades das ocorrências de atraso ou não recebimento de valores contratados.

É comum, nas abordagens clássicas, o uso de aproximações estatísticas utilizando modelos paramétricos e não-paramétricos: análise de regressão, análise discriminante, regressão Logit e regressão Probit. Diversos pesquisadores têm investigado o uso de técnicas de inteligência computacional: Redes Neurais, Árvores de Decisão, Lógica Difusa e Máquina de Suporte de Vetores para a identificação de padrões de características dos tomadores de crédito e classificação de risco.

Este relatório, inserido nesse contexto, tem por objetivo efetuar uma breve revisão das pesquisas realizadas com Inteligência Computacional (IC) e, realizar um estudo comparativo entre as técnicas escolhidas, na classificação de Dados Financeiros disponíveis.

Palavras chaves: Basiléia, IRB, Inteligência Computacional, Análise de Risco, Classificação de Risco, MLP, SVM.

Dezembro de 2008

Sumário

1. Introdução 5
1.1. Contextualização e Motivação 5
1.2. Ferramentas Utilizadas 8
1.3. Descrição dos Dados Utilizados 9
1.4. Objetivos do Relatório 10
1.5. Organização do Relatório 10
2. Revisão de Técnicas em Inteligência Computacional 11
2.1. Paradigma Conexionista 11
2.2. Paradigma Evolucionário 16
2.3. Árvores de Decisão 18
2.4. Máquina de Suporte de Vetores SVM 20
2.5. Medidas de Avaliação 24
3. Experimentos Realizados 27
3.1. Carga dos dados 27
3.2. Projeto dos Experimentos: Diretrizes 27
3.3. Pré-processamento dos Dados 29
3.4. Estrutura Geral dos Experimentos 29
3.5. Experimentos com RNA MLP 32
3.6. Experimentos com Árvores de Decisão 34
3.7. Experimentos com SVM 36
3.8. Experimentos SVM com Otimização Genética 36
3.9. Experimentos com Amostra Balanceada 38
4. Discussão dos Resultados e Trabalhos Futuros 40
4.1. Discussão dos Resultados Obtidos 40
4.2. Proposta para Trabalhos Futuros 42
5. Referências Bibliográficas 43

Índice de Figuras

Figura 1.1.1 Índice de Basiléia do SFN Fonte: Bacen 2008.....7
Figura 2.1.1 Modelo de Neurônio Artificial.....11
Figura 2.1.2 Efeito da entrada fixa bias no hiperplano.....12
Figura 2.1.3 Topologias de RNA representativas.....13
Figura 2.1.4 Topologias das RNA utilizadas por Angelini et al.....15
Figura 2.2.1 Topologia das RNA RBF.....17
Figura 2.3.1 Árvore de Decisão.....18
Figura 2.4.1 Hiperplano de máxima separação por SVM.....20
Figura 2.4.2 Linearização do espaço de entrada.....20
Figura 2.4.3 Distribuição de artigos por ano. Fonte: Yu 2008.....22
Figura 2.5.1 Matriz de confusão utilizada para definição da curva ROC.....26
Figura 3.2.1 Histograma das classes de risco do universo e amostra.....28
Figura 3.2.2 Exemplo de consulta SQL.....29
Figura 3.4.1 Estrutura geral dos experimentos com RapidMiner.....30
Figura 3.4.2 Estrutura geral na forma de arquivo XML.....30
Figura 3.4.3 Representação gráfica de uma matriz de confusão.....31
Figura 3.5.1 RNA Perceptron Multicamadas.....32
Figura 3.5.2 Matriz de confusão do experimento MLP 3.....32
Figura 3.6.1 Matriz de confusão, árvore e curva ROC universo.....35
Figura 3.6.2 Matriz de confusão, árvore e curva ROC amostra.....35
Figura 3.8.1 Modelo para os experimentos 17 a 20.....37
Figura 3.9.1 Amostra balanceada.....38
Figura 4.1.1 Acurácia das técnicas nos subconjuntos de classes.....41
Figura 4.1.2 Medidas de F-measure com α=0,5 das técnicas nas classes.....42

Índice de Tabelas

Tabela 1.1.1 Atividades desenvolvidas pelo Comitê de Basiléia em Supervisão Bancária 6
Tabela 1.1.2 Reserva de recursos por classe de risco. Fonte CMN 1999..... 7
Tabela 1.3.1 Caráter mutuamente exclusivo dos subfatores 9
Tabela 2.1.1 Cronologia no desenvolvimento de redes neurais e recentes pesquisas 14
Tabela 2.4.1 Comparativo entre técnicas de inteligência computacional..... 22
Tabela 2.4.2 Técnicas utilizadas nos artigos analisados por Yu. 22
Tabela 2.5.1 Matriz de confusão 25
Tabela 3.4.1 Planejamento de experimentos com amostra..... 31
Tabela 3.5.1 Matrizes de confusão do experimento 3 com MLP. 33
Tabela 3.5.2 Resultados dos experimentos MLP 1 a 7..... 33
Tabela 3.5.3 Características do experimento 16 com MLP..... 33
Tabela 3.5.4 Matriz de confusão do experimento 16 com MLP. 33
Tabela 3.6.1 Matrizes de confusão dos experimentos 8 com AD. 34
Tabela 3.6.2 Resultados dos experimentos AD..... 34
Tabela 3.7.1 Matriz de confusão do experimento 12 36
Tabela 3.7.2 Resultados dos experimentos SVM..... 36
Tabela 3.8.1 Experimentos com SVMAE 37
Tabela 3.8.2 Resultados dos experimentos SVM com otimização evolutiva..... 37
Tabela 3.9.1 Experimentos realizados com a amostra balanceada..... 38
Tabela 3.9.2 Matriz de confusão do experimento 24 com SVMAE..... 39
Tabela 3.9.3 Resultados obtidos nos experimentos 21 a 36. 39

1. Introdução

1.1. Contextualização e Motivação

As empresas existem em função de seus objetivos empresariais [Oldcorn e Parker 1998]. Em suas operações as empresas comercializam produtos e serviços sujeitas a Lei da Oferta e Procura, enfrentam concorrentes, são submetidas às leis e regulamentações governamentais [Chiavenato 2001].

A evolução tecnológica e, sobretudo a expansão da Internet, aumentaram o volume de informações e a rapidez de sua disseminação, contribuindo substancialmente para o movimento de Globalização [Yip e Dempster 2005]. Sistemas de negociação e transações eletrônicas permitem que negócios sejam realizados a partir de diferentes regiões geográficas de modo praticamente instantâneo.

Dos vários objetivos empresariais, destaca-se, para empresas com fins lucrativos, a maximização do retorno financeiro com a minimização dos riscos associados [Oldcorn e Parker 1998]. As empresas financeiras de um modo geral e os bancos especificamente buscam a realização desse objetivo empresarial.

Na comercialização dos produtos de crédito, os bancos captam recursos de agentes superavitários e entregam aos agentes deficitários, cobrando diferentes taxas de juros, assumindo para si os riscos e vislumbrando os lucros conforme os agentes deficitários honrem com os pagamentos.

Na linguagem comercial, os agentes deficitários que pagam em dia são denominados adimplentes e os que pagam fora do prazo ou não pagam são denominados inadimplentes [Borth 2007].

O risco inerente às operações de crédito, ou seja, o risco de inadimplência ocorre de diferentes formas: pela pré-disposição do agente não identificada pelo banco, por eventos adversos fora do controle do agente, por eventos conjunturais do país, por erro do banco na análise do risco.

As conseqüências da inadimplência para o sistema de financeiro, vão desde um aumento pontual da taxa de juros cobrada nas operações de uma instituição específica até a desestabilização do sistema financeiro de diversos países, também conhecido como crise sistêmica [Rochet 2007]. A partir desse ponto ocorrem danos à chamada economia real, aumentando a probabilidade de quebra de bancos, corrida de depositantes para retirada de recursos do sistema financeiro e retração dos agentes superavitários em fuga ao risco.

A estabilização do sistema financeiro ocorre em diferentes níveis: internamente os bancos têm interesses financeiros no bom funcionamento de seus negócios, os governos através de suas instituições buscam a estabilidade macro e microeconômica e em termos globais tratados e acordos são assinados de maneira que se atinja estabilidade em nível mundial.

O Banco Internacional de Liquidação, organismo que congrega instituições financeiras, bancos centrais ao redor do mundo, através de seu Comitê de Basiléia em Supervisão Bancária, tem desenvolvido atividades dentro da política de mitigação de risco sistêmico no sistema financeiro conforme Tabela 1.1.1.

Tabela 1.1.1 Atividades desenvolvidas pelo Comitê de Basiléia em Supervisão Bancária

Atividades desenvolvidas pelo Comitê de Basiléia para Mitigação de Risco Sistêmico	
Jul 2008	Proposta de revisão do Acordo de Basiléia II em risco de mercado
Jul 2008	Linhas gerais para computar o Risco Incremental da Carteira de Negócios
Out 2007	Documento consultivo para computar o Risco Incremental da Carteira de Negócios
Out 2006	Risco ponderado na imunização financeira internacional
Set 2006	Testes de uso de Sistemas Internos de Classificação: Contexto e Implementação
Jun 2006	Basiléia II: Convergência Internacional de Padrões de Capital e Medição de Capital: uma Revisão Abrangente
Mar 2006	Uso de Sistemas Internos de Classificação adquiridos
Nov 2005	Basiléia II: Convergência Internacional de Padrões de Capital e Medição de Capital: uma Revisão
Abr 2005	Carteira de negócios: um Sumário de Respostas
Jun 2004	Trabalho conjunto do Comitê de Basiléia e da Comissão da Organização Internacional de Seguros: um trabalho de Tratamento Prudencial em Carteira de Negócios
Jun 2004	Basiléia II: Convergência Internacional de Padrões de Capital e Medição de Capital: uma Revisão
Abr 2003	Basiléia II: o Novo Acordo de Capital de Basiléia - Terceiro Documento Consultivo
Set 2001	Atualização dos trabalhos do Novo Acordo de Capital de Basiléia
Jan 2001	Basiléia II: o Novo Acordo de Capital de Basiléia - Segundo Documento Consultivo
Nov 1999	Atualização dos trabalhos para uma Nova Adequação de Capital
Jul 1988	Convergência Internacional de Medição de Capital e Padrões de Capital

Fonte: BCBS (2008) http://www.bis.org/list/bcbs/tid_22/index.htm.

Em 2005, o Comitê de Basiléia em Supervisão Bancária, revisou as recomendações estabelecidas no acordo de 1988. Conhecida como Tratado de Basiléia II, essa revisão aprimorou os três pilares básicos da política de mitigação de risco sistêmico: Requisito Mínimo de Capital, Supervisão e Disciplina [BCBS 2005].

O Brasil é signatário dos Tratados de Basiléia, estabelecendo suas políticas, através do Conselho Monetário Nacional (CMN) e do Banco Central do Brasil (BACEN); a classificação das operações de crédito e as regras para constituição de provisão para créditos de liquidação duvidosa, estão estabelecidas na Resolução BACEN número 2682, de 21 dezembro de 1999.

A inovação e aperfeiçoamento, introduzidos em 2005, no cálculo do requisito mínimo de capital, consistem: da admissão de diferentes percentuais na provisão de recursos em função da classe de risco, anteriormente fixado em pelo menos 8% para todas as instituições[1] [Mishkin 2004]; e da recomendação para o desenvolvimento de sistemas internos de classificação e controle de risco, o que permite maior aderência da classificação em relação às características da carteira de crédito gerida pelo banco.

De acordo com o Relatório de Estabilidade Financeira, emitido pelo Banco Central em maio de 2008, o Brasil tem apresentado Índices de Basiléia superiores ao mínimo, fixado para o país, em 11%, ainda que não haja formalmente referência a esse índice nas regulamentações. Esse índice expressa uma relação entre os valores patrimoniais: apresentado e exigido da instituição frente ao seu grau de risco, incentivando as instituições a manter reservas de recursos próprios suficientes para minimizar os riscos de perdas, inesperadas e relevantes à saúde financeira da instituição [Bacen 2008]. A Figura 1.1.1 apresenta o comportamento do índice para o Sistema Financeiro Nacional[2] e a Tabela 1.1.2 apresenta os percentuais e reserva de recursos de acordo com as classes de risco, estabelecidos pela Resolução BACEN número 2682 [CMN 1999].

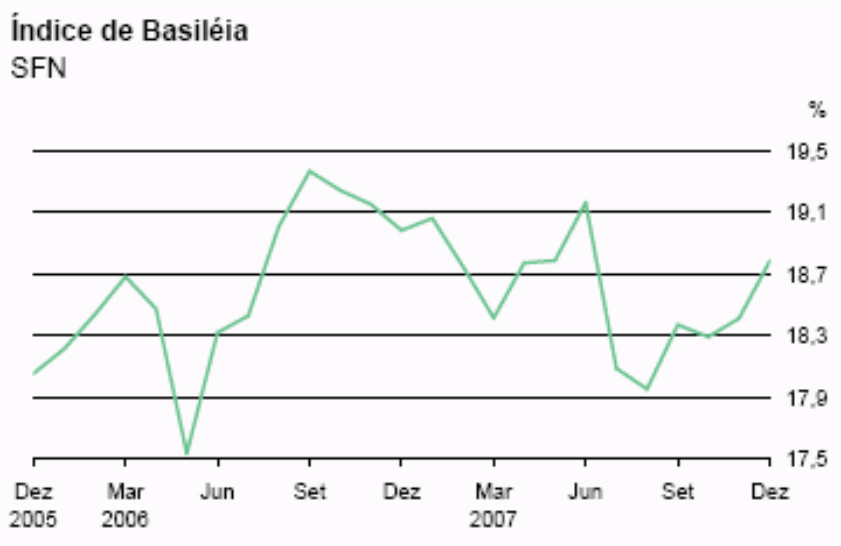

Figura 1.1.1 Índice de Basiléia do SFN Fonte: Bacen 2008

Tabela 1.1.2 Reserva de recursos por classe de risco. Fonte CMN 1999

Classe de Risco	**AA**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Percentual(%)	0	0,5	1	3	10	30	50	70	100

[1] Instituições que se mostrarem bem geridas, poderão apresentar requisito mínimo de capital inferior, conforme calculado de sua carteira de crédito.
[2] Valor médio apresentado para as instituições financeiras brasileiras

A modelagem de sistemas internos de classificação e controle de risco envolve o cálculo de probabilidades de inadimplência. É comum, nas abordagens clássicas, o uso aproximações estatísticas utilizando modelos paramétricos e não-paramétricos, dos quais podemos citar: análise de regressão, análise discriminante, regressão Logit e regressão Probit. Diversos pesquisadores têm investigado o uso de técnicas de inteligência computacional: Redes Neurais Artificiais (RNA), Árvores de Decisão (AD), Lógica Difusa (LD), Máquinas de Vetores de Suporte (SVM), para a identificação de padrões de características dos tomadores de crédito e classificação de risco nas operações realizadas.

Esses estudos são motivados, sobretudo, por três aspectos essenciais: com o desenvolvimento do mercado de crédito, tanto a complexidade, quanto o volume de operações tem crescido; tradicionalmente os sistemas de classificação baseiam-se em sistemas de escores, onde notas são atribuídas aos clientes e suas operações, tais sistemas envolvem a avaliação por especialistas humanos, sujeitos a subjetividade; com o uso das técnicas de inteligência computacional, há redução de custos, aumento de produtividade, precisão e flexibilidade na operacionalização de mudanças na estratégia de concessão de crédito [Yu et al., 2008].

1.2. Ferramentas Utilizadas

Algumas ferramentas computacionais livres e de uso acadêmico, relacionadas com mineração de dados foram utilizadas nas análises.

O Weka (Waikato Environment for Knowledge Analysis) é um software de domínio público que consiste em um conjunto de algoritmos de aprendizado de máquina [Weka 2008]. O Weka é implementado em linguagem Java, podendo ser executado em diferentes sistemas operacionais, além de possuir os benefícios inerentes a uma linguagem orientada a objetos: polimorfismo, encapsulamento, reutilização de código. Possui um livro publicado, pelo autor, explicando seus conceitos e uso [Witten e Frank 2005].

YALE (Yet Another Learning Environment), atualmente denominado RapidMiner é um software distribuído sob diferentes licenças: uma versão GPL (Licença Geral Pública – software livre) sem interface gráfica, uma versão simplificada e sem suporte e integração e, uma versão sob licença proprietária e comercial com interface gráfica, recursos para integração com outros sistemas e possuindo suporte e assistência técnica [Mierswa et al., 2006] [RapidMiner 2008].

Utilizou-se o banco de dados PostgreSQL versão 8, um software livre, gerenciador de banco de dados (SGBD) [Postgresql 2008], com o conector Java JDBC, postgresql-8.3-603.jdbc4.jar [JDBC 2008], para ligação das ferramentas Weka e RapidMiner.

Para a execução das ferramentas de análise, do banco de dados utilizou-se uma estação de trabalho convencional: microcomputador com núcleo duplo, compatível com x86, de 4,2 GHz, 4 Gbytes de memória RAM DDR2, e armazenamento serial ATA 500 Gbytes.

1.3. Descrição dos Dados Utilizados

Os dados utilizados neste trabalho, são dados reais e foram fornecidos por uma instituição brasileira de fomento de desenvolvimento regional, compreendem características de tomadores de crédito e de suas operações. Para que a instituição e os perfis de seus clientes fossem preservados, omitiu-se: o nome da instituição, os rótulos dos atributos característicos dos tomadores de créditos e as fórmulas para cálculo dos escores, utilizando-se as notas de classificação já atribuídas pela instituição.

Os dados inicialmente em um formato de banco de dados proprietário foram migrados para um formato SQL em texto plano. Selecionou-se para os experimentos, a Tabela mais representativa em termos de números de clientes; essa Tabela contém dados de empresas comerciais tomadoras de crédito, apresentando 120 atributos e 58995 registros.

As características dos tomadores são divididas em fatores e subfatores de risco. Os subfatores são mutuamente exclusivos entre si, se um subfator que apresenta o valor lógico 1, os demais subfatores que compõe o fator terão valor lógico 0, conforme exemplificado na Tabela 1.3.1.

Tabela 1.3.1 Caráter mutuamente exclusivo dos subfatores

Fator A				*Fator B*		
Subfatores				*Subfatores B*		
A1	*A2*	*A3*	*A4*	*B1*	*B2*	*B3*
0	0	0	1	0	1	0
0	0	0	1	0	1	0
0	0	0	1	0	1	0
0	0	0	1	0	1	0

1.4. Objetivos do Relatório

O presente trabalho está inserido em um contexto amplo de pesquisa em Inteligência Computacional e mais especificamente de Computação Bioinspirada (BioCom), do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo. As principais linhas de pesquisa estão relacionadas ao Aprendizado de Máquina, Computação Evolutiva, Redes Neurais, Sistemas Dinâmicos, Sistemas Inteligentes, Robótica Inteligente, BioInformática e Mineração de Dados, Aquisição de Conhecimento, Sistemas Baseados em Conhecimento, Sistemas *Fuzzy* e Sistemas Inteligentes Híbridos. Alguns trabalhos relacionados à análise de risco de crédito foram publicados e são citados no Capítulo 2.

Este trabalho tem por objetivo efetuar uma revisão das pesquisas realizadas com Inteligência Computacional e, realizar um estudo comparativo entre as técnicas escolhidas, na classificação dos Dados Financeiros disponíveis.

1.5. Organização do Relatório

No Capítulo 2, é apresentada uma revisão das principais técnicas de Inteligência Computacional investigadas e das pesquisas realizadas com Inteligência Computacional em Análise de Risco de Crédito.. São descritas as seguintes técnicas: Árvores de Decisão, Redes Neurais Artificiais e as Máquinas de Vetores de Suporte.

No Capítulo 3 são apresentados os experimentos para classificação da base de dados financeiros, juntamente com uma análise comparativa das técnicas utilizadas.

No Capítulo 4 é apresentada uma discussão sobre os resultados obtidos e trabalhos futuros.

2. Revisão de Técnicas em Inteligência Computacional

Para o pesquisador Munakata, a Inteligência Artificial (IA) está para o raciocínio assim como a Revolução Industrial para a força física e divide-se fundamentalmente em duas abordagens: uma tradicional, dominante, simbólica, de alto nível de abstração e macroscópica e outra numérica, de baixo nível, microscópica com ênfase em fisiologia e genética. A abordagem microscópica contempla Redes Neurais Artificiais e Algoritmos Evolutivos. Em adição a esta divisão, técnicas relativamente novas buscam aumentar a flexibilidade no processamento inteligente: Lógica Difusa, Teoria dos Conjuntos Nebulosos e Sistemas Caóticos, compondo com as técnicas microscópicas a Inteligência computacional (IC) [Munakata 2008].

A inteligência computacional permite aos sistemas inteligentes apresentarem características inerentes ao raciocínio humano: solução de problemas complexos, extração de conhecimento dos dados, aprendizado por exemplos, generalização do conhecimento aprendido, processamento de informação incerta e imprecisa, explicação das decisões tomadas, adaptação e evolução das soluções por otimização de parâmetros e pesquisa paralela. Os principais paradigmas de modelagem da IC, segundo Jain et al., são: Conexionista, Lógica Difusa, Evolucionário e Híbrido [Jain et al., 2008].

2.1. Paradigma Conexionista

As RNA representam 65 anos de pesquisas em IC conexionista, tendo como marco inicial as pesquisas McCulloch e Pitts em 1943, sobre a modelagem matemática de um neurônio artificial [Haykin 2001]. Na Figura 2.1.1 é apresentado um modelo esquemático de neurônio artificial.

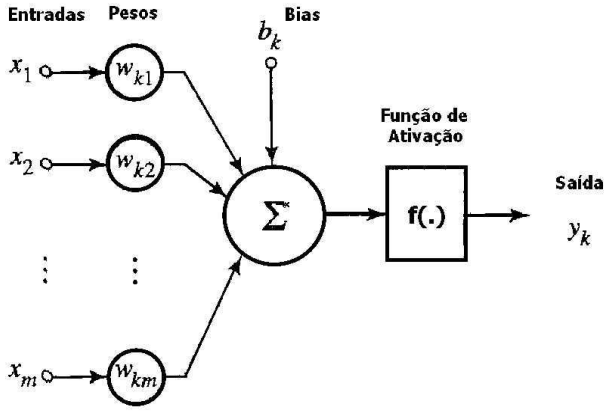

Figura 2.1.1 Modelo de Neurônio Artificial.

A saída **y** de cada neurônio individual é calculada por uma função de ativação **f** aplicada às entradas $\mathbf{x_i}$ ponderadas por seus pesos $\mathbf{w_k}$ e bias $\mathbf{b_k}$, conforme equação 2.1.1.

$$y := f\left(\sum_{i=1}^{n}(x_i\, w_{ki}) + b_k\right) \tag{2.1.1}$$

As funções de ativação mais comuns são a sigmóide e a tangente hiperbólica, apresentadas nas equações 2.1.2 e 2.1.3.

$$f(v) := \frac{1}{1 + \mathbf{e}^{(-a\,v)}} \tag{2.1.2}$$

$$f(v) := \tanh(v) \tag{2.1.3}$$

Individualmente um neurônio é capaz de trabalhar com problemas linearmente separáveis. A entrada fixa bias é utilizada para deslocar o hiperplano de separação, conforme mostra a Figura 2.1.2, melhorando essa separação.

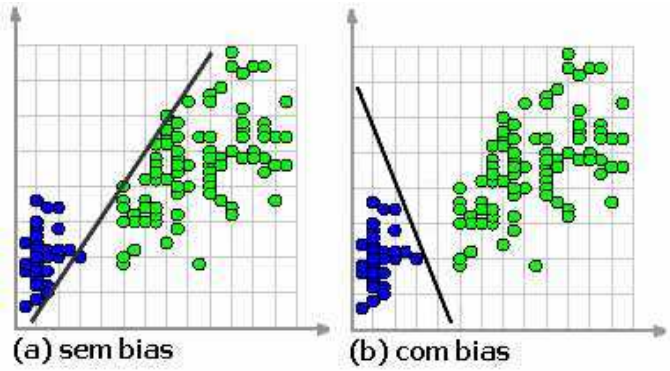

Figura 2.1.2 Efeito da entrada fixa bias no hiperplano

Neurônios artificiais interconectados formam redes e imitam o funcionamento do cérebro. Formalmente são sistemas de processamento paralelo e distribuído não-lineares. O conceito de aprendizagem através de ajustes nos pesos das conexões foi introduzido por Hebb em 1949. Os pesos das conexões, inicialmente aleatórios são ajustados automaticamente, através de um algoritmo de aprendizado [Haykin 2001].

As topologias das RNA definem: a maneira com que seus neurônios individuais são conectados, a existência ou não de realimentações, a disposição das camadas, caso haja mais de uma e, o número de neurônios.

Diversas topologias de redes foram desenvolvidas e estudadas no decorrer desses anos de pesquisa. A introdução de camada intermediária resolveu a limitação do neurônio individual em trabalhar apenas com problemas linearmente separáveis. A Figura 2.1.3 apresenta algumas topologias representativas de RNA.

Segundo Haykin, uma das arquiteturas de RNA mais utilizadas é a de topologia Perceptron Multicamadas (MLP), com algoritmo de Retropropagação (*Backpropagation*) [Haykin 2001], em conseqüência, dessa constatação, em termos de paradigma conexionista, este trabalho se limitará a essa arquitetura.

O Algoritmo de Retropropagação apresentado a seguir, utiliza a regra delta generalizada para atualização dos pesos da rede, conforme equações 2.1.4. e 2.1.5.

$$\Delta w_{kj}(n+1) = \alpha w_{kj}(n) + \eta \delta_j y_j \tag{2.1.4}$$

$$w(n+1) = w(n) + \Delta w_{kj}(n) \tag{2.1.5}$$

α - é a constante de momentum, quando $\alpha = 0$,
esta função funciona como a regra delta comum;
η - é a taxa de aprendizado;
δj - é o erro da unidade;
yj - é a saída produzida pela unidade j

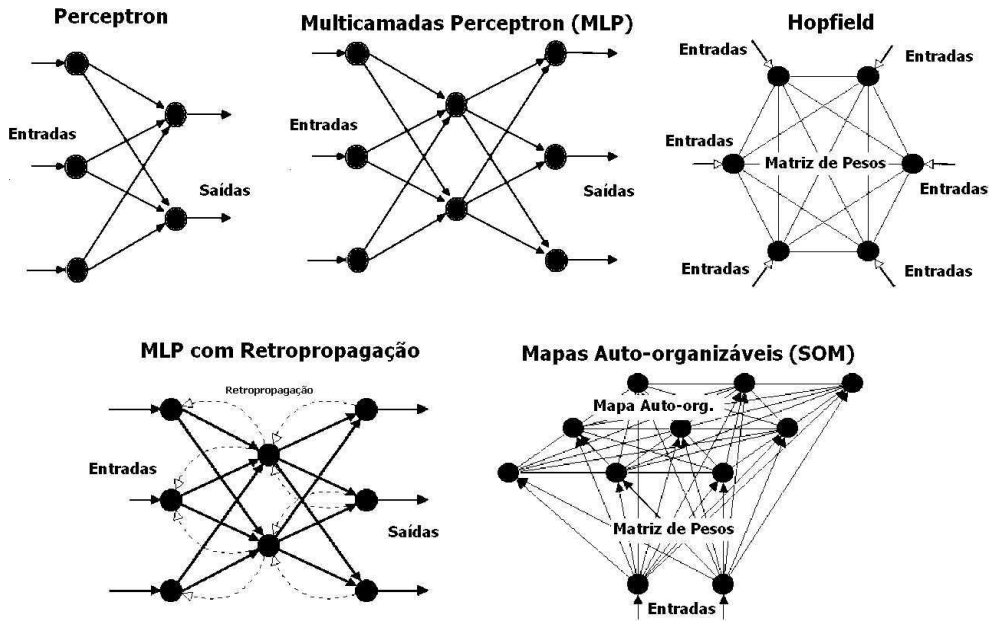

Figura 2.1.3 Topologias de RNA representativas.

Algoritmo Retropropagação[Haykin 2001]

```
1: Iniciar pesos da rede com valores aleatórios;
2: define valor limite para erro
3: repita
4:      Inicia erro total = 0;
5:      para todo padrão de treinamento faça
6:              para todo camada i da rede, i = 1, 2, ..., n faça
7:                      para todo  nó nij da i-ésima camada faça
8:                              Calcular saída produzida do nó;
9:                      fim-para
10:             fim-para
11:             calcula erro = saída desejada - saída produzida;
12:             para todo cada camada i da rede, i = n, n − 1, ..., 1 faça
13:                     para todo nó nij da i-ésima camada faça
14:                     Ajustar pesos do nó;
15:                     fim-para
16:             fim-para
17:             calcula erro total = erro total + erro;
18:     fim-para
19: até que erro total > valor limite para erro
```

Algumas das limitações das RNA são: o conhecimento está embutido nos pesos das conexões entre neurônios, o sistema é do tipo caixa-preta, na busca da solução esta pode convergir para um máximo (ou mínimo) local, o projeto da rede apresenta aspectos experimentais e exige experiência do projetista [Haykin 2001]. Estudos com sistemas híbridos buscam a solução dessas limitações.

A Tabela 2.1.l apresenta uma cronologia do desenvolvimento em Redes Neurais.

Tabela 2.1.1 Cronologia no desenvolvimento de redes neurais e recentes pesquisas

Ano	Rede
1943	Modelagem matemática do neurônio
1949	Aprendizado Hebbiano - Introdução do aprendizado pelo ajuste dos pesos
1958	Perceptron - Rosemblatt
1960	Rede Adaline - Widrow e Hoff
1982-1984	Rede de Hopfield - Hopfield
1982-1988	Mapas auto-organizáveis - Kohonen
1986	Multicamadas de Perceptrons (MLP) - Rumelhart et al.
1989	Redes de Função de Base Radial (RBF) -Moody e Darken
1987	Redes de Ressonância Adaptativas - Carpenter e Grossberg
2006	Rede Adiante Baseada na função de Lyapunov - Behera et al.
2007	Rede Neural Probabilistica ponderada - Song et al.
2007	Politipo ARTMAP - Amorin et al.
2007	GRNN de Densidade Dirigida - Goulermas et al.
2008	Rede RBF auto-organizável - Lian et al.
2008	Rede Neural de Hopfield Atrasada - Mou et al.

Segundo Jain, as pesquisas em termos de RNA podem ser categorizadas em duas principais linhas: uma de modelagem matemática do sistema nervoso em nível microscópico de neurônios e sinapses e outra de desenvolvimento de algoritmos de aprendizagem de máquina (*Machine Learning*) em nível macroscópico em termos de imitar certas funções do funcionamento do cérebro[Jain et al., 2008].

Pesquisas Realizadas com RNA MLP

Os pesquisadores Angelini et al. discutiram o uso de RNA (Redes Neurais Artificiais) em sistemas internos de classificação de risco de crédito, com dados reais de pequenas e médias empresas italianas [Angelini et al., 2007].

Foram testadas duas topologias de rede, uma totalmente e outra parcialmente conectada, com duas camadas ocultas e utilizando-se do paradigma de aprendizado supervisionado e algoritmo retro propagação. As topologias são apresentadas na Figura 2.1.4.

Os resultados mostraram-se promissores, com classificação correta de entradas e muito baixa taxa de erro. Semelhante a outros trabalhos, as entradas consistiam de razões econométricas e, a saída uma classificação binária em adimplentes e inadimplentes. As tarefas de análise e pré-processamento dos dados de entrada se mostraram fundamentalmente importantes para o sucesso dos experimentos, sobretudo em termos de tratamento de dados faltantes e normalização de valores. Os pesquisadores observaram experimentalmente a aplicabilidade de RNA para o problema de classificação de risco de crédito, as redes classificaram corretamente as entradas com uma taxa de erro muito baixa. Angelini et al. propõe estudos futuros para: o desenvolvimento de metodologia em técnicas procedimentais de análise de dados e de otimização de parâmetros, julgados fundamentais para o sucesso de experimentos e a investigação da aplicabilidade de RNA recorrentes.

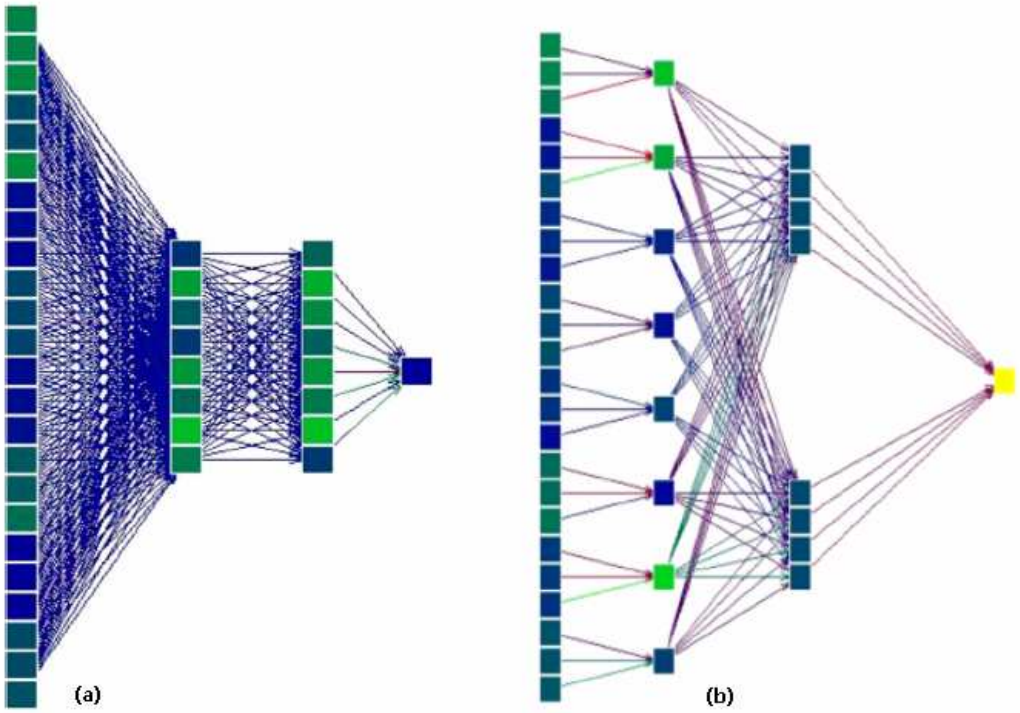

Figura 2.1.4 Topologias das RNA utilizadas por Angelini et al.

2.2. Paradigma Evolucionário

O paradigma evolucionário abrange a modelagem computacional de genética e evolução herdados da Biologia; os principais conceitos envolvidos são: seleção, reprodução e mutação. Os sistemas procuram as melhores soluções baseadas em adaptações de soluções existentes, de acordo com as variações ambientais e interações entre indivíduos.

Algoritmo Genético [Goldberg1989]

```
1: inicio
2: estabelecer critério de parada da evolução
3: escolher  população inicial
4: repita
5:      para todo indivíduo da população;
6:              selecionar indivíduos mais aptos;
7:                      para indivíduos selecionados aplicar operadores genéticos
8:                      fim-para
9:              avaliar aptidão dos indivíduos;
10:     fim-para
11: até  critério de parada seja atingido
12: fim do algoritmo
```

Pesquisas Realizadas com RNA RBF e Algoritmos Genéticos

"Evolutionary Radial Basis Functions for Credit Assessment", um artigo publicado em 2005 através da cooperação de pesquisadores do ICMC-USP e UFPE, discute o projeto de RNA do tipo RBF através de algoritmos genéticos para classificação crédito [Carvalho et al., 2005].

Carvalho et al. apresentam uma abordagem para projeto de RNA do tipo RBF (*Radial Basis Functions*) através do uso de algoritmos genéticos, discute técnicas relacionadas presentes na literatura e realiza experimentos com a abordagem proposta. Na Figura 2.2.1 é apresentada a topologia de utilizada, a camada oculta é compostas por funções $h_i(x)$ de base , sendo as mais comuns as Gaussianas, apresentadas na forma da equação 2.2.1 e, a saída por uma combinação linear dessas funções, dada pela equação 2.2.2

$$h_i(\mathbf{x}) = \exp \frac{-\|\mathbf{x} - \mathbf{c}_i\|^2}{2\sigma_i^2} \tag{2.2.1}$$

$$\hat{f}(x) = \sum_{i=1}^{m} w_i h_i(\mathbf{x}) \tag{2.2.2}$$

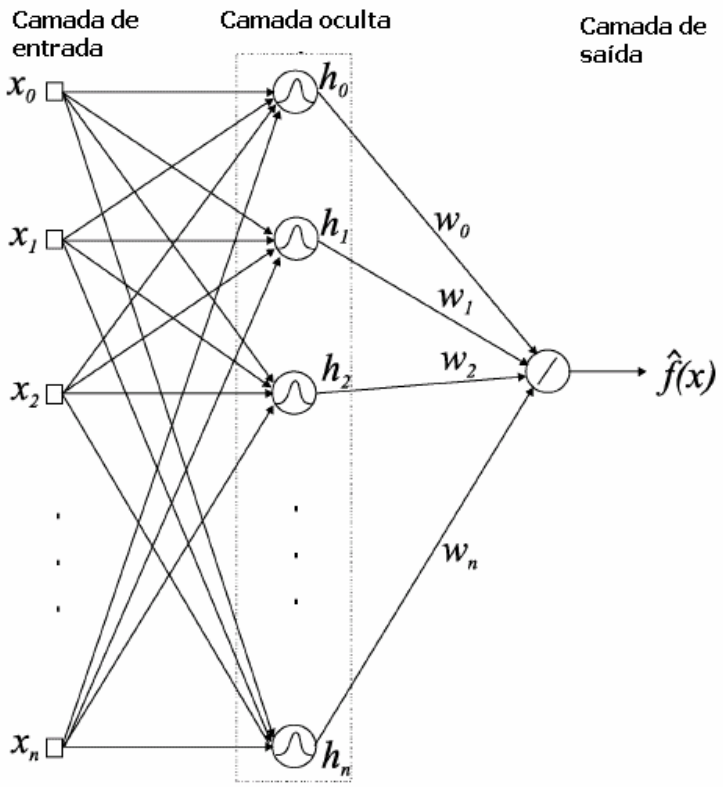

Figura 2.2.1 Topologia das RNA RBF.

O treino da rede ocorre em duas etapas: uma seleção dos raios e centros das funções $h_i(x)$ e o ajuste dos pesos w_i. Detalhes experimentais poderão ser consultados em sua referência [Carvalho et al., 2005]. Segundo Carvalho et al. os resultados obtidos sugerem superioridade da abordagem genética proposta, com RNA RBF resultantes com menor número médio de nós nas camadas ocultas e menor taxa média de erro, comparativamente aos resultados obtidos com outras RNA, SVM e RNA RBF treinadas por outros métodos.

.

2.3. Árvores de Decisão

Em contraposição às RNA, cujo conhecimento fica embutido na estrutura da rede, as Árvores de Decisão (AD) têm a vantagem de expressar o conhecimento adquirido por meio de conjuntos de regras encadeadas do tipo “SE-ENTÃO-SENÃO” ou grafo. Na Figura 2.3.1 é apresentada uma AD nas formas de grafo e de lista de regras. Num grafo os nós correspondem aos atributos, as folhas às classes e os ramos aos caminhos de decisão; busca-se nas sub-árvores geradas a partir de um determinado nó a maior uniformidade interna em cada sub-árvore e a maior diferenciação entre elas. O pesquisador Tom M. Mitchell afirma que o aprendizado por AD é um dos mais utilizados em inferência indutiva, encontrando aplicação em diversas áreas inclusive em risco de crédito, [Mitchell 1997]. As AD são robustas em relação a dados incompletos e ruídos, idealmente os dados devem ser representados por instâncias de atributos e seus valores, os valores objetivos da função de aprendizado devem ser valores discretos.

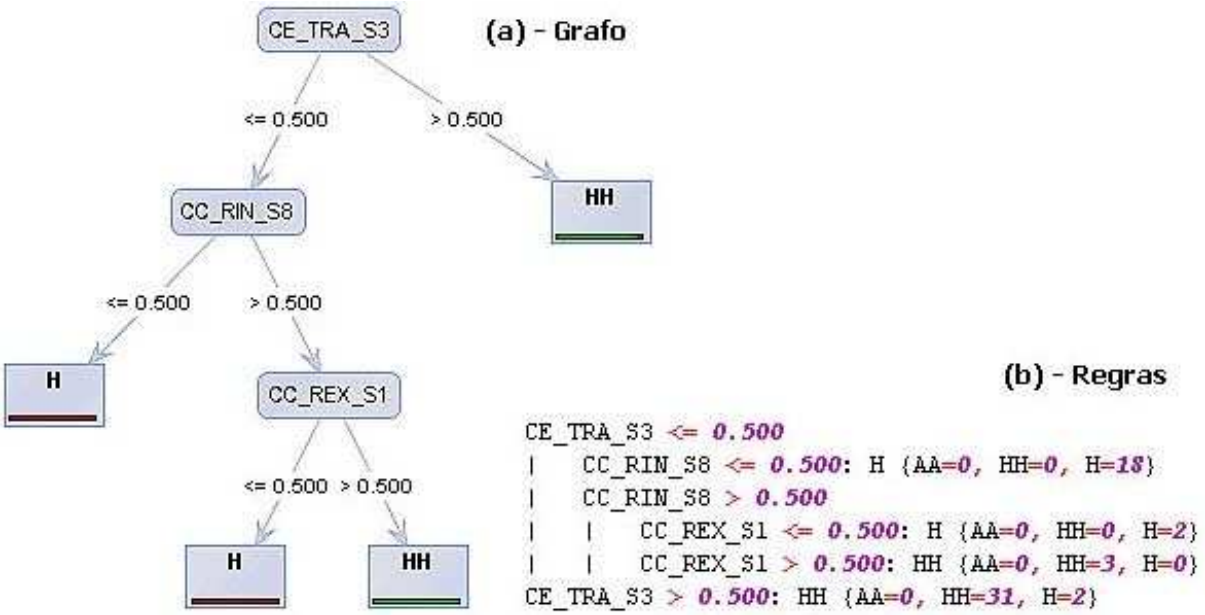

Figura 2.3.1 Árvore de Decisão

Para a formação da árvore, o algoritmo de aprendizado busca o maior ganho de informação possível, e seu cálculo envolve os conceitos da teoria da informação e entropia.

Um dos algoritmos mais conhecidos é o ID3 apresentado abaixo.

Algoritmo ID3 [Mitchell 1997]

```
inicio
para dados de treinamento disponíveis
calcular ganho de informação dos atributos;
selecionar  atributo dos dados de treinamento com maior ganho de informação;
atribuir nó
        se todos os exemplos são da mesma classe
                rotular nó com a classe desses exemplos;
                encerra execução do algoritmo;
        senão
                para todo valor de atributo do nó
                        criar novo nó conectado;
                        aplicar ID3 aos exemplos cobertos por cada nó criado;
                fim-para;
        fim-se
fim-para
retornar árvore;
fim do algoritmo
```

Revisão de Pesquisas Realizadas com Extração de Conhecimento de RNA MLP.

"Knowledge Acquisition Using Symbolic and Connectionist Algorithms for Credit Evaluation", um artigo publicado em 1998 por pesquisadores do ICMC-USP. Esse trabalho investigou o desempenho de RNA do tipo MLP e algoritmos de aprendizado simbólico na avaliação de crédito [Rezende et al., 1998].

Nesse trabalho Rezende et al. investigaram o desempenho de RNA do tipo MLP na avaliação de crédito e comparou-se o conhecimento obtido dessas RNA com o conhecimento gerado pelos algoritmos de aprendizado simbólico C4.5 e CN2. A extração do conhecimento das RNA realizou-se por uma ferramenta externa programada em linguagem PROLOG, desenvolvida por outros pesquisadores do ICMC-USP [Rezende et al., 1998].

Os pesquisadores concluíram que o projeto da RNA, que apresente resultados razoáveis: não consiste de tarefa trivial, é dependente do conjunto de dados, necessita de um bom entendimento do domínio ou um especialista, que auxilie na compreensão dos dados e seleção de atributos e exemplos relevantes e, o uso de ferramenta de extração de conhecimento pode suprir a limitação da RNA em explicar seus resultados.

2.4. Máquina de Suporte de Vetores SVM

As máquinas de suporte de vetores SVM utilizam-se do conceito de hiperplanos linearmente separáveis, citado anteriormente com a técnica MLP, e traz fundamentalmente dois aprimoramentos: a maximização da separação pela busca do hiperplano ótimo para a separação e a utilização de funções de mapeamento para a geração de separação linear, se necessária [Haykin 2001], apresentados respectivamente nas Figuras 2.4.1 e 2.4.2.

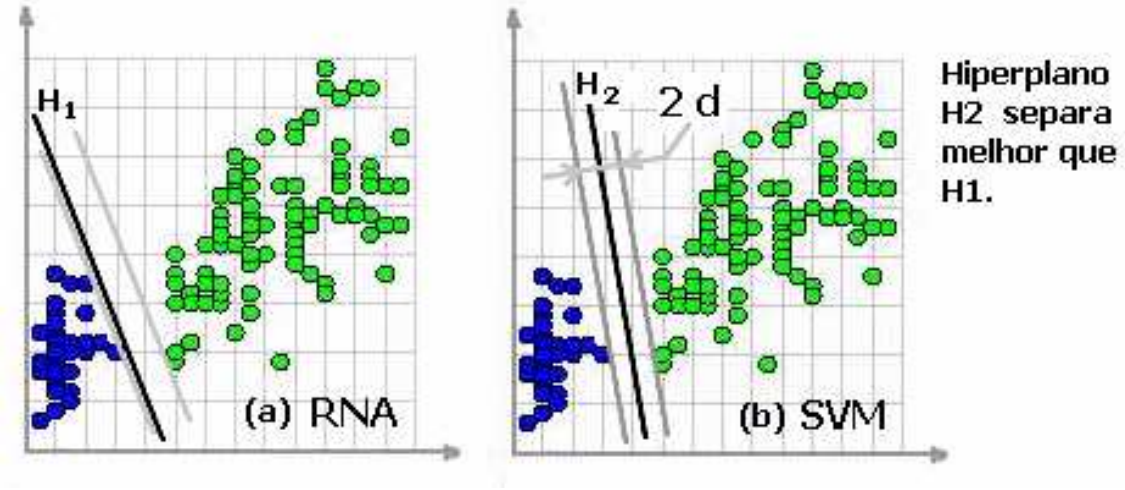

Figura 2.4.1 Hiperplano de máxima separação por SVM

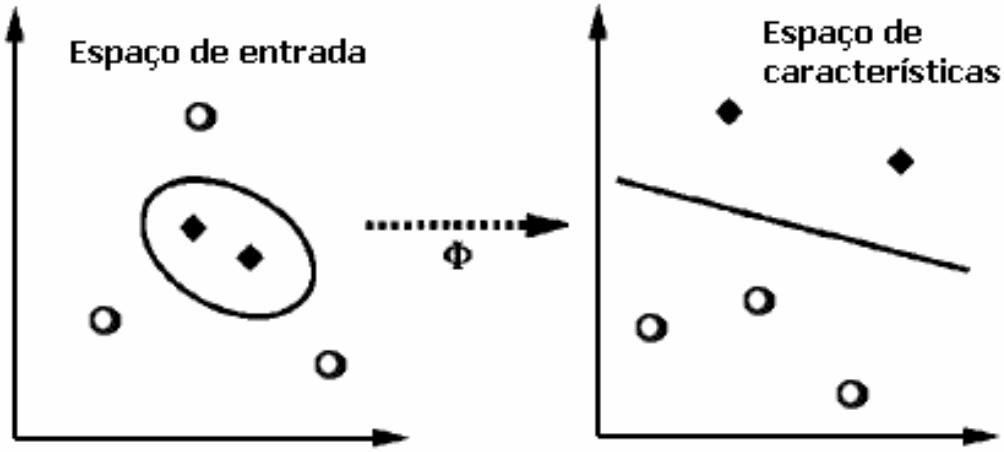

Figura 2.4.2 Linearização do espaço de entrada

Devido o fato do conjunto de funções $\Phi = \{ \Phi_1 .. \Phi_n\}$ de mapeamento poder ser maior que o espaço de entradas e as funções complexas, realiza-se a aproximação por uma função K denominada *Kernel*, sendo os principais tipos: linear, sigmoidal, RBF e polinomial [Haykin 2001], descritos pelas equações 2.4.1 a 2.4.4, respectivamente:

$$K(x_i, x_j) = x_i x_j \tag{2.4.1}$$

$$K(x_i, x_j) = \tanh(\gamma x_i x_j + r) \tag{2.4.2}$$

$$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2), \gamma > 0 \tag{2.4.3}$$

$$K(x_i, x_j) = (\gamma x_i x_j + r)^d, \gamma > 0 \tag{2.4.4}$$

Os valores γ, r e d são parâmetros de controle da função núcleo. Para a função linear (γ , r, d)= (1,0,1). Para as funções sigmoidal e RBF o valor do expoente d=1 e para a função polinomial, o valor de d representa o grau da função.

Revisão de Pesquisas Realizadas com SVM

"Bio-Inspired Credit Risk Analysis: Computational Intelligence with Support Vector Machine", uma monografia, de quatro partes, publicada em 2008, faz um levantamento de trabalhos publicados relacionados ao uso das técnicas de inteligência computacional em geral e investiga a aplicação de SVM em sistemas híbridos com outras técnicas [Yu et al., 2008].

Na primeira parte da monografia Yu et al. [Yu et al., 2008] descreve o problema de risco de crédito e apresenta uma breve explanação da metodologia utilizada no levantamento bibliográfico. Os pesquisadores realizaram buscas de referências bibliográficas em bases científicas: Science Direct, JSTOR, Science Citation Index, IEEE Xplore, Wiley InterScience, com as expressões "*credit scoring OR credit risk evaluation*", com retorno de aproximadamente 600 artigos relacionados a risco de crédito, em um período compreendido entre 1970 e 2007, classificou-se 233 por ano de publicação conforme Figura 2.4.3, e selecionou-se 35 contendo as expressões "SVM" ou "support vector machine".

Para o comparativo qualitativo das diversas técnicas de inteligência computacional utilizadas, Yu et al. estabeleceram quatro critérios: precisão, interpretabilidade, simplicidade e flexibilidade; 32 artigos foram selecionados para o estudo das técnicas utilizadas num período de 1996 a 2007, a Tabela 2.4.1 apresenta os resultados comparativos.

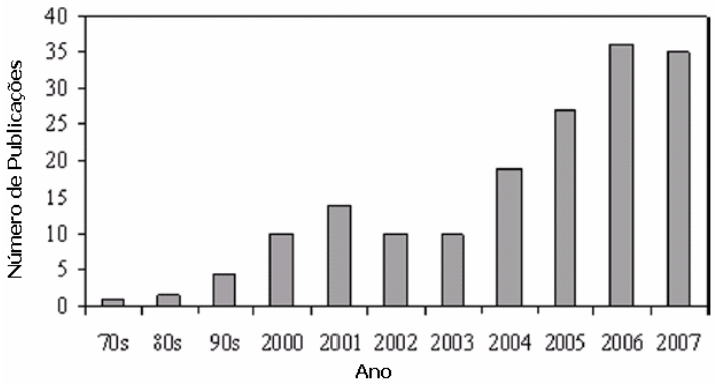

Figura 2.4.3 Distribuição de artigos por ano. Fonte: Yu 2008.

Tabela 2.4.1 Comparativo entre técnicas de inteligência computacional.

Método	Precisão	Interpretabilidade	Simplicidade	Flexibilidade
LDA,LOG, PR	★★	★★★	★★★	★
DT	★★	★★★	★★	★
KNN	★	★★★	★★★	★
LP	★	★★★	★★	★★★
NN	★★★	★	★	★
EA	★★	★	★	★
RS	★★	★	★★	★
SVM	★★★	★★	★	★★★
Hybrid/ensemble	★★★	★	★	★★

LDA - Analise Discriminante Linear
LOG - Regressão Logística
PR - Regressão Probit
DT - Árvores de Decisão
KNN - K-vizinhos mais próximos
LP - Programação Linear
NN - Redes Neurais
EA - Algoritmos Evolucionários
RS - *Rough Sets*
SVM - Máquinas de Vetores de Suporte
Hybrid/Ensemble – Sistemas Híbridos

Fonte: Yu2008, com adaptação de legenda de siglas.

Segundo Yu et al. cada uma das diferentes técnicas de inteligência computacional se destaca em um ou mais dos aspectos qualitativos, sendo o mais relevante para o problema de risco de crédito a precisão. Decorrente dessa constatação, o pesquisador investigou as precisões obtidas nos 32 trabalhos. As maiores precisões foram obtidas com SVM e com sistemas Híbridos. Um resumo do número de experimentos com cada técnica utilizada é apresentado na Tabela 2.4.2, estratificada de [Yu 2008 pp. 20-22].

Tabela 2.4.2 Técnicas utilizadas nos artigos analisados por Yu.

Técnica	**LDA**	**LOG**	**PR**	**DT**	**K-NN**	**LP**	**NN**	**ET**	**RS**	**SVM**	**HY**
Número Experimentos	7	15	1	14	4	1	23	6	2	10	13

Na parte II da monografia, Yu et al. descrevem experimentos com SVM unitários com busca de parâmetros ótimos, através dos algoritmos: ponto vizinho mais próximo (NPA) com projeto de experimento (DOE), pesquisa direta (DS), genético (GA) e pesquisa em grade (GS); Yu et al., relatam que os resultados obtidos com essas técnicas foram melhores que aqueles com SVM unitários simples. Nas partes III e IV, são apresentados respectivamente modelos híbridos e montados com SVM. Os principais objetivos investigativos de Yu et al. foram: constatar a importância dos ajustes nos parâmetros do SVM, averiguar a superioridade dos sistemas híbridos, pesquisar as vantagens da abordagem de montagem de SVM unitários combinando modelos ou aplicando estratégias de aprendizado e programação evolutiva.

Yu et al. ressaltam a escolha de Máquinas de Vetores de Suporte (SVM) em seu trabalho, sobretudo, devido ao alto poder de classificação de padrões e suas características de funcionamento. SVM é uma classe de métodos auto-adaptáveis, não lineares, dirigidos pelos dados sem que haja a necessidade de suposições específicas (por exemplo, distribuição estatística normal), no qual se pode construir uma função de discriminante linear para substituir a função não linear de mapeamento de entrada, com a vantagem de uma dimensão mais baixa que a original. Em conseqüência dessa linearidade da função discriminante, a complexidade computacional não fica dependente da dimensão da amostra. Outra característica importante é a boa capacidade de generalização obtida pelos métodos, através da maximização da separação de hiperplanos, evitando mínimos locais.

2.5. Medidas de Avaliação

Alguns conceitos e medidas comuns para diversas técnicas de aprendizado:

- *Aprendizado supervisionado e não supervisionado*: O aprendizado supervisionado é realizado através de treinamento com exemplos rotulados, onde são informadas as classes corretas para cada exemplo do treinamento. No aprendizado não supervisionado, ocorre a exploração das características dos dados, sem que haja rótulos indicando classes.
- *Medidas em relação ao desempenho ou taxa de acerto (ou erro)*: Para que possa efetivamente ocorrer a aprendizagem é necessário que haja respostas melhores conforme a experiência é adquirida. Medições são executadas para que se controle o desempenho do aprendizado. Três medidas de uso comum em diversas técnicas são a precisão (*precision*), a acurácia (*accuracy*) e a cobertura (*recall*), calculadas pelas proporções de acertos e erros de predição, constantes em uma Tabela denominada matriz de confusão. Um exemplo de matriz de confusão é apresentado na Tabela 2.5.1.
- Complexidade: a complexidade da técnica e ajuste paramétricos deve ser compatível com a complexidade dos dados a serem processados.
- Partição dos dados em subconjuntos de treino e teste: Os dados a serem processados são divididos em subconjuntos de exemplos de treino e teste, devem ser representativos do domínio de dados, homogêneos em relação ao todo e com suficiente diversidade interna.
- *Ajustamento e generalização:* durante o projeto e medições de desempenho, deve-se buscar o equilíbrio entre a capacidade de generalização e ajustamento aos exemplos de treino. O aumento do ajustamento implica na diminuição da generalização e vice-versa. Os extremos configuram situações que impedem o funcionamento correto da técnica utilizada com considerável queda da taxa de acerto para dados novos. Um extremo é o super-ajustamento aos dados de treinamento (*overfitting*), onde o desempenho é satisfatório apenas com o próprio conjunto de treino; o outro extremo é o sub-ajustamento (*underfitting*), onde a técnica não consegue, a partir do treino, aprender as características dos dados a serem processados. Complexidade da técnica elevada em relação aos dados ou partição de treino muito grande em relação aos dados a serem processados induz ao super-ajustamento e, complexidade da técnica baixa em relação a dos dados ou partição de treino pequena induz ao sub-ajustamento.

Tabela 2.5.1 Matriz de confusão

accuracy: 86.33% +/- 0.67% (mikro: 86.33%)					
	true AA	true A	true B	true C	class precision
pred. AA	902	15	27	6	94.95%
pred. A	33	3115	207	280	85.69%
pred. B	0	285	1500	4	83.85%
pred. C	0	17	4	27	56.25%
class recall	96.47%	90.76%	86.31%	8.52%	

Seja a matriz de confusão A, quadrada de ordem k igual ao número de classes do problema, com linhas i e colunas j, a precisão de uma classe i será dada pela equação 2.5.1, a cobertura pela equação 2.5.2 e a acurácia pela equação 2.5.3:

$$P_i := \frac{a_{i=j}}{\sum_{i=0}^{k} a_{i,j}} \tag{2.5.1}$$

$$R_i = \frac{a_{i=j}}{\sum_{j=0}^{k} a_{i,j}} \tag{2.5.2}$$

$$\mathbf{Acc} = \frac{\sum_{j=0}^{k} a_{i=j}}{\sum_{i=0}^{k} \sum_{j=0}^{k} a_{i,j}} \tag{2.5.3}$$

A medida de precisão representa a proporção de exemplos preditos corretamente e o número de exemplos preditos para a classe (somatória na linha), a cobertura é a razão entre o número de exemplos preditos corretamente e o total de exemplos pertencentes à classe (somatória na coluna). A acurácia é a proporção do número de exemplos recuperados em todas as classes (somatória na diagonal principal) e o número total de exemplos.

A avaliação da importância das medidas de precisão e cobertura depende das características do problema abordado, pois classificações com alta precisão e baixa recuperação ou o inverso podem representar resultados inadequados ao problema abordado.

Para a avaliação conjunta da precisão e cobertura, pesquisadores desenvolveram novas medidas que considerem os dois aspectos, uma dessas medidas é a *F-measure*, apresentada na equação 2.5.4. Quando α=1 F_α=P e quando α=0 F_α =R Normalmente utiliza-se α=0,5 [Manning et al.2008].

$$F_\alpha = \frac{PR}{(1-\alpha)P + \alpha R} \tag{2.5.4}$$

- *Curva ROC:* utilizadas desde a década de 40 em sistemas de comunicação e em psicologia, as curvas ROC, representam uma medida de quantificação da precisão obtida em uma classificação binária, tendo como característica importante não serem sensíveis à distribuição de probabilidade dos exemplos de treino ou teste. O traçado da curva é feito a partir das relações de acertos e erro, relacionadas com as medidas de desempenho e matriz de confusão. [Eberhart 2007].

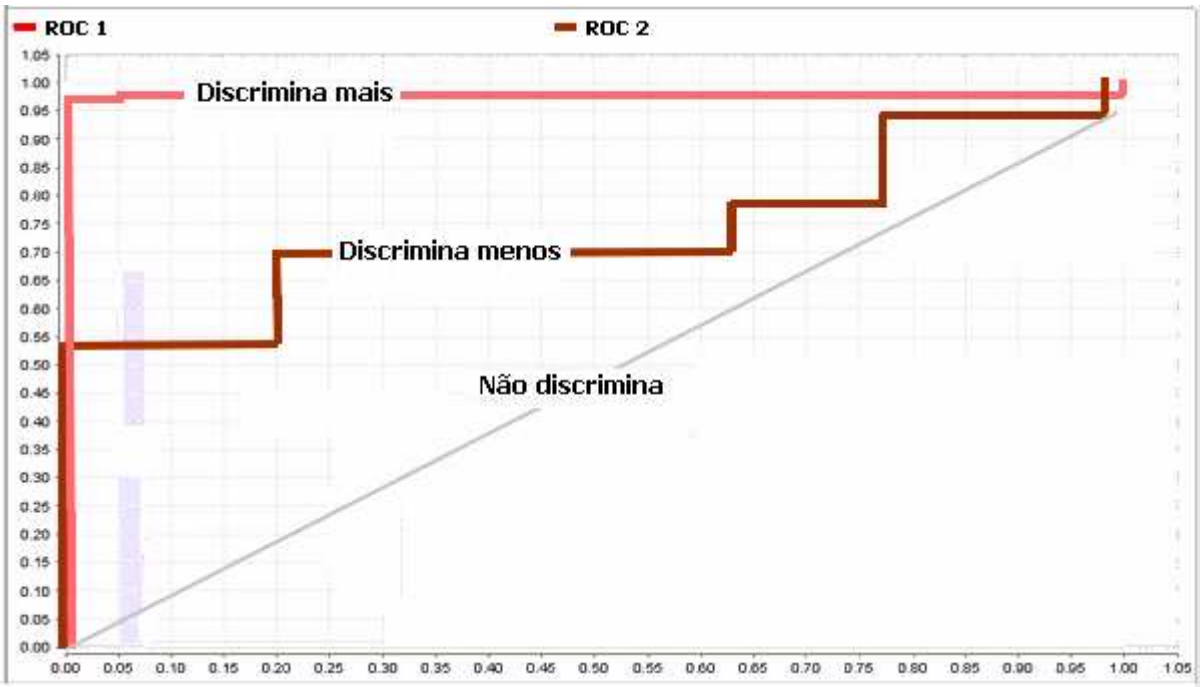

Figura 2.5.1 Matriz de confusão utilizada para definição da curva ROC

A Curva ROC1 discrimina mais que a curva ROC2, sobre a diagonal não há discriminação, pois os exemplos classificados teriam 50% dos exemplos em cada classe.

3. Experimentos Realizados

3.1. Carga dos dados

Os dados, inicialmente em um formato de banco de dados proprietário, foram exportados para um formato de texto plano CSV. Instalou-se o banco de dados PostgreSQL8 em microcomputador com processador de núcleo duplo, compatível com x86, de 4,2 GHz, 4 Gbytes de memória RAM DDR2, e armazenamento serial ATA 500 Gbytes.

Da importação dos dados, do arquivo CSV para o banco de dados, resultou em 26 Tabelas, das quais, selecionou-se para os experimentos, a Tabela mais representativa em número de clientes. Essa Tabela contém dados de empresas comerciais tomadoras de crédito, apresentando um universo de 58995 registros (linhas) com 120 atributos.(colunas) em cada linha.

3.2. Projeto dos Experimentos: Diretrizes

Alguns aspectos relacionados aos dados disponíveis e ao conhecimento do domínio, serviram de linhas gerais e foram considerados no projeto dos experimentos:

- *Tipo de aprendizado*: o tipo de aprendizado escolhido para os experimentos foi o ***supervisionado,*** pois as classes de risco são bem definidas nos dados de treino disponíveis.
- *Quantidades de tuplas*: o ***universo*** dos dados disponíveis relativo às empresas comerciais tomadoras de crédito totalizou 58995 registros (ou tuplas), sendo relativamente grande, para o poder e tempo de processamentos disponíveis, decidiu-se pela colhida de uma ***amostra*** de 6853 registros, para testes com os algoritmos. Observou-se a manutenção das proporções de classes da amostra em relação ao universo; os histogramas de classes de risco são apresentados na Figura 3.2.1. Universo e amostra são Tabelas distintas e sua escolha realizada pela clausula FROM do arquivo de consulta SQL. Decidiu-se que a amostra estaria contida no universo, devido ao uso de validação cruzada.
- *Quantidade de atributos*: Para a otimização do processamento, efetuou-se uma seleção de atributos de entrada comparando-se o resultado obtido pelo algoritmo de seleção *weka. attributeSelection.CfsSubsetEval* e pesquisa *weka.ttribute Selection.BestFirst* -D

1 -N 5, da ferramenta Weka, com a estatística dos atributos geradas no RapidMiner. Constatou-se uma correspondência para a maioria dos atributos. A partir dessa seleção de atributos, e da comparação de resultados preliminares em ensaios de teste, estabeleceram-se dois conjuntos iniciais de atributos: ***todosat.sql***, contendo todos os 85 subfatores e a classe e, ***1asat.sql***, com 50 subfatores e a classe. A seleção dos atributos é feita pela clausula SELECT da consulta SQL.

- *Classes processadas nos experimentos*: através de variações na clausula WHERE da consulta SQL efetuou-se a escolha de conjuntos de classes participantes dos ensaios. Os grupos básicos de classes correspondem à: todas as classes[3] {AA,B,C,D, E,F,G,H,HH}, extremos {AA, A, H,HH}, extremo de notas altas {AA,A,B,C} e extremo de notas baixas {D, E, F, G, H, HH}.

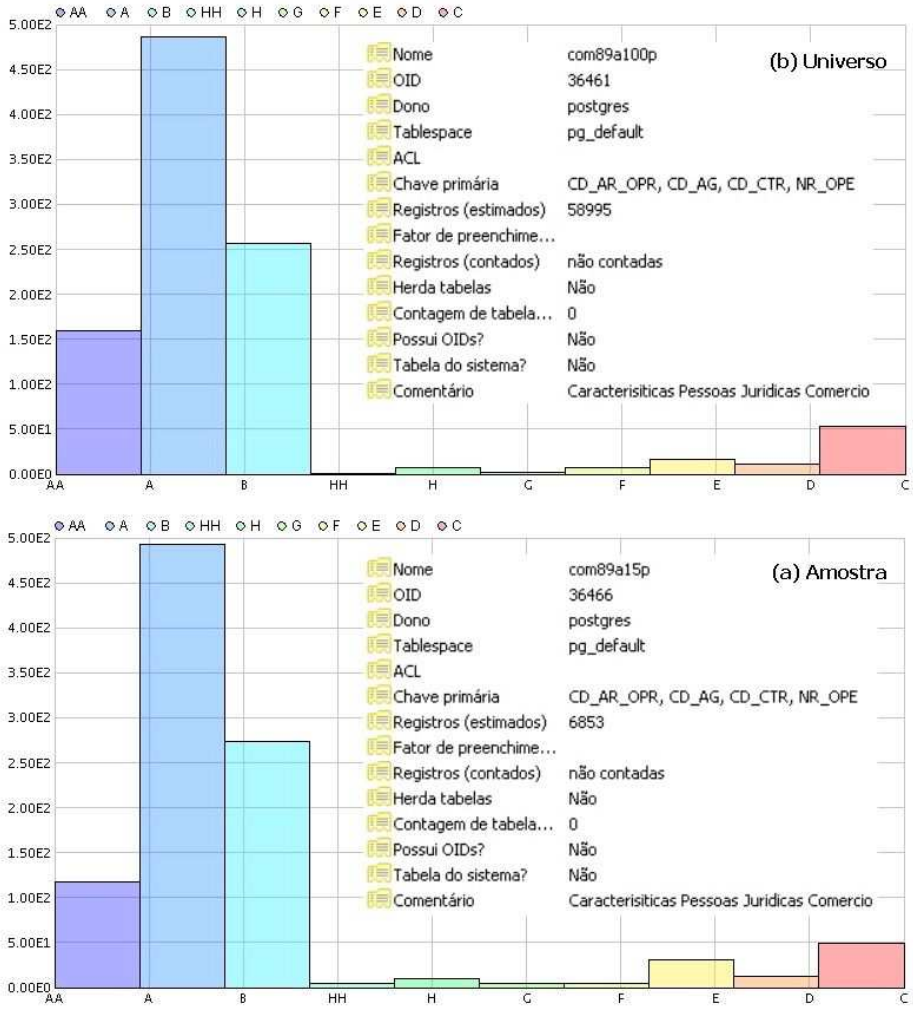

Figura 3.2.1 Histograma das classes de risco do universo e amostra

[3] De acordo com a resolução BACEN 2692 de 21/12/99 no Brasil classificam-se riscos de crédito em 9 classes, entretanto no contexto de Basiléia II admitem-se mais classes. Os dados recebidos para os experimentos apresentam-se em 10 classes, sendo que na classe de maior risco HH, não ocorre operações de crédito.

A seleção dos dados presentes nos experimentos, é realizada através de consultas em SQL. A Figura 3.2.2 mostra um exemplo de consulta SQL, no SELECT são selecionados os atributos e a classe e no WHERE quais os valores da classe serão analisados.

todas.sql

```
1    SELECT "NT_RSC_AJT_06",
2        "CC_RIN_S1",
3        "CC_RIN_S2",
4        "CC_RIN_S3",
5        "CC_RIN_S4",
6        "CC_RIN_S5",
7        "CC_RIN_S6",
8        "CC_RIN_S7",
9        "CC_RIN_S8",
10       "CC_RIN_S9",
11       "CC_RIN_S10",
12    FROM "com89a15p"
13    WHERE ( "NT_RSC_AJT_06"='AA'
14    OR "NT_RSC_AJT_06"='A'
15    OR "NT_RSC_AJT_06"='B'
16    OR "NT_RSC_AJT_06"='C'
17    OR "NT_RSC_AJT_06"='D'
18    OR "NT_RSC_AJT_06"='E'
19    OR "NT_RSC_AJT_06"='F'
20    OR "NT_RSC_AJT_06"='G'
21    OR "NT_RSC_AJT_06"='H'
22    OR "NT_RSC_AJT_06"='HH');
```

M:\Dados\Analise2\MLP\todas\todas.sql

Figura 3.2.2 Exemplo de consulta SQL.

3.3. Pré-processamento dos Dados

Através da inspeção dos dados, retiraram-se atributos irrelevantes para a análise: códigos internos, datas e números de documentos, mantendo-se apenas os atributos: chaves primárias, a classe de risco e os conjuntos de subfatores representativos dos fatores de risco. Adicionou-se uma nova classe binária, para resumir as classes originais: AA, A, B e C correspondendo ao valor S e, as classes HH, H, G, F, E, D correspondendo ao valor N, o significado dessa classe é emprestar para S e não emprestar para N. A Tabela resultante apresenta 90 atributos relativos aos subfatores de risco de crédito. Devido à formatação dos dados, utilizada pela instituição, a tarefa de pré-processamento foi reduzida e os dados apresentaram-se adequados aos processamentos.

3.4. Estrutura Geral dos Experimentos

Através da ferramenta RapidMiner, montou-se a estrutura geral utilizada pela maioria dos experimentos com as diversas técnicas. O conceito de montagem de experimentos assemelha-se ao *knowledgeFlow* da ferramenta Weka, onde define-se um fluxo de elementos de processamento. Os elementos centrais são: acesso à base de dados, validação cruzada, algoritmo de aprendizado, e extratores de resultados. A Figura 3.4.1 mostra o diagrama da estrutura geral dos experimentos e a Figura 3.4.2 o diagrama na forma XML.

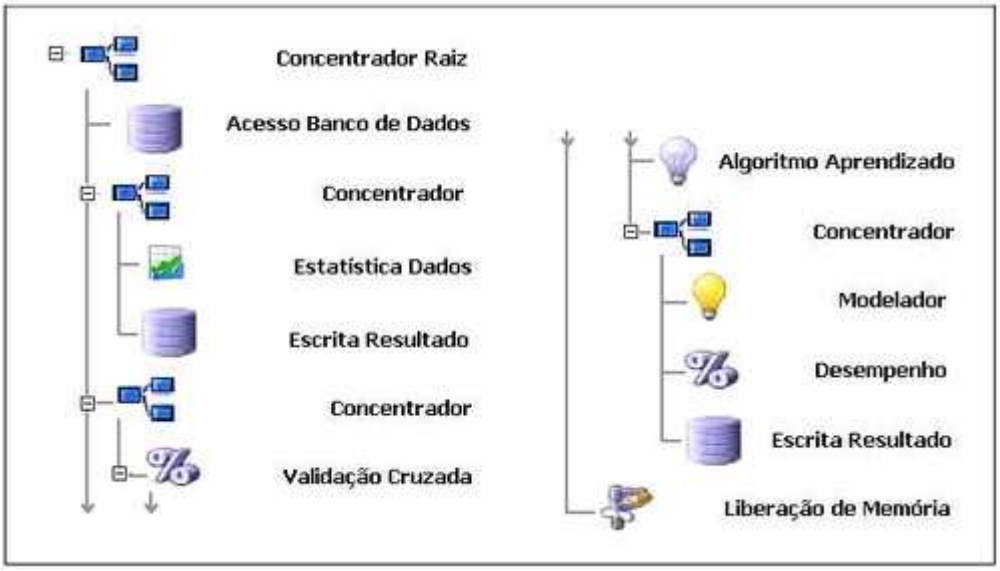

Figura 3.4.1 Estrutura geral dos experimentos com RapidMiner.

mlp.todas.xml

```
1   <?xml version="1.0" encoding="windows-1252"?>
2   <process version="4.2">
3
4     <operator name="Concentrador Raiz" class="Process" expanded="yes">
5         <operator name="Base de Dados Risco" class="DatabaseExampleSource">
6             <parameter key="database_system"  value="PostgreSQL"/>
7             <parameter key="database_url" value="jdbc:postgresql://localhost:5432/Risco"/>
8             <parameter key="label_attribute"  value="NT_RSC_AJT_06"/>
9             <parameter key="password" value="plLkadf7wvg="/>
10            <parameter key="query"    value="M:\SQL\todas.sql"/>
11            <parameter key="username" value="postgres"/>
12        </operator>
13        <operator name="Concentrador Estatistica" class="OperatorChain" expanded="yes">
14            <operator name="Calculo Estatistico" class="DataStatistics">
15            </operator>
16            <operator name="Grava Estatistica" class="ResultWriter">
17                <parameter key="result_file"  value="M:\Estatisticas\Estatisticas.res"/>
18            </operator>
19        </operator>
20        <operator name="Concentrador Algoritmo" class="OperatorChain" expanded="yes">
21            <operator name="Validacao Cruzada" class="XValidation" expanded="yes">
22                <parameter key="average_performances_only"    value="false"/>
23                <parameter key="number_of_validations"    value="3"/>
24                <operator name="Algoritmo MLP" class="W-MultilayerPerceptron">
25                </operator>
26                <operator name="Concentrador Desempenho" class="OperatorChain" expanded="yes">
27                    <operator name="Aplicacao de Modelo" class="ModelApplier">
28                        <list key="application_parameters">
29                        </list>
30                    </operator>
31                    <operator name="Medida de Desempenho" class="Performance">
32                        <parameter key="use_example_weights"  value="false"/>
33                    </operator>
34                    <operator name="Grava Desempenho" class="ResultWriter">
35                        <parameter key="result_file"  value="M:\Resultado\Matriz.Confusao.res"/>
36                    </operator>
37                </operator>
38            </operator>
39        </operator>
40        <operator name="Libera Memoria" class="MemoryCleanUp">
41        </operator>
42    </operator>
43
44  </process>
```

M:\Dados\Analise2\MLP\todas\mlp.todas.xml

Figura 3.4.2 Estrutura geral na forma de arquivo XML.

De acordo com as diretrizes de projeto e estrutura geral de experimentos, planejou-se a execução de séries de experimentos com técnicas individuais. A Tabela 3.4.1 apresenta um resumo de experimentos numerados de 1 a 15 e está organizada por cinco aspectos: quantidade de exemplos, quantidade de atributos, conjuntos de classes, técnica e ferramenta utilizada. Nos conjuntos de classes, {AA..HH} representa o conjunto das classes AA,A,B,C,D,E,F,G,H e HH; {AA..C} representa o conjunto de classes AA,A,B e C, analogamente {D..HH} representa o conjunto com as classes D, E, F, G, H e HH.

Tabela 3.4.1 Planejamento de experimentos com amostra

Experi-	*Qtde de tuplas*		*Qtde de atributos*		*Conjunto de Classes*				*Técnica utilizada*			*Ferramenta*	
mento	*universo*	*amostra*	*todos*	*1a.Seleção*	*{AA..HH}*	*{AA,A,H,HH}*	*{AA..C}*	*{D..HH}*	*MLP*	*AD*	*SVM*	*RMiner*	*Weka*
1		o	o		o				o			o	
2		o	o		o				o				o
3		o		o	o				o			o	
4		o		o		o			o			o	
5		o		o			o		o			o	
6		o		o				o	o			o	
7		o	o		o			o	o			o	
8		o		o	o					o		o	
9		o		o		o				o		o	
10		o		o			o			o		o	
11		o		o				o		o		o	
12		o		o	o						o		
13		o		o		o					o		
14		o		o			o				o		
15		o		o				o			o		

As apresentações de resultados dos experimentos utilizarão as matrizes de confusão e suas representações gráficas. Nas representações gráficas serão omitidas as legendas das classes que seguirão o padrão apresentado na Figura 3.4.3, exceto se desordenada.

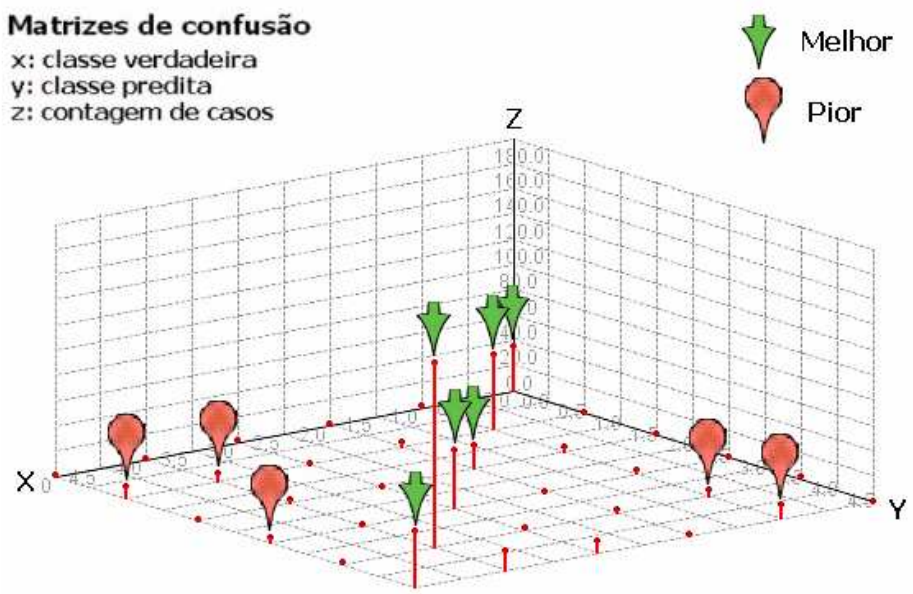

Figura 3.4.3 Representação gráfica de uma matriz de confusão.

3.5. Experimentos com RNA MLP

Para a execução dos experimentos de números 1 a 7, utilizando a técnica RNA MLP, definiu-se o número de neurônios da camada oculta igual a média aritmética do número de entrada p e saídas m. Os parâmetros de ajuste do algoritmo de retro propagação foram ajustados com os valores: momento α=0,2; taxa de aprendizado η=0,3 e número de épocas n=500. Na Figura 3.5.1 é apresentado a topologia da rede, com p=85, m=10 e 47 neurônios da camada oculta.

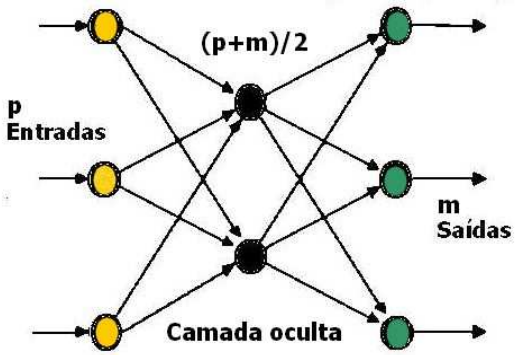

Figura 3.5.1 RNA Perceptron Multicamadas.

Através da comparação dos resultados dos experimentos 1 e 2 constatou-se que diversos subfatores (atributos na Tabela) possuíam valores lógicos idênticos e iguais a zero para todas as tuplas, situação até então desconhecida. Eliminaram-se esses atributos irrelevantes da consulta SQL e realizou-se o experimento 3. A Figura 3.5.2 e Tabela 3.5.1 mostram o viés de classificação para as classes mais numerosas no experimento 3.

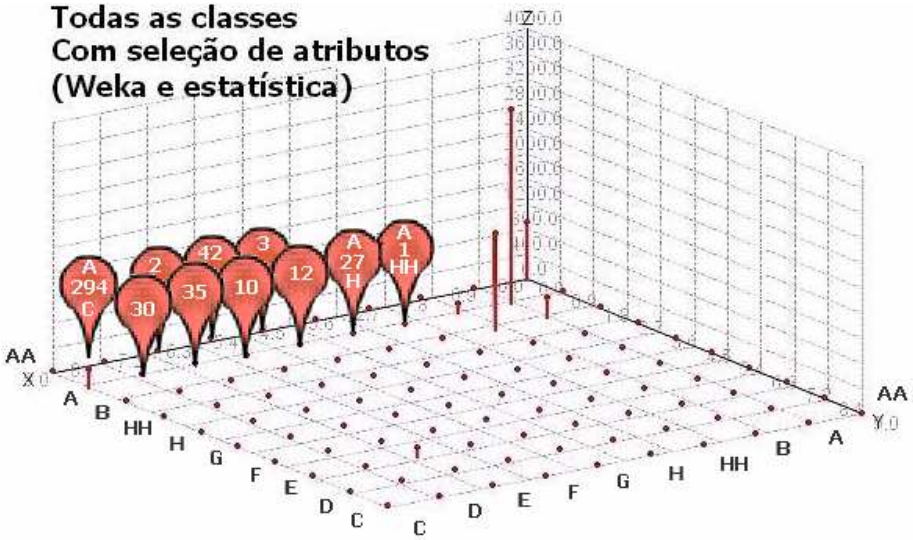

Figura 3.5.2 Matriz de confusão do experimento MLP 3.

Tabela 3.5.1 Matrizes de confusão do experimento 3 com MLP.

accuracy: 83.12% +/- 1.06% (mikro: 83.12%) kappa: 0.736 +/- 0.016 (mikro: 0.736)

	true AA	true A	true B	true HH	true H	true G	true F	true E	true D	true C	class precis
pred. AA	882	13	17	1	0	2	42	3	1	7	91.12%
pred. A	33	3082	174	1	27	12	10	35	30	294	83.34%
pred. B	20	324	1534	8	6	4	1	3	3	5	80.40%
pred. HH	0	0	0	11	0	0	2	0	0	0	84.62%
pred. H	0	8	6	3	42	1	2	1	6	4	57.53%
pred. G	0	0	1	0	0	3	0	0	0	1	60.00%
pred. F	0	0	0	0	0	0	0	0	0	0	0.00%
pred. E	0	2	4	1	0	3	0	122	3	3	88.41%
pred. D	0	0	0	6	5	0	0	4	17	0	53.12%
pred. C	0	3	2	5	2	0	1	1	1	3	16.67%
class recall	94.33%	89.80%	88.26%	30.56%	51.22%	12.00%	0.00%	72.19%	27.87%	0.95%	

A descrição do experimento 16 e os resultados dos experimentos com MLP 1 a 7 estão resumidos nas Tabelas 3.5.2 e 3.5.3. Na Tabela 3.5.4 é apresentada a matriz de confusão resultante do experimento 16.

Tabela 3.5.2 Resultados dos experimentos MLP 1 a 7.

Experi-mento	*Cobertura (%)* AA	A	B	C	D	E	F	G	H	HH		
1	94,3	89,8	88,3	1,0	27,8	72,2	0,0	12,0	51,2	30,5		
2												
3	93,8	89,8	88,4	2,5	31,2	72,2	0,0	8,0	46,3	58,3		
4	93,4	99,5							45,1	75,0		
5	94,3	87,1	88,7	29,0								
6					75,4	85,8	79,3	60,0	75,6	94,4		
7					72,1	86,4	79,3	72,0	71,9	94,4		
Experi-mento	*Precisão (%)* AA	A	B	C	D	E	F	G	H	HH	*Acurácia Global*	*Desvio Padrão*
1	91,1	83,3	80,4	16,7	53,1	88,4	0,0	60,0	57,5	84,6	83,12	1,06
2												
3	91,3	83,7	79,8	38,1	59,4	91,0	0,0	25,0	62,3	77,8	83,26	0,68
4	97,9	96,9							86,1	100,0	96,99	0,50
5	95,7	88,1	81,0	44,9							85,69	0,44
6					64,0	90,6	86,8	68,2	72,9	87,2	80,73	3,85
7					57,1	91,8	86,8	66,7	75,6	91,9	80,50	2,90

Tabela 3.5.3 Características do experimento 16 com MLP.

Critério	*Qtde de tuplas*		*Qtde de atributos*		*Conjunto de Classes*				*Técnica utilizada*			*Ferramenta*		
#	*universo*	*amostra*	*todos*	*1a.Seleção*	*{AA..HH}*	*{AA,A,H,HH}*	*{AA..C}*	*{D..HH}*	*MLP*	*AD*	*SVMAE*	*RMiner*	*Weka*	*Tanagra*
16	o		o		o				o			o		

Tabela 3.5.4 Matriz de confusão do experimento 16 com MLP.

curacy: 73.09% +/- 0.20% (mikro: 73.09%)

	true AA	true C	true D	true E	true F	true G	true H	true HH	true B	true A	class pr
AA	6339	270	104	202	207	97	149	0	424	963	72.40%
C	0	258	7	5	1	0	2	1	25	14	82.43%
D	0	8	20	1	0	0	6	0	11	16	32.26%
E	0	4	1	628	2	1	2	2	66	4	88.45%
F	0	0	0	0	0	0	0	0	0	0	0.00%
G	0	0	0	0	0	0	0	0	0	0	0.00%
H	0	1	5	4	3	1	40	14	5	5	51.28%
HH	0	0	0	1	0	1	3	35	0	0	87.50%
B	869	1175	108	55	8	17	71	95	10798	3580	64.37%
A	991	1115	235	71	37	18	191	19	4581	25003	77.50%
s recall	77.31%	9.11%	4.17%	64.94%	0.00%	0.00%	8.62%	21.08%	67.87%	84.51%	

3.6. Experimentos com Árvores de Decisão

Nos experimentos de números 8 a 11, configurou-se o elemento de aprendizado na ferramenta RapidMiner, para AD, com poda automática e ganho de informação, altura da árvore com valor 10 e no mínimo duas folhas. Semelhante aos resultados com RNA MLP, a matriz de confusão obtida do experimento 8 é apresentada na Tabela 3.6-1 e sua representação gráfica muito semelhante ao da Figura 3.5-2, apresentando significativo viés para as classes mais numerosas. A Tabela 3.6-2 apresenta as precisões e recuperação dos experimentos.

Tabela 3.6.1 Matrizes de confusão dos experimentos 8 com AD.

accuracy: 81.96% +/- 0.72% (mikro: 81.96%)

	true AA	true A	true B	true HH	true H	true G	true F	true E	true D	true C	class p
pred. AA	901	17	30	0	2	0	42	6	1	9	89.38%
pred. A	34	3072	270	2	26	13	13	35	35	284	81.18%
pred. B	0	318	1422	2	9	5	2	2	3	6	80.38%
pred. HH	0	0	0	32	1	0	0	0	0	0	96.97%
pred. H	0	10	7	0	35	0	0	0	3	2	61.40%
pred. G	0	0	2	0	0	6	0	0	0	4	50.00%
pred. F	0	0	0	0	0	0	0	0	0	0	0.00%
pred. E	0	1	3	0	1	0	1	122	3	1	92.42%
pred. D	0	0	1	0	0	0	0	3	16	0	80.00%
pred. C	0	14	3	0	8	1	0	1	0	11	28.95%
class recall	96.36%	89.51%	81.82%	88.89%	42.68%	24.00%	0.00%	72.19%	26.23%	3.47%	

Os resultados para os experimentos com AD estão resumidos na Tabela 4.6-2

Tabela 3.6.2 Resultados dos experimentos AD

Experi-mento	*Cobertura (%)* *AA*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*HH*		
8	96,4	89,5	81,8	3,5	26,2	72,2	0,0	24,0	42,7	88,9		
9	96,8	99,2							50,0	94,4		
10	96,6	91,0	79,6	4,1								
11					57,4	81,7	84,5	60,0	74,4	94,4		
Experi-mento	*Precisão (%)* *AA*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*HH*	*Acurácia Global*	*Desvio Padrão*
8	89,4	81,2	80,4	29,0	80,0	92,4	0,0	50,0	61,4	97,0	81,96	0,72
9	98,1	98,0							85,4	91,9	97,79	0,38
10	94,0	83,1	82,8	43,3							84,43	0,18
11					60,3	85,2	68,1	75,0	75,3	89,5	77,02	2,56

Nos experimentos planejados inicialmente, os ensaios de extremos incluíam 4 classes {AA, A, H, HH}, ensaiou-se com AD extremos de duas classes {AA,HH}, cujos resultados são apresentados nas Figuras 3.6.1 e 3.6.2.

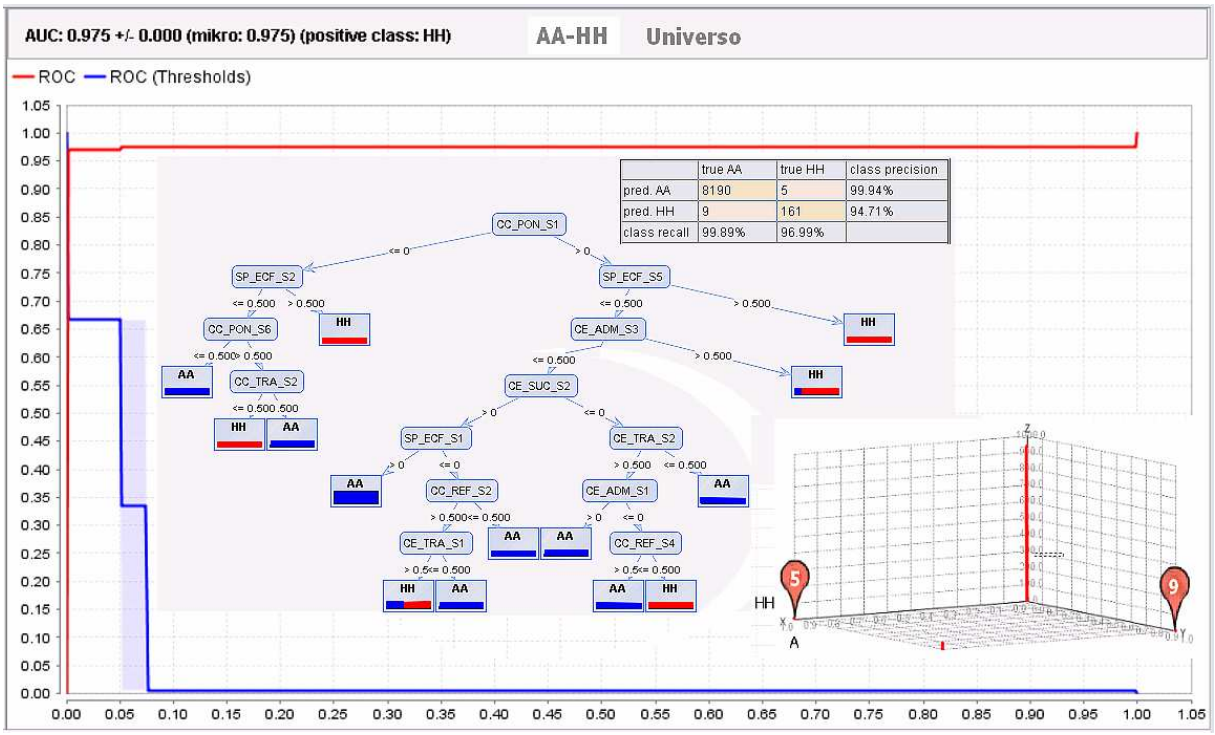

Figura 3.6.1 Matriz de confusão, árvore e curva ROC universo.

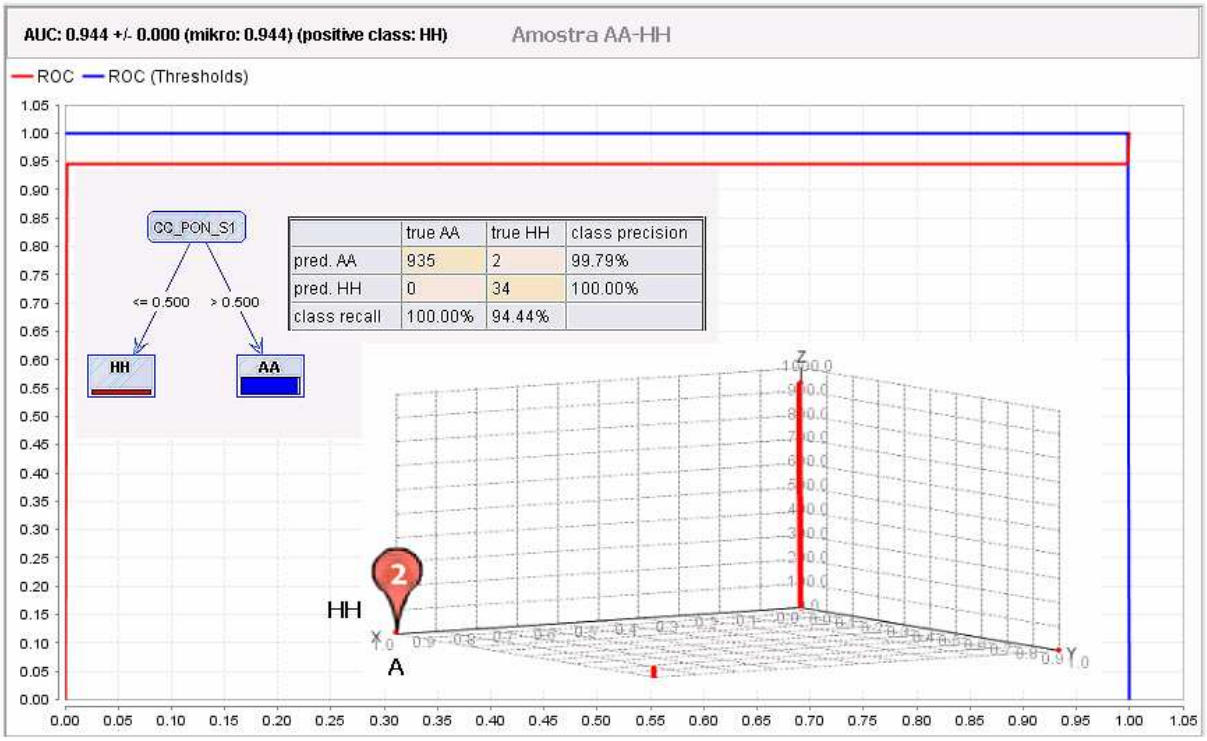

Figura 3.6.2 Matriz de confusão, árvore e curva ROC amostra.

3.7. Experimentos com SVM

Nos experimentos de números 12 a 15, utilizou-se a técnica de SVM, configurou-se o elemento de aprendizado para SVM do tipo C-SVC, com núcleo do tipo RBF e função polinomial de grau 3. A matriz de confusão é apresentada na Tabela 3.7.1.

Tabela 3.7.1 Matriz de confusão do experimento 12

accuracy: 83.48% +/- 0.60% (mikro: 83.48%)

	true AA	true A	true B	true HH	true H	true G	true F	true E	true D	true C	class pre
pred. AA	899	15	26	0	0	0	42	3	1	6	90.62%
pred. A	35	3085	200	4	38	9	11	35	36	290	82.42%
pred. B	0	308	1498	0	2	4	0	1	2	2	82.44%
pred. HH	0	0	0	32	0	0	0	0	0	0	100.00%
pred. H	0	8	7	0	35	0	0	0	3	2	63.64%
pred. G	0	0	0	0	0	12	0	1	0	3	75.00%
pred. F	0	0	0	0	0	0	5	0	0	0	100.00%
pred. E	0	1	2	0	0	0	0	126	3	1	94.74%
pred. D	1	0	0	0	0	0	0	3	16	0	80.00%
pred. C	0	15	5	0	7	0	0	0	0	13	32.50%
class. recall	96.15%	89.89%	86.19%	88.89%	42.68%	48.00%	8.62%	74.56%	26.23%	4.10%	

Os resultados para os experimentos com SVM estão resumidos na Tabela 3.7.2

Tabela 3.7.2 Resultados dos experimentos SVM

Experimento	*Cobertura (%)* AA	A	B	C	D	E	F	G	H	HH		
12	96,2	89,9	86,2	4,1	26,2	74,6	8,6	45,0	42,7	89,9		
13	95,3	98,9							52,4	94,4		
14	96,0	90,6	85,6	1,0								
15					63,9	86,4	84,5	72,0	92,7	94,4		
Experimento	*Precisão (%)* AA	A	B	C	D	E	F	G	H	HH	*Acurácia Global*	*Desvio Padrão*
12	90,6	82,4	82,4	32,5	80,0	94,7	100,0	75,0	63,6	100,0	83,45	0,60
13	98,5	97,6							64,2	97,1	97,24	0,50
14	95,0	84,8	82,8	23,1					64,2	97,1	85,60	0,12
15					72,2	94,2	96,1	90,0	65,0	100,0	83,99	1,75

3.8. Experimentos SVM com Otimização Genética

Adicionou-se à estrutura dos experimentos de números 12 a 15, um elemento de otimização de parâmetros por algoritmo evolucionário, originando-se a série de experimentos apresentada na Tabela 3.8.1. O operador evolutivo foi aplicado ao parâmetro γ do SVM

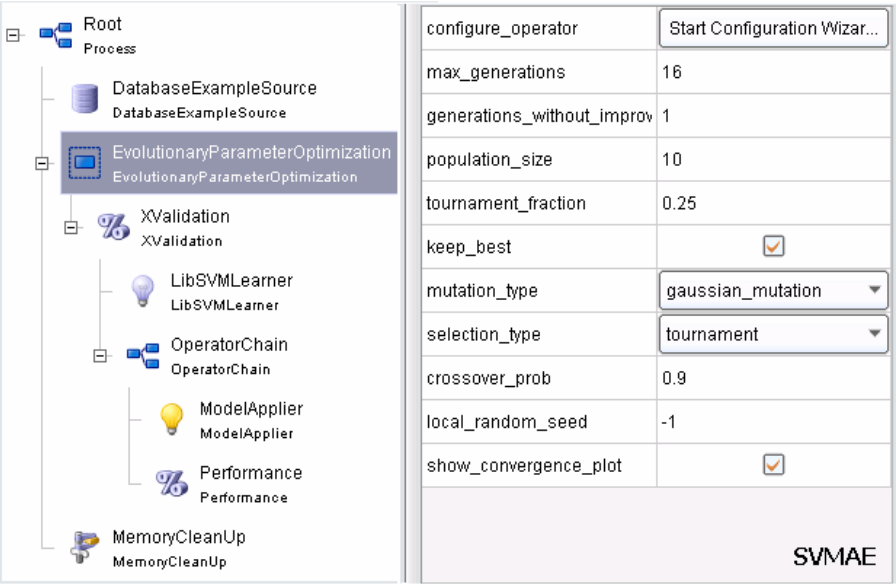

Figura 3.8.1 Modelo para os experimentos 17 a 20.

Tabela 3.8.1 Experimentos com SVMAE

Critério	*Qtde de tuplas*		*Qtde de atributos*		*Conjunto de Classes*				*Técnica utilizada*			*Ferramenta*		
#	*universo*	*amostra*	*todos*	*1a.Seleção*	*{AA..HH}*	*{AA,A,H,HH}*	*{AA..C}*	*{D..HH}*	*MLP*	*AD*	*SVMAE*	*RMiner*	*Weka*	*Tanagra*
17		o		o	o						o			
18		o		o		o					o			
19		o		o			o				o			
20		o		o				o			o			

Os resultados para os experimentos com SVM estão resumidos na Tabela 3.8.2

Tabela 3.8.2 Resultados dos experimentos SVM com otimização evolutiva.

Experi-	*Cobertura (%)*											
mento	*AA*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*HH*		
16	77,3	84,5	67,9	9,1	4,2	64,9	0,0	0,0	8,6	21,1		
17	96,3	90,7	85,9	6,6	34,4	74,6	8,6	48,0	45,1	94,4		
18	96,3	99,3							45,1	94,4		
19	96,5	90,8	86,3	8,5								
20					72,1	85,8	89,7	72,0	92,7	94,4		
Experi-	*Precisão (%)*										*Acurácia*	*Desvio*
mento	*AA*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*HH*	*Global*	*Padrão*
16	72,4	77,5	64,4	82,4	32,3	88.45	0,0	0,0	51,3	87,5	73,09	0,20
17	90,6	82,7	83,4	48,8	84,0	97,7	100,0	75,0	64,9	97,1	84,05	1,34
18	98,4	97,7							82,2	100,0	97,66	0,09
19	95,0	85,7	83,9	56,3							86,33	0,67
20					75,9	96,7	96,3	85,7	66,7	100,0	85,62	1,78

3.9. Experimentos com Amostra Balanceada

Através da observação dos resultados dos experimentos 1 a 20 e da constatação do viés de classificação para as classes mais numerosas, decidiu-se pela realização de uma amostragem dos dados buscando-se o balanceamento do número de exemplos entre as classes.

A amostra balanceada representada na Figura 3.9.1 apresentou 135 exemplos de cada classe, totalizando 1350 exemplos.

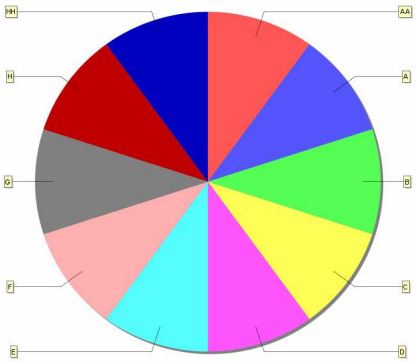

Figura 3.9.1 Amostra balanceada.

As características dos experimentos de números 21 a 36 realizados com a amostra balanceada são apresentadas na Tabela 3.9.1.

Tabela 3.9.1 Experimentos realizados com a amostra balanceada.

Experi-	*Qtde de tuplas*	*Qtde de atributos*		*Conjunto de Classes*				*Técnica utilizada*			*Ferramenta*	
mento	*Amostra Balanceada*	*todos*	*1a.Seleção*	*{AA..HH}*	*{AA,A,H,HH}*	*{AA..C}*	*{D..HH}*	*MLP*	*AD*	*SVM*	*SVMAE*	*RMiner*
21	o		o	o				o				o
22	o		o	o					o			o
23	o		o	o						o		o
24	o		o	o							o	o
25	o		o		o			o				o
26	o		o		o				o			o
27	o		o		o					o		o
28	o		o		o						o	o
29	o		o			o		o				o
30	o		o			o			o			o
31	o		o			o				o		o
32	o		o			o					o	o
33	o		o				o	o				o
34	o		o				o		o			o
35	o		o				o			o		o
36	o		o				o				o	o

Na Tabela 3.9.2 é apresentada a matriz de confusão obtida do experimento 24 com a amostra balanceada, com todas as classes, utilizando a técnica SVMAE. Os resultados obtidos com todos os experimentos com a amostra balanceada são apresentados na Tabela 3.9.3.

Tabela 3.9.2 Matriz de confusão do experimento 24 com SVMAE.

accuracy: 66.30% +/- 2.81% (mikro: 66.30%)

	true AA	true A	true B	true C	true D	true E	true F	true G	true H	true HH	class precision
pred. AA	80	0	0	6	0	0	0	0	0	0	93.02%
pred. A	0	74	12	1	1	0	0	0	1	1	82.22%
pred. B	1	31	110	12	15	13	14	8	36	5	44.90%
pred. C	0	11	1	71	1	0	0	0	3	0	81.61%
pred. D	0	2	3	25	114	10	1	1	12	0	67.86%
pred. E	0	4	6	12	0	108	0	5	3	0	78.26%
pred. F	53	9	2	0	0	0	118	94	12	0	40.97%
pred. G	1	0	0	4	1	3	0	24	0	0	72.73%
pred. H	0	4	1	4	3	1	2	3	68	1	78.16%
pred. HH	0	0	0	0	0	0	0	0	0	128	100.00%
class recall	59.26%	54.81%	81.48%	52.59%	84.44%	80.00%	87.41%	17.78%	50.37%	94.81%	

Tabela 3.9.3 Resultados obtidos nos experimentos 21 a 36.

Experi-mento	*Cobertura (%)* *AA*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*HH*		
20					72,13	85,80	89,66	72,00	92,68	94,44		
21	59,26	55,56	63,70	45,93	84,44	81,48	83,70	16,30	52,59	85,19		
22	59,26	27,41	68,99	41,48	69,63	62,96	89,63	8,15	37,78	88,15		
23	59,26	52,59	79,26	46,67	80,00	74,07	89,63	15,56	46,67	94,81		
24	59,26	54,81	81,48	52,59	84,44	80,00	87,41	17,78	50,37	94,81		
25	100,00	80,00							71,85	90,37		
26	100,00	79,26							67,41	93,33		
27	98,52	67,41							85,93	94,81		
28	98,52	75,56							88,67	94,81		
29	98,52	55,56	63,70	85,19								
30	99,26	40,00	79,26	63,70								
31	98,52	46,67	85,19	78,52								
32	98,52	57,04	83,70	79,26								
33					85,19	79,26	88,15	20,00	65,93	88,89		
34					80,00	78,52	87,41	13,33	51,11	91,85		
35					80,00	77,04	88,15	17,04	79,26	92,59		
36					84,44	80,00	93,33	20,74	79,26	94,81		
Experi-mento	*Precisão (%)* *AA*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*HH*	*Acurácia Global*	*Desvio Padrão*
20					75,86	96,67	96,30	85,71	66,67	100,00	85,62	1,78
21	84,21	64,66	61,43	69,66	62,64	70,51	38,70	59,46	65,74	85,19	62,81	2,82
22	81,63	55,22	39,41	58,95	56,29	89,47	36,67	39,29	53,12	86,23	55,33	1,61
23	93,02	82,56	41,47	67,74	65,85	80,65	40,47	70,00	76,83	100,00	63,85	2,98
24	93,02	82,22	44,90	81,61	67,86	78,26	40,97	72,73	78,16	100,00	66,30	2,81
25	86,54	85,04							77,60	92,42	85,56	1,36
26	81,33	84,25							85,85	89,36	85,00	2,53
27	86,93	91,00							72,95	100,00	86,67	1,20
28	86,93	92,73							78,52	100,00	88,89	0,91
29	87,50	70,75	68,80	73,25							75,74	0,94
30	85,35	75,00	54,04	76,11							70,56	2,83
31	88,08	87,50	61,50	81,54							77,22	3,27
32	88,67	88,51	65,70	81,68							79,63	2,77
33					74,68	85,60	51,52	72,97	70,63	87,59	71,23	4,13
34					72,00	71,62	47,58	58,06	75,00	87,94	67,04	3,79
35					81,82	87,39	52,42	85,19	59,44	100,00	72,35	3,03
36					81,43	95,58	53,85	87,50	65,64	100,00	75,43	0,97

4. Discussão dos Resultados e Trabalhos Futuros

4.1. Discussão dos Resultados Obtidos

A estrutura dos dados fornecidos pela instituição financeira mostrou-se adequada para o processamento através de diferentes técnicas da Inteligência Computacional, devido a três de suas características: os exemplos apresentam atributo com valor de classes bem definidas, os atributos relativos aos subfatores, dentro de um fator de risco, possuem valores mutuamente exclusivos e, os dados apresentam-se com valores normalizados. Essas características dos dados, somados ao poder explanatório da técnica de Árvores de Decisão conduzem a sua escolha, entretanto, o número de regras pode tornar-se muito elevado e outras técnicas mostram-se mais precisas na discriminação das classes. O pré-processamento dos dados foi bastante reduzido devido às características dos dados.

Computacionalmente, o desafio deste trabalho apresentou-se na forma de problema de classificação; sob o ponto de vista financeiro, sua solução tem relevância crescente, sobretudo, no contexto da atual crise financeira internacional e também pela possibilidade de aumento da rentabilidade das instituições financeiras: pela liberação de parte dos recursos do capital mínimo requerido, pelo aumento de eficiência e precisão do sistema de classificação, pelo aumento da agilidade na execução de novas políticas de classificação do risco de crédito e pela eliminação de viés de subjetividade de avaliação do operador de crédito.

É natural imaginar-se que as instituições desejem operar com os clientes de mais baixo risco de crédito, os dados fornecidos para este trabalho apresentam-se alinhados com essa hipótese conforme se verifica no histograma da Figura 3.2.1; as proporções entre as notas melhores AA, A e B e, as piores F, G, H e HH, levam os algoritmos à tendências de classificação reveladas pelas matrizes de confusão dos experimentos com todas as classes (Figuras 3.5.1, 3.6.1 e 3.7.1) e também observáveis no gráfico da Figura 4.1.1.

Essas tendências sugerem a necessidade de uma melhora no pré-processamento dos exemplos, não no sentido de formatação dos dados, mas de um maior equilíbrio nas proporções das classes presentes nos exemplos. A redução do número de exemplos das classes predominantes poderia ser uma estratégia, entretanto ocorreria uma diminuição significativa do volume de dados, outra estratégia possível seria a geração de exemplos artificiais [Batista 2003].

A partir dos experimentos constataram-se os melhores desempenhos, tanto em termos de acurácia quanto em termos de *F-measure*, com as técnicas de máquina de vetores de

suporte SVM, sugerindo que a estratégia de melhorar os hiperplanos de separação das classes utlizadas sejam responsáveis por essa melhoria nos resultados. A expectativa inicial, de que o algoritmo genético realizasse a otimização do parâmetro γ do SVM, também pode ser verificada em termos experimentais com os dados disponíveis. Os gráficos das Figuras 4.1.1 e 4.1.2 resumem os resultados de acurácia e F-measure de 32 experimentos. O eixo das abscissas de cada gráfico está organizado por combinações do tipo de amostra, do subconjunto de classes e da técnica utilizada em cada experimento, formando séries de quatro experimentos (AD, ML, SVM e SVMAE). percebe-se a superioridade do SVMAE.

Comparando-se os resultados das medidas obtidas nos experimentos com a amostra balanceada (Tabela 3.3.3) com aqueles obtidos com a amostra não balanceada (Tabelas 3.8.2, 3.7.2, 3.6.2 e 3.5.2) constata-se maior uniformidade nos valores das medidas de cobertura e acurácia para a amostra balanceada, com aumento de alguns valores de cobertura e diminuição da acurácia. Essa diminuição da acurácia era esperado pela diminuição do viés de classificação das classes mais numerosas presentes na amostra não balanceada.

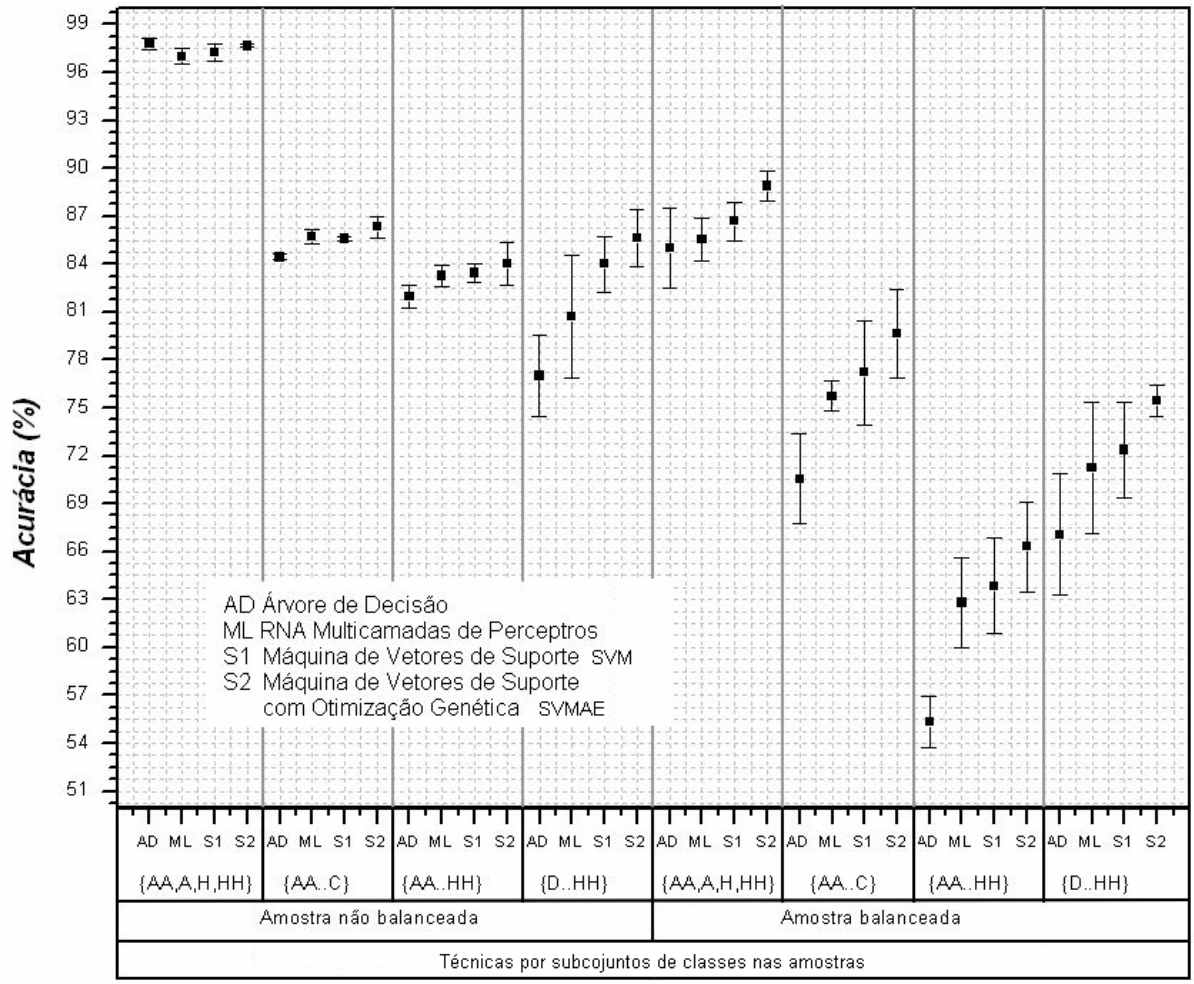

Figura 4.1.1 Acurácia das técnicas nos subconjuntos de classes.

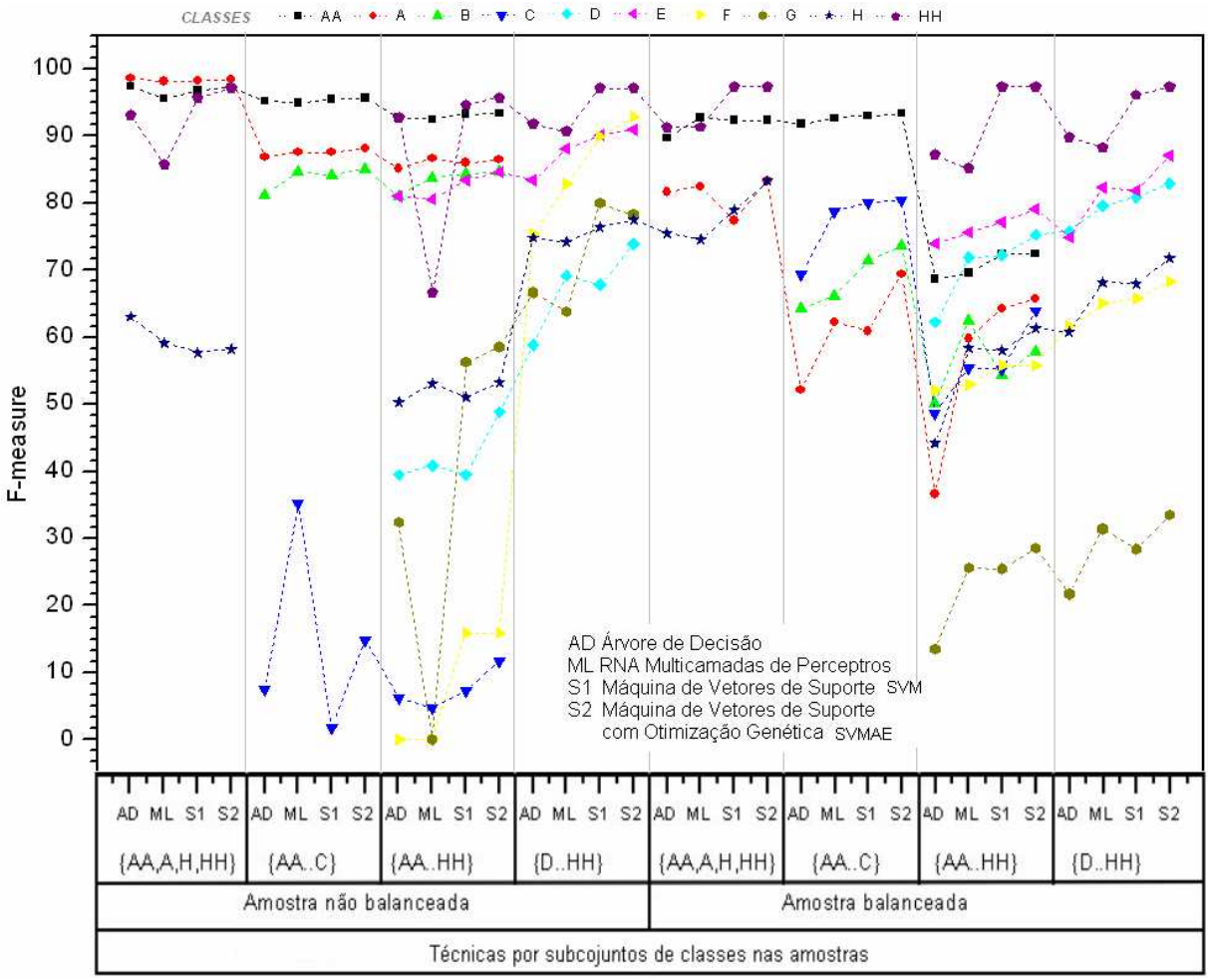

Figura 4.1.2 Medidas de F-measure com α=0,5 das técnicas nas classes.

4.2. Proposta para Trabalhos Futuros

Alinhado com os resultados obtidos, em termos de técnicas, trabalhos futuros poderiam explorar: a aplicação de paradigmas evolucionários com as outras técnicas utilizadas AD e MLP; o aperfeiçoamento de seu uso com SVM, limitado no presente trabalho ao parâmetro γ do SVM. Extensões do presente trabalho poderiam incluir o paradigma nebuloso (*Fuzzy*) para aumentar a flexibilidade de classificação, possivelmente numa arquitetura neuro-fuzzy ou neuro-fuzzyAE e o uso de exemplos gerados artificialmente para balanceamento de classes.

Em termos de contexto do problema, trabalhos futuros poderiam estudar a migração de notas durante o ciclo de vida da operação de crédito.

5. Referências Bibliográficas

[Angelini et al., 2007] ANGELINI, E.; DI TOLLO, G.; ROLI, A. ***A neural network approach for credit risk evaluation***, The Quarterly Review of Economics and Finance (2007), doi:10.1016/j.qref.2007.04.001

[CMN 1999] CONSELHO MONETARIO NACIONAL. ***Resolução 2682, de 21 de dezembro de 1999***. Dispõe sobre critérios de classificação das operações de crédito e regras para constituição de provisão para créditos de liquidação duvidosa. Brasília, 1999.

[Bacen 2008] BANCO CENTRAL DO BRASIL. Relatório de Estabilidade Financeira. Volume 7 número 1, Maio de 2008. Brasilia, 2008.

[Batista 2003] BATISTA, G. E. A. P. A. ***Pré-processamento de Dados em Aprendizado de Máquina Supervisionado***. Tese de Doutorado, ICMC-USP. São Carlos: 2003. Disponivel em <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102003-160219/>. Acessado em 27 out 2008.

[BCBS 2005] BASEL COMMITE ON BANKING SUPERVISION, ***Studies on the Validation of Internal Rating Systems. Working Paper No.14***. Disponível em <http://www.bis.org/publ/bcbs_wp14.htm >. Acessado em: 07 jul. 2008.

[Borth 2007] BORTH, N. L. ***Inadimplência****: Construção de Modelos de Previsão.* São Paulo: Editora Nobel, 2004.

[Carvalho et al., 2005] DE CARVALHO, A. P. L. F.; LACERDA, E. G. M.; BRAGA, A. P.; and LUDERMIR, T. ***Evolutionary Radial Basis Functions for Credit Assessment***, pages 167-181, Vol. 22, No. 3, Applied Intelligence, ISNN 0924-669X, Springer, May 2005.

[Chiavenato 2001] CHIAVENATO, I.; ***Introdução à Teoria Geral da Administração***, 4. ed. São Paulo: Editora Campus, 2001.

[Eberhart 2007] EBERHART, R. C., SHI, Y. ***Computational Intelligence****: Concept to Implementations*. pp 404-406. New York: Morgan Kaufmann, 2007.

[Engelmann e Rauhmeier 2006] ENGELMANN, B.; RAUHMEIER, R.; ***The Basel II Risk Parameters****: Estimation, Validation and Stress Testing.* Berlin: Springer, 2006.

[Goldberg 1989] GOLDBERG, D.E. ***Genetic Algorithms in Search, Optimization and Machine Learning.*** Addison Wesley, New York: 1989.

[Heffernan 2005] HEFFERMAN, S. ***Modern Banking***, John Wiley & Sons, 2005.

[Haykin 2001] HAYKIN, S. ***Redes Neurais: Princípios e Prática***. Tradução de Paulo Martins Engel. 2. ed. Porto Alegre: Bookman, 2001. Título original: Neural Networks: a comprehensive foundation, 2/E.

[Jain et al., 2008] JAIN L. C.; SATO, M.; VIRVOU, M.; TSIHRINTZIS G. A.; BALAS, V. E.; ABEYNAYAKE C. ***Computational Intelligence Paradigms****: Innovative Applications*. Berlin: Springer-Verlag, 2008.

[JDBC 2008] Página oficial do Conector Java para Dados PostgreSQL na Internet, Disponível em <http://jdbc.postgresql.org/download.html>. Acesso em: 07 ago. 2008

[Manning et al.2008] MANNING, C.D.; RAGHAVAN P.; SCHUTZE, H. ***Introduction to Information Retrieval***. Cambrige University Press, New York: 2008.

[Mierswa et al., 2006] MIERSWA, I.; WURST, M.; KLINKENBERG, R.; SCHOLZ, M.; and EULER, T. YALE: ***Rapid Prototyping for Complex Data Mining Tasks***, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.

[Mishkin 2004] MISHKIN, F.S., ***The Economics of Money, Banking and Financial Markets***. pp. 260-269. Addison-Wesley, 7th.ed. USA: 2004.

[Mitchell 1997] MITCHELL, T.M. ***Machine Learning***. McGraw Hill, New York: 1997.

[Munakata 2008] MUNAKATA, T. ***Fundamentals of the New Artificial Intelligence****: Neural, Evolutionary, Fuzzy and More*. Cleveland, Ohio, USA: Springer-Verlag, 2008.

[Oldcorn e Parker 1998] OLDCORN, R.; PARKER, D. ***Decisão Estratégica para Investidores***, pág. 21. São Paulo: Editora Nobel, 1998.

[Postgresql 2008] Página oficial do Banco de Dados PostgreSQL na Internet, Disponível em < http://www.postgresql.org/>. Acesso em: 05 ago. 2008

[RapidMiner 2008] Página oficial do software RapidMiner na Internet. Disponível em: <http://rapid-i.com/index.php?lang=en>. Acesso em: 05 ago. 2008

[Rezende et al., 1998] REZENDE S.; HORST, P.; PADILHA, T.; ROCHA, C.; and DE CARVALHO; A. C. P. F. ***Knowledge Acquisition Using Symbolic and Connectionist Algorithms for Credit Evaluation***. Proceedings of the IEEE World Congress on Computational Intelligence, WCCI´98, Anchorage, USA, May 1998.

[Rochet 2007] ROCHET, J. ***Why Are there So Many Banking Crisis***. New Jersey, USA: Princeton University Press, 2007.

[Sun e Wang 2005] SUN, M.; WANG, S. ***Validation of credit rating model****: a preliminary look at methodology and literature review*, China, 2005.

[Witten e Frank 2005] – WITTEN, I. H.; FRANK, E. ***Data Mining Practical Machine Learning Tools and Techniques***. Morgan Kaufmann, 2005

[Weka 2008] Página oficial do software Weka na Internet, Disponível em <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 01 ago. 2008

[Yip e Dempster 2005] YIP, G.; DEMPSTER, A. ***Using the Internet to Enhance Global Strategy.*** European Management Journal, Volume 23, Issue 1, February 2005, Pages 1-13 doi:10.1016/j.emj.2004.12.005

[Yu et al., 2008] YU, L.; WANG, S.; LAI, K. K.; ZHOU, L. ***Bio-Inspired Credit Risk Analysis: Computational Intelligence with Support Vector Machines***, Springer-Verlag, Berlin Heidelberg, 2008.