UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
DEPARTAMENTO DE MÁQUINAS COMPONENTES, E SISTEMAS INTELIGENTES
Laboratório de Controle e Sistemas Inteligentes

Modelos Para Previsão do Risco de Crédito

Cristiano Roberto de Souza
Orientador : Prof. Dr. Gilmar Barreto

Dissertação de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica. Área de concentração: Automação

Comissão Examinadora :
Prof. Dr. Gilmar Barreto - DMCSI-FEEC-UNICAMP - Presidente
Profa. Dra. Cicília Yuko Wada - IMECC - UNICAMP
Prof. Dr. Akebo Yamakami, - DT-FEEC-UNICAMP

Campinas, São Paulo, Brasil
Março de 2010

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

So89m

Souza, Cristiano Roberto de
Modelos para previsão do risco de crédito / Cristiano Roberto de Souza. –Campinas, SP: [s.n.], 2010.

Orientador: Gilmar Barreto.
Dissertação de Mestrado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Créditos - Avaliação de riscos. 2. Redes neurais (Computação). 3. Processamento de linguagem natural (Computação). I. Barreto, Gilmar. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês:	Models to forecast financial risk
Palavras-chave em Inglês:	Computer - human interaction, Neural networks, Natural language processi
Área de concentração:	Automação
Titulação:	Mestre em Engenharia Elétrica
Banca Examinadora:	Akebo Yamakami, Cicília Yuko Wada
Data da defesa:	31/03/2010
Programa de Pós Graduação:	Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidato: Cristiano Roberto de Souza

Data da Defesa: 31 de março de 2010

Título da Tese: "Modelos para Previsão do Risco de Crédito"

Prof. Dr. Gilmar Barreto (Presidente): _____

Profa. Dra. Cicília Yuko Wada: _____

Prof. Dr. Akebo Yamakami: _____

Resumo

Os modelos computacionais para previsão do risco financeiro têm ganhado grande importância desde 1970. Com a atual crise financeira os governos tem discutido formas de regular o setor financeiro e a mais conhecida e adotada é a de Basiléia I e II, que é fortemente suportada por modelo de previsão de risco de crédito. Assim este tipo de modelo pode ajudar os governos e as instituições financeiras a conhecerem melhor suas carteiras para assim criarem controle sobre os riscos envolvidos.

Para se ter uma idéia da importância destes modelos para as instituições financeiras a avaliação de risco dada pelo modelo é utilizada como forma de mostrar ao Banco Central a qualidade da carteira de crédito. Através desta medida de qualidade o Banco Central exige que os acionistas do banco deixem depositados um percentual do dinheiro emprestado como garantia dos empréstimos duvidosos criando assim o Índice de Basiléia.

Com o objetivo de estudar as ferramentas que atualmente auxiliam no desenvolvimento dos modelos de risco de crédito iremos abordar:

1. Técnicas tradicionais Estatísticas,

2. Técnicas Não Paramétricas,

3. Técnicas Computação Natural

Palavras-chave: *Credit Scoring, Behaviour Scoring*, Redes Neurais, Computação Natural, Risco de Crédito.

Abstract

The computer models to forecast financial risk have gained great importance since 1970 [1]. With the current crisis Financial government has discussed ways to regulate the financial sector, and the most widely known and adopted form is Basel I and II, which is strongly supported by the forecasting models of credit risk. This type of model can help governments and financial institutions to better understand their portfolios so they can establish control over the risks involved.

To get an idea of the importance of this models for financial institutions, the risk assessment given by the model is used as a way of showing the central bank quality of credit portfolio. This measure of quality the Central Bank requires that the shareholders of the bank no longer paid a percentage of the borrowed money as collateral in problem loans and thus creating the index of Basel.

In order to study the tools that actually support the development to models of credit risk we will cover:

1. Statistics techniques,
2. Non-Parametric Techniques,
3. Natural Computation Techniques

Keywords: Credit Scoring, Behavior Scoring , Neural Networks, Natural Computing, Credit Risk. Technical Reports.

Agradecimentos

Agradeço primeiramente ao meu orientador pelo tempo dedicado, pela paciência e pela amizade. Sinto que meus conhecimentos sobre a pesquisa e a vida em muito foram expandidos pela nossas conversas.

A minha família pelo apoio durante esta jornada.

A minha esposa Letícia pela compreensão das horas dedicadas.

A Minha Esposa

Sumário

Lista de Figuras

Lista de Tabelas

Trabalhos Publicados Pelo Autor

1. Cristiano Roberto de Souza, Gilmar Barreto. “7 Brazilian Conference on Dynamics Controland Applications, May 2008, Presidente Prudente, Brazil”. *Ensemble Methods and Selection of Models for Credit Risk Control*.

Capítulo 1

Introdução

1.1 Introdução

A falta de previsão do risco de crédito é um dos principais componentes da atual crise financeira, pois as grandes instituições financeiras tiveram perdas de crédito não previstas em um curto espaço de tempo. Isso significa que uma grande quantidade de clientes deixaram de pagar as prestações de seus empréstimos levando as instituições financeiras a não terem dinheiro suficiente para honrar o dinheiro depositado nas contas dos outros clientes.

A existência de modelos de crédito, que são usados desde 1970 [1], permitem uma correta avaliação dos risco envolvido nas operações de Crédito permitem que as instituições possam garantir a otimização do risco retorno para a empresa e ao mesmo tempo terem seus riscos conhecidos e controlados.

Os modelos utilizados para se prever o risco de um cliente podem ser classificados como *Credit Scoring e Behaviour Scoring*. O primeiro é utilizado quando o cliente é novo na instituição e pouca informação sobre ele é conhecida, basicamente são utilizadas informações fornecidas pelo próprio cliente. A segunda classe de modelo é utilizada quando o cliente já tem informações de comportamentos internas a instituição hà algum tempo e assim é possível se utilizar estas informações, o que faz com o modelo tenha uma maior precisão.

Uma grande discussão sobre como os governos devem fiscalizar e controlar os bancos deve ser iniciada após o controle dos efeitos da crise. Atualmente existe o acordo internacional de Basiléia [2] que visa a regulamentação do setor. Este acordo tem três grandes pilares.

1. Capital
2. Revisão pelo Supervisor
3. Disciplina de Mercado.

Neste trabalho discutir algumas da técnicas para se medir o risco de crédito, apresentarmos alguns exemplos de utilização e um modelo computacional capaz de auxiliar na tomada de decisões na concessão do crédito, e também na avaliação do risco dos créditos já concedidos. Estas técnicas são a base para que o primeiro pilar de Basiléia seja atendido .

1.2 Organização do Texto

No segundo capítulo abordamos alguns conceitos de construção dos modelos de Crédito, assim como sua história, e assim as diferenças entre *Credit Scoring*e *Behaviour Scoring*.

No terceiro capítulo discutimos os conceitos das técnicas mais utilizadas assim como exemplos de cálculo e utilização das mesmas.

No quarto capítulo apresentamos a proposta de uma técnica de computação natural e uma aplicação pratica.

No quinto capítulo fizemos uma comparação dos resultados dos modelos utilizados.

No sexto capítulo apresentamos as conclusões e sugestões para trabalhos futuros.

Capítulo 2

Credit Scoring, Behaviour Scoring

Os modelos de *Scoring* são as classificações do tipo de modelo utilizado para ajudar as organizações a decidir sobre a concessão de crédito e mais atualmente também o quanto vai ser cobrado pelo crédito.

Existem dois tipos de modelos utilizados para tomada de decisão:

- ***Credit Scoring*** : A instituição deve ou não conceder crédito a novos clientes.
- ***Behaviour Scoring***: A instituição deve concordar em aumentar o crédito de um cliente? Quais as ações que o departamento de marketing deve tomar para aumentar o lucro com o cliente? Se o cliente começa a dar sinais de que vai atrasar, quais as ações que a empresa deve tomar?

Entre as áreas atualmente em desenvolvimento no *credit scoring* podemos citar :

- áreas que criam técnicas para os modelos se adaptarem as mudanças da economia,
- áreas que tentam maximizar o lucro,
- áreas buscando metodologias para se fazer a inferência de clientes não contratados.

Na primeira área as técnicas estão preocupada em identificar o risco do cliente de modo que quando ocorrer uma mudança na economia elas possam ser automaticamente incorporadas, ou quando

o cliente mudar seu comportamento o modelo possa perceber tal mudança e corrigir a previsão do risco do cliente.

A segunda área não está mais preocupadas em reduzir as taxas de sinistro mas sim em aumentar o retorno sobre o capital emprestado, por exemplo, oferecendo taxas maiores para clientes com maior risco ou exigindo garantias.

Na terceira área busca-se métodos de diminuir ou acabar com o efeito de que a somente uma parte dos clientes é permitido contratar, o que faz com que se a população utilizada no desenvolvimento do modelo não seja representativa da população solicitante de crédito criando se um vício amostral. Este vício causa que características ruins no modelo atual podem se tornar neutras em modelos futuros, pois os clientes que criavam esta característica deixam de contratar e aparecer na amostra.

A informações disponíveis para se construir tais modelos de *Scoring* são:

- Cadastrais,
- De mercado
- Informações sobre a operação.
- Informações de comportamento de pagamento.
- Quantidade de produtos.
- Relacionamento.

As instituições guardam milhões de informações de clientes que no passado tiveram crédito aprovado e agora podem ser utilizadas para se construir os modelos de previsão. No entanto existe um problema com estas informações, elas são viciadas pois existem muitos clientes que solicitam crédito e são negados não sendo possível assim obter o desempenho destes clientes. Assim as informações existem somente para uma parte da população alvo, fato que de ser levando em conta na construção dos modelos e principalmente na criação das previsões das taxa de inadimplência, que atualmente são exigidas pelos bancos centrais dos pais que já adotaram o acordo de Basiléia II.

2.1 História

Credit Scoring é essencialmente uma ferramenta de reconhecimento de diferentes grupos na população quando não é possível verificar a característica que separa os grupos mas características correlacionadas. A idéia de discriminar grupos em uma população foi introduzida por Fisher [3].

Os primeiros estudos para se utilizar técnicas de discriminação para separar entre bons e maus pagadores foram feitos na década de 40 [4]. Nesta época as casas financeiras perceberam que poderiam utilizar as técnicas de discriminação de grupos para melhorar o seu desempenho, porém muitos especialistas foram alocados em pesquisas militares, o que acabou atrasando o seu desenvolvimento. Inicialmente as casas financeiras usavam regras julgamentais para decidir o crédito. Após a guerra as casas financeiras encontraram nos modelos estatísticos grandes benefícios para aumentar seus lucros e controlar o risco. A primeira consultoria foi formada por Bill Fair e Earl Isaac em 1950.

Em 1960 foram construídos os primeiros modelos de *Credit Scoring* e isso permitiu ampliar a oferta de Cartão de Crédito, pois estes modelos permitem a análise de grandes volumes de clientes em pouco tempo, automatizando assim a decisão de crédito. Não demorou muito para as instituições financeiras perceberem que os *Credit Scoring* eram muito mais eficientes para separar bons e maus pagadores do que qualquer método julgamental, isso fez que as taxas de sinistro reduzirem 50 % [5].

Com o sucesso de aplicação dos *Credit Scoring* nos Cartões de Crédito as instituições começaram a utilizar os modelos de Crédito para outros tipos de empréstimos como Cheque Especial, Financiamento de Veículos e Crédito Pessoal Parcelado.

Nos década de 90 os modelos de classificação começaram a ser utilizados para aumentar o retorno em campanhas de publicidade, onde os modelos deciden quais os clientes tem maior probabilidade de responder positivamente a campanha reduzindo assim o custo de emissão da campanha e aumentando o retorno ao mesmo tempo. No ramo de seguros também estão sendo utilizados como ferramentas para atribuir um valor ao risco, ou seja, de acordo com o perfil do cliente, como sua localidade, idade, profissão e outros cobra-se prêmios diferentes.

Mais recentemnente a resposta do modelo tem se expandido além do risco de o cliente pagar ou

não um empréstimo, mas também a questões tais como: O cliente irá fechar sua conta? A transação sendo efetuada é uma fraude? O cliente irá aumentar a utilização do cheque especial? O cliente que comprou tal produto também comprou qual outro? Assim as empresas cada vez mais podem especializar o atendimento a cliente oferecendo o que ele quer e quando ele quer.

Com os avanços da modelagem computacional de dados foi possível utilizar técnicas mais avançadas como regressão logística e programação linear na construção dos modelos. Mais recentemente tem se testado técnicas de inteligência artificial e de computação natural. Estas técnicas não só exigem computadores potentes como também pessoal especializado para a construção e utilizacão dos modelos em todos os níveis hierárquicos da instituições. O avanço computacional também permitiu o armazenamento de gigantescas quantidades de informações em bancos de dados, para que assim possam ser utilizadas nos modelos melhorando seus desempenhos.

No Brasil, as instituições financeiras começaram a utilizar as técnicas de *credit scoring* nos meados da década de 90 quando a inflação foi controlada pelo Plano Real e elas deixaram de ganhar dinheiro nas operação de *Over Night* e foram obrigadas a gerar receita com empréstimos a pessoas Físicas e Jurídicas. Inicialmente foram feitos estudos para se conhecer as carteiras de crédito das instituições e criação de bancos de dados históricos para se construir os modelos. Os primeiros modelos foram trazidos por consultorias estrangeiras, porém logo percebeu se que o público brasileiro tinha perfil muito diferente e os modelos precisavam serem construídos internamente.

As instituições usam atualmente os modelos para medir o risco de crédito, *marketing*, risco operacional entre outros. Muitas delas tem áreas específicas para construção dos modelo e de pesquisas de novas técnicas e metodologia de construção, além de um forte investimento em banco e dados e auditoria das fórmulas utilizadas.

Com a determinação que todos os bancos até 2011 terão que cumprir os requisitos do acordo de Basiléia II, os modelos de risco de crédito ganham ainda mais importância, pois o acordo é fortemente baseado na auto regulamentação das instituições e, para isso, elas tem que ter modelos de previsão de risco eficientes e com documentação atualizada.

2.2 *Credit Scoring*

Os *Credit Scoring*, em geral são utilizados para se calcular o risco do cliente no processo de decidir se vai ou não ser concedido o empréstimo ao cliente novo. O *credit score* permite que a decisão tenha seu risco medido e controlado, ou seja, a instituição sabe antecipadamente qual o risco do cliente e assim pode decidir quanto quer perder/ganhar com esta carteira de clientes. Assim é feito também um estudo financeiro que completa o estudo para decidir quais clientes serão aprovados ou não. Por exemplo, aprovar só os clientes com risco de não pagar inferior a 10% .

Para se construir o modelo geralmente são utilizados os seguintes tipos de informações:

- **Informações Cadastrais**: sexo, estado civil e outras,
- **Informações da Operação**: número de parcelas, valor, tipo de garantia e outras,
- **Informações de Mercado**: negativação no SPC (Serviço de Proteção a Crédito) e SERASA
- **Informações Internas**: situação de outros produtos de crédito, histórico de pagamento em créditos anteriores, informações de parceiros.

A figura 2.1 ilustra o funcionamento do *credit scoring*.

Fig. 2.1: Diagrama do *Credit Scoring*

O modelo é construído a partir do comportamento de uma amostra de clientes que contrataram(Assinar o Contrato) o produto no passado, a qual poder ser de algumas centenas a milhares de clientes. Normalmente esse não é um problema pois instituições financeiras costumam ter milhões de clientes em sua base histórica. Para cada cliente na amostra são necessárias as informações

descritas anteriormente e seu histórico de crédito por um tempo fixo após a contratação do produto, normalmente 12, 18 ou 24 meses dependendo do tipo de empréstimo. Com este histórico é possível decidir se o cliente é um mau pagador ou não, onde a definição de mau pagador é geralmente dada quando o cliente atrasa mais de 90 dias no pagamento da parcela ou dos juros no caso de rotativos. A Figura 2.2 ilustra este processo onde N é o mês de referência de contratação.

Fig. 2.2: Dados Selecionados do Cliente para *Credit*

O horizonte de tempo para se prever o risco do cliente é uma questão a ser estudada, no entanto este item será objeto de estudos futuros. Estudos mostram que a taxa de sinistro como função do tempo tende a se estabilizar após 12 meses. Portanto períodos menores que 12 meses podem não refletir a taxas reais. No entanto períodos superiores a 24 meses utilizados para predição não são confiáveis pois a população não é estática e neste período as suas características podem ter mudado.

Um exemplo de modelos de *Credit Scoring* muito utilizado é para a concessão de Cheque Especial, no qual um cliente solicita um limite de crédito na abertura de uma conta corrente.

2.3 *Behaviour Scoring*

Os modelos de *Behaviour Scoring* são utilizados para medir o risco do cliente em um produto já contratado, ou seja, como o cliente esta se comportando na utilização deste produto. A diferença entre os modelos de *Behaviour* e *Credit Scoring* é que os modelos de *Credit Scoring* utilizam principalmente as informação de cadastro do cliente e os modelos de *Behaviour* utilizam informações de como esta sendo utilizado o produto, como por exemplo, dias de atraso, saldo da conta corrente, etc. Por utilizar informações de comportamento e mais atualizadas os modelos de *Behaviour Scoring*

tem uma capacidade de identificar os maus pagadores até duas vezes maior que os modelos de *Credit Scoring*, ou seja sua precisão na probabilidade de pagamento da dívida é maior.

Fig. 2.3: Diagrama do Behaviour Scoring

Os modelos de *Behaviour Scoring*, assim como os modelos *Credit Scoring*, utilizam amostras passadas de comportamento dos clientes. Porém neste caso fixa se um ponto no tempo e a partir deste ponto no tempo observa se o comportamento do cliente para saber se ele é bom pagador ou não, e antes deste ponto todas as informações de pagamento disponíveis. As informações utilizadas nos modelos são coletadas anteriormente a este ponto no tempo, como na figura 2.4.

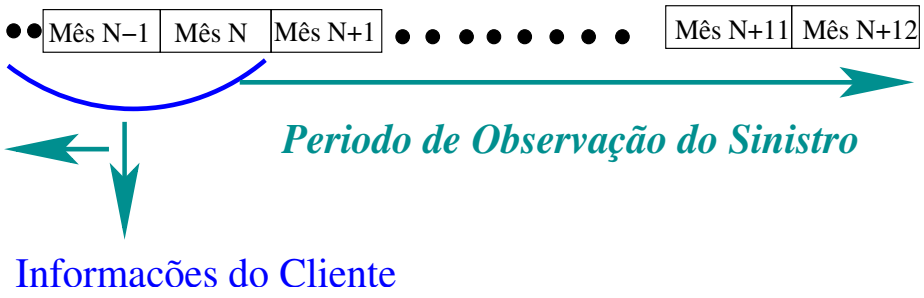

Fig. 2.4: Dados Selecionados do Cliente para *Behaviour*

Um exemplo de *Behaviour Scoring* é quando este é utilizado para a prevenção, ou seja, quando se verifica que um cliente tem grande chance de não quitar a sua dívida atual é tomada uma atitude preventiva como por exemplo oferecer um parcelamento da dívida com juros menores. Dessa forma, enquanto os modelos de *Credit Scoring* controlam o risco exposto a novos clientes os modelos de *Behaviour Scoring* oferecem uma ferramenta para gerenciamento da carteira atual de clientes.

Os modelos de *Collect Scoring* são utilizados na recuperação de clientes inadimplentes, completando assim o ciclo de vida do cliente numa instituição financeira. Estes modelos são utilizados para

se prever quais os clientes devem ser cobrados primeiro e como esses clientes devem ser cobrados a partir do momento que eles entram em cobrança, aumentando assim a eficiência da cobrança. Nos modelos de cobrança um fator ao qual as instituições estão preocupadas é em identificar os clientes que aprenderam o processo de cobrança e agora estão viciados em descontos.

Os modelos de *Collect Scoring* são casos particulares dos modelos de *Behaviour Scoring*, pois estes usam informações de comportamento recente do cliente porém o objetivo neste caso é orientar a ação de cobrança, identificando os clientes que não é necessário cobrar e quais os clientes é preciso uma ação mais intensa.

A Figura 2.5 descreve o ciclo de vida de um cliente e com qual modelo ele está sendo avaliado a cada momento.

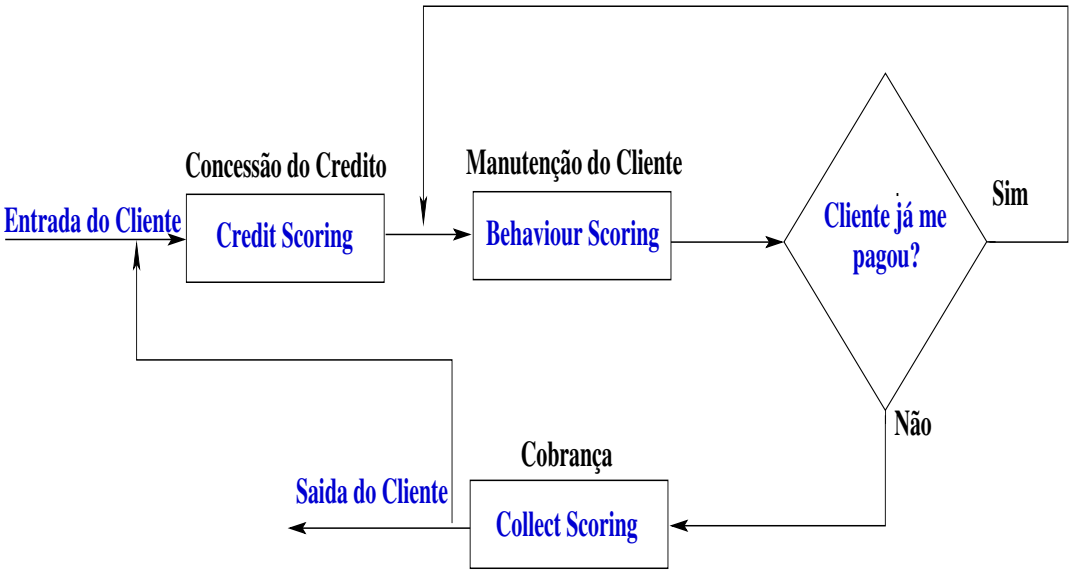

Fig. 2.5: Ciclo de Vida de Um Cliente

Através da figura 2.5 pode se ver que a instituição está a todo momento re-avaliando o risco do cliente com alguns dos tipos de modelos descritos anteriormente. Com está avaliação é possível periodicamente rever a necessidade de alocação e capital e informar ao regulador a qualidade da carteira de crédito.

2.4 Sumário

Neste capitulo apresentamos umabreve histórico da utilização dos modelos de crédito assim como algumas definições de *Credit Scoring* e *Behviour Scoring*. Essas definições parecem simples, mas são de extrema importância para que na pratica não ocorram erros de, por exemplo, selecionar variáveis futuras para se construir um modelo.

Capítulo 3

Algumas Técnicas Utilizadas para modelos de *Behaviour e Credit Score*

Os modelos de *Behaviour e Credit Score* tem como principal objetivo inferir o risco do cliente pagar o empréstimo a partir de uma amostras destes clientes com crédito no passado. Para isso fazem uso das técnicas que a partir de algumas características do objeto em estudo e uma amostra deles seja possível tal inferência.

Umas da áreas mais conhecidas atualmente que utiliza estimativas de risco é a pesquisa médica. Na medicina, um médico pode está interessado em saber qual o risco de um paciente contrair uma doença, por exemplo um câncer. Ele tem uma amostra de pessoas com câncer e uma amostra de pessoas saudáveis, assim ele coleta informações de ambas as amostras e calcula o modelo de risco.

A técnica normalmente utilizada por para estes estudos é a regressão logística, pois ela tem uma série de propriedades interessantes, que serão discutidas na seção 3.1 assim como sua teoria, tornando se assim uma das técnicas mais utilizadas atualmente em modelos de *Scoring*.

Outra área reconhecidamente interessada em relacionar variáveis de entrada com uma ou mais saídas é a engenharia. Os interesses vão desde em controle de processos, no qual as informações de sensores indicam condições e uma ação deve ser tomada para corrigir o processou ou torna-lo mais eficiente, até métodos de inteligência artificial na qual queremos que equações matemáticas tenham

propriedades semelhantes as do cérebro humano.

Nos modelos de crédito não são relacionados uma séria de informações provenientes de sensores, neste caso características dos clientes, com as ação de conceder ou não crédito, criando assim um sistema de controle do risco envolvido. Deste modo as técnicas de redes neurais e outras podem ser aplicadas ao processo de modelos de crédito.

Neste capítulo pretendemos utilizar técnicas de aprendizado de máquina, ou inteligência artificial, para que possamos criar modelos que não tenham objetivos conhecidos como de reduzir o erro da estimativa de risco, e sim o objetivo de maximizar o lucro dos acionistas apontando caminhos como eles fariam. Desta maneira podemos adotar técnicas de Estratégias Evolutivas para construir modelos que tenham tais propriedades.

3.1 Regressão Logística

A regressão logística é uma ferramenta muito conhecida nas áreas biológicas e médicas, pois ela permite descrever razoavelmente a relação entre um resultado (variável dependente ou variável resposta) e as características do conjunto de variáveis independentes (preditoras ou explicativas). No contexto de crédito ela é muito útil pois permite utilizar estas relações para explicar os que faz um cliente ser bom ou ruim, assim torna possível tirar conclusões de quais características do cliente o modelo está utilizando para dizer se ele é bom ou mau pagador facilitando assim a interpretação por pessoas não técnicas. Ela também fornece uma estimativa linear da probabilidade de pagamento do cliente.

O que diferencia um modelo de regressão logística do modelo de regressão linear é a variável resposta, que na regressão logística é binária ou dicotômica e na regressão linear ela é contínua. está diferença entre regressão logística e linear é refletida tanto na escolha de um modelo paramétrico como nas suas suposições. Uma vez que está diferença é detectada, os métodos empregados em uma análise usando regressão logística seguem os mesmos princípios usados na regressão linear.

Muitas distribuições têm sido propostas para serem usadas na análise de uma variável resposta

dicotômica. Existe duas razões principais para se escolher a distribuição logística:

1. Do ponto de vista matemático ela é extremamente flexível e fácil de ser usada,

2. Ela é adequada para a interpretação dos resultados.

Supondo a existência de um conjunto de dados formado por uma resposta Y e uma matriz de variáveis X, a fim de simplificar a notação será usado a quantidade $\pi(x) = E(x)$, onde $E(x)$ é o operador esperança, para representar a média condicional de Y dado X quando a distribuição logística é usada. A fórmula específica do modelo de regressão que será usado é:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{3.1}$$

onde β são os parâmetros e x são as variáveis explicativas.

A transformação de $\pi(x)$ que será o centro do estudo da regressão logística é a transformação logito. está transformação em termos de $\pi(x)$ é dada por:

$$g(x) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x \tag{3.2}$$

A importância desta transformação é que $g(x)$ tem muitas propriedades desejáveis de um modelo de regressão linear. O logito é linear em seus parâmetros, é contínuo eestáa definido entre $-\infty$ e $+\infty$.

Outra diferença importante entre o modelo de regressão linear e logístico é com respeito a distribuição condicional da variável resposta. No modelo linear uma observação de variável resposta pode ser expresso como $y = E(y/x) + \epsilon$. A quantidade ϵ é chamado de erro e expressa como um desvio da observação em relação a média condicional. A suposição comum é que ϵ segue uma distribuição normal com média $E(Y/x)$, e uma variância que é constante. Isso implica que a distribuição condicional da variável resposta dado x será normal com média $E(Y/x)$, e variância que é constante. Este não é o caso com uma variável resposta dicotômica que pode assumir apenas dois valores. Nesta situação podemos expressar o valor da variável resposta dado x como $y = \pi(x) + \epsilon$. Aqui a quantidade ϵ pode assumir um dos dois possíveis valores:

- se $y = 1$, então $\epsilon = 1 - \pi(x)$ com probabilidade $\pi(x)$
- se $y = 0$, $\epsilon = \pi(x)$ então com probabilidade $1 - \pi(x)$

Assim, ϵ tem uma distribuição com média zero e variância igual a $\pi(x)\,[1 - \pi(x)]$, isto é, a distribuição condicional da variável resposta segue uma distribuição binomial com probabilidade dada pela média condicional $\pi(x)$.

Para ajustar o modelo é necessário supor uma amostra de n observações independentes do par (X_i, Y_i), i =1, 2,. . . , n, onde Y_i denota o valor de uma variável resposta dicotômica e X_i o valor da variável independente para o i-ésimo sujeito. Além disso, assuma que a variável resposta tenha sido codificada como zero ou 1, representando ausência ou presença da característica, respectivamente. Ajustar o modelo de regressão na equação para um conjunto de dados requer que os valores dos parâmetros desconhecidos, β_0 e β_1 sejam estimados

Na regressão linear o método mais usado para estimação de parâmetros é o de mínimos quadrados. Neste método, encontram se valores dos parâmetros β_0 e β_1 que minimizam a soma dos quadrados de desvios de valores observados de Y dos valores preditos baseados no modelo. Sob as suposições usuais para regressão linear o método de mínimos quadrados produz estimadores com um número desejável de propriedades estatísticas. No entanto, quando o método de mínimos quadrados é aplicado para um modelo com um resultado dicotômico os estimadores não têm mais estas mesmas propriedades.

O método geral de estimação que leva à função de mínimos quadrados sob o modelo de regressão linear (quando os erros são normalmente distribuídos) é conhecido como estimativa de máxima verossimilhança. Este método fornecerá o fundamento para nossa aproximação, a estimação com o modelo de regressão logística. Em um sentido mais geral, o método de máxima verossimilhança produz valores para os parâmetros desconhecidos que maximizam a probabilidade de obtenção dos conjuntos de dados observados, assim a partir da observação dos dados podemos encontrar os parâmetros desconhecidos.

Para aplicar este método devemos primeiro construir uma função, chamada de função de ve-

rossimilhança. está função expressa a probabilidade dos dados observados como uma função de parâmetros desconhecidos. Os estimadores de máxima verossimilhança destes parâmetros são escolhidos para serem aqueles valores que maximizem está função. Assim, os estimadores resultantes são aqueles em que os dados observados tem maior chance de acontecer.

Se Y é codificado como zero ou 1, então a expressão para $\pi(x)$ dada na equação 3.1 fornece a probabilidade condicional que Y é igual a 1 dado x, este será denominado como $P(y = 1/x)$. Dai segue que a quantidade $1 - \pi(x)$ dá a probabilidade condicional que Y é igual a zero dado x, $P(y = 0/x)$. Assim, para aqueles pares (x_i, y_i), onde $y_i = 1$ a contribuição para a função de verossimilhança é $\pi(x_i)$, e para $y_i = 0$ a contribuição para a função de verossimilhança é $1 - \pi(x_i)$. Uma forma conveniente de expressar a contribuição de verossimilhança para o par (x_i, y_i) é pelo termo:

$$\zeta(x_i) = \pi(x_i)^{y_i} * [1 - \pi(x_i)]^{1-y_i} \tag{3.3}$$

Considerando as observações independentes, a função de verossimilhança é obtida como o produto dos termos dados acima:

$$l(\beta) = \prod_{i=1}^{n} (\zeta(x_i)) \tag{3.4}$$

O princípio de máxima verossimilhança determina que seja usado como estimativa de β o valor que maximiza a expressão na equação anterior. Porém, é matematicamente mais fácil trabalhar com o logaritmo desta equação, sendo queestáa transformação preserva o ponto de máximo nos parâmetros a serem estimados. está expressão, o log da verossimilhança, é definida como

$$\ln(\beta) = \ln(l(\beta)) = \sum_{i=1}^{n} (y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]) \tag{3.5}$$

Na regressão linear as equações de verossimilhança, obtidas por derivadas da função da soma de quadrados dos desvios com respeito a β, são lineares em parâmetros desconhecidos e assim são

facilmente solucionadas. Para as expressões de regressão logística essas derivadas são não-lineares em β e assim requerer métodos especiais para suas soluções. Estes métodos são de natureza iterativa e têm sido programados em software disponíveis de regressão logística.

A regressão logística traz a vantagem de que os parâmetros estimados tem interpretação clara, dependendo de como foi feita a amostragem dos dados. Os parâmetros da regressão logística podem ser interpretados como sendo uma razão de risco do evento em análise acontecer.

Por exemplo: Vamos supor que o evento em estudo seja o cliente pagar o empréstimo (Resposta "0") e o cliente não pagar o empréstimo (Resposta "1"). As variáveis explicativas são os indicadores de o cliente ser solteiro e se o cliente é Mulher.

Os parâmetros estimados são 0.1 para o indicador de mulher e -0.3 para o indicador de solteiro. Assim o fato do cliente ser mulher melhora em 10% a chance dele não sinistrar, já o fato do cliente ser solteiro aumenta em 30% a chance do cliente não pagar.

A regressão logística se encaixa perfeitamente no contexto de *credit scoring*, pois o evento em estudo é dicotônico (Bom/Mau) e a resposta é um probabilidade do evento ocorrer. Assim é possível atribuir a cada cliente a probabilidade dele pagar o empréstimo e com está probabilidade fazer cálculos financeiros que suportam a decisão de aprovar ou não o crédito.

3.1.1 Um Exemplo de Máxima Verossimilhança aplicado a Crédito

O método de estimação de parâmetros utilizados pela regressão logística pode ser apresentado por um estudo de probabilidade simples [6] aplicado a um modelo de regressão, ao invés de somente a um modelo de probabilidade. O estudo proposto é mostrar a estimativa dos parâmetros de um conjunto de clientes que com a características X=sexo, que tem as possibilidades Masculino(M) e Feminino (F) e um conjunto Y de resposta se o cliente pagou(0) ou não(1) o empréstimo.

Vamos considerar uma amostra de 100 pessoas do sexo masculino, sendo que 10 delas não pagaram o empréstimo e outra amostra também de 100 clientes do sexo feminino sendo que 20 delas não pagaram o empréstimo.

OBS	Sexo	Resposta
1	M	0
2	M	0
3	M	0
4	M	1
5	M	0
6	M	0
⋮	⋮	⋮
97	M	0
98	M	0
99	M	1
100	M	1

Tab. 3.1: Dados Exemplo Máxima Verossimilhança Masculino

OBS	Sexo	Resposta
1	F	0
2	F	1
3	F	1
4	F	1
5	F	0
6	F	0
⋮	⋮	⋮
97	F	0
98	F	0
99	F	1
100	F	1

Tab. 3.2: Dados Exemplo Máxima Verossimilhança Feminino

Inicialmente na análise dos dados pode se observar que a taxa de maus pagadores dos homens é de 10% e das mulheres é de 20%. Assim pode se concluir prematuramente que a taxa de maus clientes das mulheres é 2 vezes maior que a dos homens. Porém é preciso observar que estes dados não contemplam toda a população e sim um amostra da mesma, assim é necessário se calcular qual a probabilidade de cada um dos clientes cair na amostra e assim a relação estimada desta relação.

O método de máxima verossimilhança aplicado a um modelo de regressão logística irá responder a seguinte pergunta: qual é o valor da constante e da relação entre a taxa de maus clientes dos homens e mulheres que maximiza a chance desta amostra acontecer?

Inicialmente, para calcular o modelo é necessário escolher qual a característica de X que será modelada como referência, pois como as informações de Masculino e Feminino são complementares se colocarmos as duas informações elas seriam colineares com o intercepto do modelo. Arbitrariamente será escolhida a característica Masculino como nossa referência neste estudo. A característica recebe este nome de referência pois após a estimativa do parâmetro as explicações de chance são dadas com relação a ela. Para efeitos de cálculo Masculino será o valor 0 e Feminino o valor 1. Assim a matriz de dados X(Ou Matriz de Planejamento é:

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \text{ e, } Y = \begin{pmatrix} 0 \\ 1 \\ 1 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

Observa se que na matriz X temos duas colunas uma com os dados e uma com a coluna preenchido com uns. Esta coluna serve para criar a constante do modelo que estará ligada a β_0

Pela análise das equações 3.1 e 3.3 pode se notar que elas dependes exclusivamente dos valores de x_i e y_i. Neste exemplo x_i pode assumir apenas 2 valores (0 ou 1) e y_i também, assim temos apenas quatro equações para 3.3.

Masculino bom cliente:

$$a = 1 - \frac{e^{\beta_0}}{1 + e^{\beta_0}} \tag{3.6}$$

Masculino mau cliente:

$$b = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \tag{3.7}$$

Feminino bom cliente:

$$c = 1 - \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}} \tag{3.8}$$

Feminino mau cliente:

$$d = \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}} \tag{3.9}$$

Combinando estas equações na forma de 3.4 e levando as quantidades em consideração temos que a equação de máxima verossimilhança e pegando o logaritmo temos:

$$L = (a^{90}) * (b^{10}) * (c^{80}) * (d^{20}) = \frac{e^{30\,\beta_0+20\,\beta_1}}{(1 + e^{\beta_0})^{100} (1 + e^{\beta_0+\beta_1})^{100}} \tag{3.10}$$

$$Ln(L) = \ln\left(\left(1 - \frac{e^{\beta_0}}{1 + e^{\beta_0}}\right)^{90} \left(e^{\beta_0}\right)^{10} \left(1 - \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}}\right)^{80} \left(e^{\beta_0+\beta_1}\right)^{20} \left(1 + e^{\beta_0}\right)^{-10} \left(1 + e^{\beta_0+\beta_1}\right)^{-20}\right) \tag{3.11}$$

A figura 3.1 mostra a função de máxima verossimilhança (No título é apresentada a forma da equação de $Ln(L)$) e a figura 3.2 mostra as curvas de níveis. Para se achar o máximo desta função é necessário utilizar métodos de busca iterativos pois as derivadas parciais de cada um dos parâmetros

não tem solução fechada. Aplicando um método podemos chegar aos valores de β_0 e β_1:

$$\beta_0 = -2.197$$

$$\beta_1 = 0.8109$$

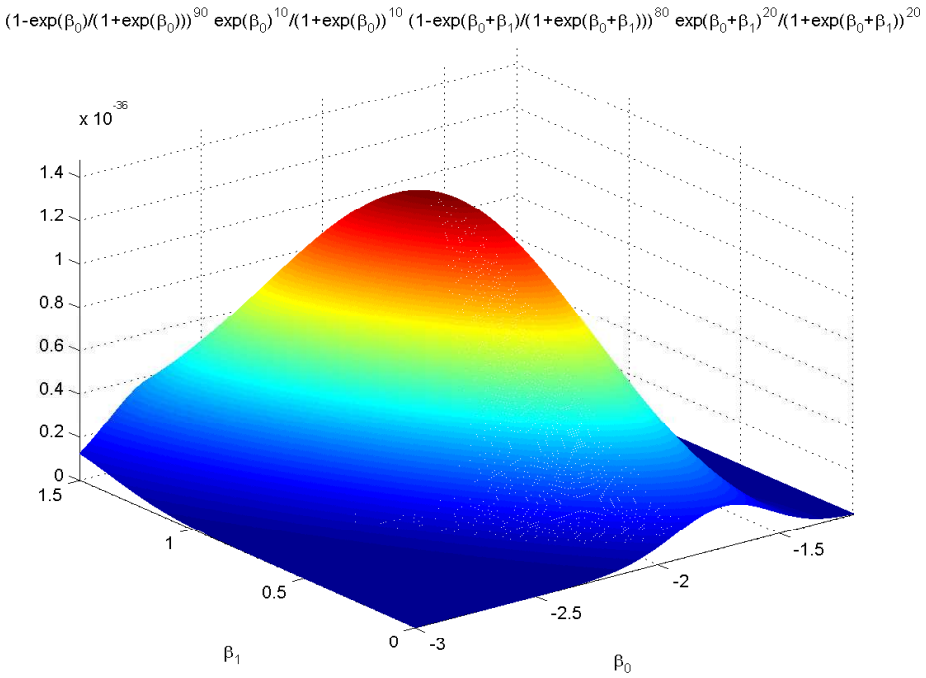

Fig. 3.1: Curva Máxima Verossimilhança Exemplo Logística.

A interpretação do valor de β_1 indica que neste exemplo fictício os homens são 81.09% melhores pagadores que as mulheres, ou que as mulheres são $1/0.8109 - 1 = 23.32\%$ mais ariscados que os homens e não 50% mais arriscado como a análise inicial das taxas de maus clientes podem sugerir. Assim quando estamos falando de amostras, e 99% das vezes é uma mostra, tem que se levar em consideração a probabilidade de cada clientes ser selecionado na amostra.

Está interpretação também nos ajuda a explicar a pessoas não técnicas como o modelo está calculando a probabilidade de o cliente não pagar. Neste exemplo a mulher tem uma maior chance de não pagar, no mundo real não é o que acontece, pois observa se que as mulheres são melhores pagadoras

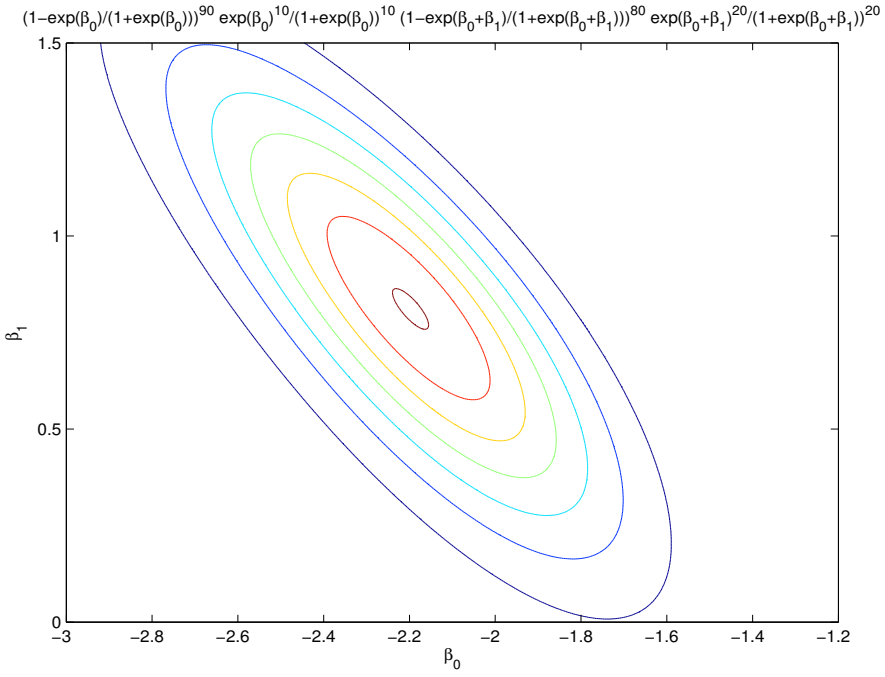

Fig. 3.2: Curva de Níveis para a função de Máxima Verossimilhança

que os homens. Assim num caso deste seria necessário melhores explicações para que o modelo fosse aceito, como por exemplo se tratar de uma população específica com características diferenciadas.

3.2 Redes Neurais Artificiais

Com o amplo sucesso do uso da regressão logística em aumentar o lucro dos bancos com as carteiras de crédito, as instituições financeiras e empresas que vendem soluções começaram a buscar em outros campos técnicas capases de conseguir melhores desempenhos ou suprir necessidades que a regressão logística falhe, como em reconhecimento de eventos raros por exemplo.

Antes de passarmos diretamente a utilização dela em modelos de crédito, vamos contar um pouco de seu surgimento e teorias expostas.

A idéia inicial das redes neurais artificiais era reproduzir o funcionamento do cérebro humano num esforço inicial de se entender o funcionamento do mesmo. O objetivo era produzir mecanismos

artificiais capazes de funcionar da mesma maneira, aprendendo, tomando decisões, reconhecessem padrões armazenados anteriormente, produzindo correlações nunca vistas antes, etc.

A redes artificiais são baseadas em como o cérebro humano é organizado pois ele espectacularmente superior a computação digital. Atualmente os chips são capazes de calcular operações simples em nano segundos enquanto o cérebro demora milisegundos, porém o cérebro é capaz de reconhecer padrões antigos por um ângulo nunca visto antes. A principal diferença não está na velocidade mas em como o processamento da informação é organizado. No cérebro o processamento é massivamente paralelo e um neurônio está conectado a muitos outros através das conexões sinápticas. Outra característica da atividade neurológica é a capacidade de adaptar e de se auto organizar. A medida que se vai se adquirindo novas experiências o cérebro tem que se adaptar para poder assimilar novas perspectivas. Também é conhecido que dependendo da atividade apenas uma região do cérebro é ativada.

O primeiro trabalho sobre redes neurais artificiais foi publicado por MacCulloch and Pitts [7], onde eles consideraram que o neurônio irá executar uma função lógica binária. Eles imaginaram o neurônio funcionando com a lógica das proposições e todo seu trabalho foi suportado por elas. Os conhecimento atuais sobre o cérebro permitem afirmar que os cérebro não trabalha desta forma e este neurônio focou conhecido como um caso particular de neurônio.

Exitem algumas definições para uma Rede Neural Artificial, entre as mais aceita está a que uma rede neural artificial (RNA) é um sistema massivamente paralelo e distribuído, composto por unidades de processamento simples que possuem uma capacidade natural de armazenar e utilizar conhecimento. As RNAs tem várias características em comum com o cérebro humano:

- As informações são processadas em unidades simples, neurônios.
- Os neurônios são ligados uns aos outros criando uma rede.
- As informações quase sempre são armazenadas nos pesos.
- Existem um processo de aprendizagem que guardam as informações nos pesos.

A estrutura básica das RNAs é o neurônio apresentado na figura 3.3:

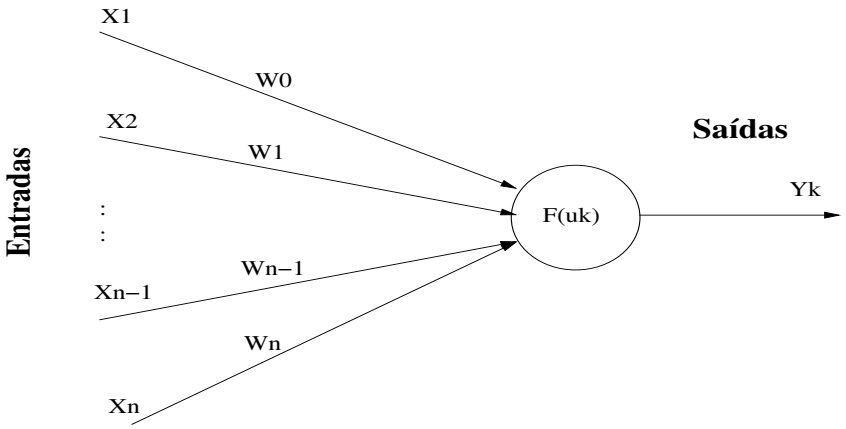

Fig. 3.3: Estrutura do Neurônio

As principais características de um neurônio são:

- A junção somadora;
- A função de ativação $f(u_k)$;
- Os valores de entrada x_i;
- Os pesos w_i;
- A sinapses que ligam a entrada ao neurônio.

A junção somadora soma todas as entradas multiplicadas pelos pesos, criando uma combinação linear dos pesos. Após a união, a função de ativação é aplicada com a finalidade de limitar a saída e/ou introduzir não linearidade ao modelo. As sinapses entregam a informação aos neurônios seguintes tornando se assim, entradas em novos neurônios.

De maneira geral a relação entre a entrada e saída tem a forma:

$$y = f(\phi(x, w)) \tag{3.12}$$

onde f e ϕ são funções previamente definidas, x representa o vetor de entrada e *w* são os pesos das conexões sinápticas. A função f é chamada de *função de ativação*.

Nesta notação f e ϕ parecem ser redundantes porem em algumas aplicações elas têm características importantes. Usualmente ϕ tem a forma linear e f é escolhida a partir de um pequeno conjunto de funções, por exemplo:

- $f(u) = sgn(u)$ Gera saída binária -1,+1.
- $f(u) = (sgn(u) + 1)/2$ Gera saída binária O ou 1.
- $f(u) = (1 - e^{(-u)})^{-1}$ Normal Padrão
- $f(u) = tanh(u)$ Função Sigmoidal

Os neurônios podem ser inter conectados de várias maneiras gerando as diferentes arquiteturas possíveis de redes:

- As redes feedforward de uma única camada;
- As redes feedforward de múltiplas camadas;
- As redes recorrentes.

A figura 3.4 exemplifica uma arquitetura de rede feedforward de uma única camada. A rede é denominada feedforward porque a informação trafega em apenas um sentido, da entrada para a saída.

A figura 3.5 exemplifica uma arquitetura de rede multi camadas. As redes multi camadas podem ter uma ou mais camadas, teoricamente aumentando o poder de processamento da rede de uma única camada.

O projeto de construção de um modelo através de uma RNA envolve três fases:

- a escolha de um conjunto de neurônios artificiais,
- a definição de um padrão de conectividade,

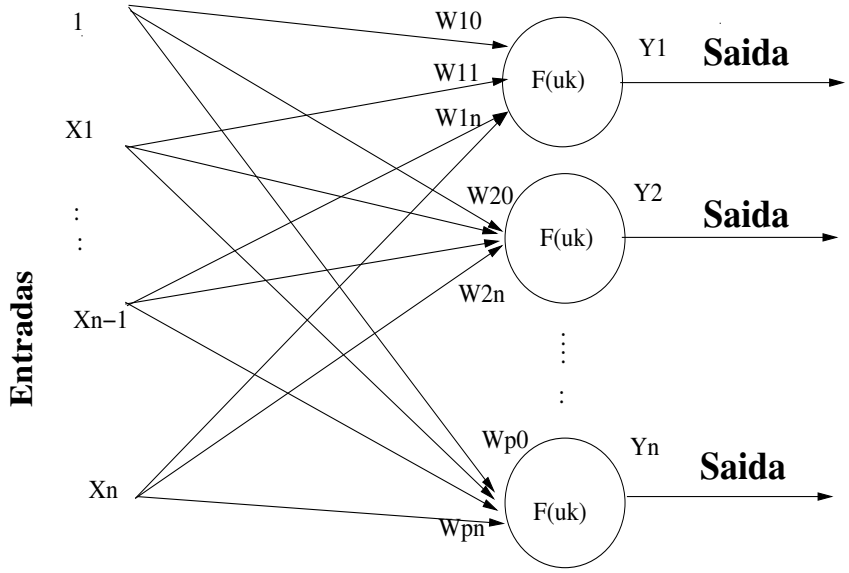

Fig. 3.4: Arquitetura de Redes Neurais feedforward

- a definição de um algoritmo de treinamento.

A escolha da quantidade de neurônios depende da necessidade de generalização da rede, quanto mais neurônios maior o poder de generalização, porém neurônios demais podem deixar a rede instável e difícil de treinar. O padrão de conectividade define como a informação irá trafegar através da rede e como a conhecimento deve ser armazenado.

Existem vários algoritmos para a estimação dos parâmetros das redes neurais. O algoritmo depende da rede e do poder computacional disponível e do tipo de rede a ser treinada.

A arquitetura e o método de estimação utilizado define também a a utilização da rede. No caso de estimar o risco de crédito de um cliente a arquitetura que melhor se adapta é a do tipo MLP, que é uma das redes mais conhecidas que será tratada na próxima seção, onde serão discutidos os métodos de estimativa e utilização.

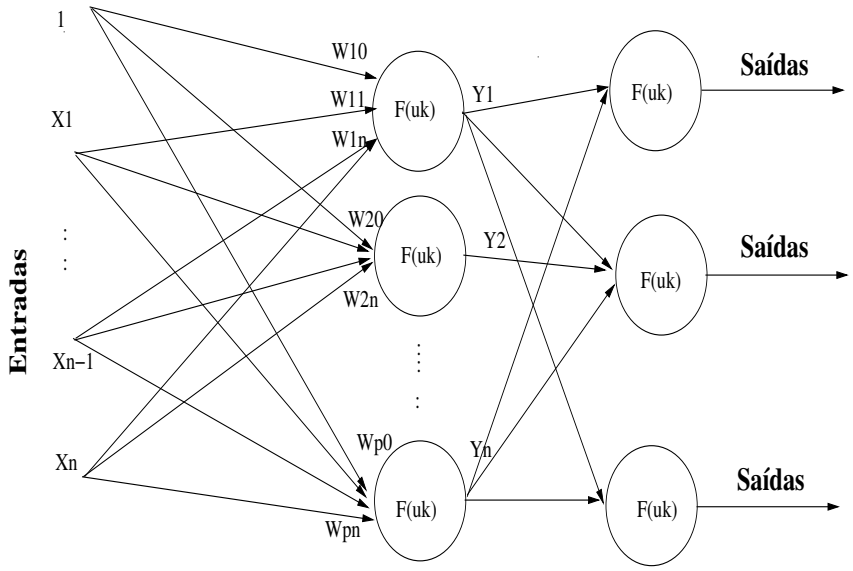

Fig. 3.5: Arquitetura de Redes Neurais

3.2.1 Redes do tipo Perceptron de Múltiplas Camadas (MLP) aplicadas a modelos

Para se construir um modelos de previsão de risco de crédito é necessário mapear as informações de entrada com a resposta de se o cliente pagou ou não a dívida. Uma das arquiteturas mais conhecidas de redes que faz este tipo de mapeamento é do tipo MLP que faz este mapeamento não linear entre a entrada e saída.

Perceptron de Múltiplas Camadas é uma rede do tipo perceptron com pelo menos uma camada intermediária, originalmente proposta no livro Perceptrons [8]. Este tipo de rede pode ser treinada para que os pesos da conexões permitam construir um mapeamento entre uma matriz de entrada X e uma ou várias saídas Y_i. Assim é possível utilizar as entradas da matiz X para se prever as saídas Y, criando um mapeamento.

Uma rede MLP comum possui as características:

- Os neurônios tem geralmente funções de ativação não linear.
- A rede possui camadas intermediárias.

- Todas as entradas são ligadas a todos os neurônios e toda saida de uma camada intermediário é ligada em um outro neurônio.

Cybenco [9] demonstrou que uma rede MLP com apenas uma camada intermediária é suficiente para aproximar qualquer função contínua que se encaixe em um hipercubo unitário. Neste contexto a MPL pode ser usada para regressão Linear/Não Linear, construção de funções discriminantes e previsão de séries temporais. É importante notar que os resultados mostrados por ele garantem a existência, não determinando que a rede MLP sejá otima no sentido de tempo nessário para se encontrar a solução.

O algoritmo utilizado para se encontrar os pesos dos neurônios(ou treinar a rede) nas redes do tipo MPL é o *backpropagation*. Este algoritmo consiste em uma fase de propagação positiva do sinal de entrada e uma fase de retropropagação do erro encontrado. Na fase de propagação positiva um sinal é aplicado a entrada da rede e é observado o valor de saída, na segunda fase este valor obtido é comparado com o valor desejado e então é gerado o erro. Este erro é passado as camadas anteriores da rede através do gradiente do erro com relação aos vetores dos pesos. Assim ajustando os pesos de forma que eles caminhem no sentido oposto do gradiente iremos minimizar o erro da saída. Para propagar o erro entre as camadas da rede é utilizada a regra delta generalizada.

A regra delta, que é utilizada para determinar os pesos, encontra o gradiente do vetor de erro em relação aos pesos de entrada dos neurônios. Este vetor gradiente será utilizado para diminuir o erro entre a saída da rede e os dados verdadeiros Y. Como já é conhecido, atualizando os pesos da rede no sentido contrário ao crescimento do gradiente estaremos minimizando o erro da rede. A regra delta para a primeira camada pode ser definida como:

$$w_{i,j,h}(t+1) = w_{i,j,h}(t) - \alpha \frac{\delta E(t)}{\delta w_{i,j,h}(t)} \tag{3.13}$$

onde $w_{i,j,h}(t+1)$ são os pesos do neurônio i da entrada j na camada h da interação t+1, $w_i(t)$ são os pesos do neurônio i da entrada j na camada h da interação t,α é o tamanho do passo e $E(t)$ é o vetor de erros no tempo t.

A regra delta inicialmente calcula o gradiente do erro entre a saída da primeira camada de neurônios, então para passar o erro às outra camadas consideramos a saída da camada anterior a entrada da próxima, propagando o erro entre as camadas. Então a saída Y pode ser escrita como:

$$Y_{h+1} = f_{h+1}(w_{h+1}Y_h) \tag{3.14}$$

para todas as camadas h da rede.

No algoritmo do *backpropagation* é necessário cálculo de todas as derivadas parciais das camadas intermediárias, para isso é necessário o uso da regra da cadeia. Assim temos:

$$\frac{\delta E}{\delta w_{i,j,h}} = \frac{\delta E}{\delta u_{i,h}} X \frac{\delta u_{i,h}}{\delta w_{i,j,h}} \tag{3.15}$$

O termo $u_{i,h} = \sum_{j=1}^{S} w_{i,j,h} y_{j,h-1}$, então:

$$\frac{\delta u_{i,h}}{\delta w_{i,j,h}} = y_{j,h-1} \tag{3.16}$$

Então a equação de atualização dos pesos em qualquer camada pode ser descrita como:

$$w_{i,j,h}(t+1) = w_{i,j,h}(t) - \alpha \frac{\delta E}{\delta u_{i,h}} y_{j,h-1} \tag{3.17}$$

O termo $\frac{\delta E}{\delta u_{i,h}}$ é conhecido como sensibilidade da rede e o seu cálculo é necessário para formula final. O seu cálculo depende da camada anterior para ser calculado, assim então é necessário o uso de outra regra da cadeia. Está estrutura de cálculo do erro é que dá origem ao nome retro-propagação, pois a camada anterior depende da próxima e assim por diante.

Então cada elemento a ser calculado tem a forma:

$$\frac{\delta u_{i,h+1}}{\delta u_{i,h}} = w_{i,j,h+1} \frac{\delta y_{j,h}}{\delta u_{i,h}} \tag{3.18}$$

assim pode se observar que a sensibilidade está indo da camada mais próxima da saída para a primeira

camada mais próxima das entradas

$$E_h \rightarrow E_h - 1 \rightarrow E_h - 2 \ldots \rightarrow E_2 \rightarrow E_1$$

Assim é possível definir um algoritmo padrão para cálculo dos pesos da rede.

Procedimento de calculo de W da MPL;
  Inicializar todos os pesos w_i, j, h com valores aleatórios pequenos;
  $t \leftarrow 1$;
  $max_t \leftarrow$ *O Número Máximo de Interações ;*
  $minerro \leftarrow$ *O Erro mínimo aceitável entre a saída da rede e os valores Y;*
  $\alpha \leftarrow$ *O tamanho do passo a ser dado a cada interação ;*
  Enquanto $t < max_t$ *e* $MSE < minerro$,
    Faça de i = 1 até N ***Para cada amostra no conjunto de treinamento***,
      $y_0 \leftarrow x_i$;
      $y_{i,h+1} \leftarrow f_{h+1}(W_{h+1}y_{i,h})$;
      $\delta_{i,h} = -2F_h(u_{i,h})(d_i - y_i)$ ***Para Última Camada de Neurônios***;
      $\delta_{i,h} = -2F_h(u_{i,h})W_{h+1}\delta_{i,h+1}$ ***Para as outras Camadas de Neurônios***;
      Atualiza os pesos;
      $W_h \leftarrow W_{h-1} - \alpha delta_{i,h}(y_{i,h-1})$;
      Calcula o E : $E_i \leftarrow (d_i - y_i)$;
    Fim
    Calcula o Erro :$MSE \leftarrow \sum \frac{E_i}{N}$;
    $t \leftarrow t + 1$;
  Fim
Fim

Na utilização prática do treinamento de uma rede MLP precisa se de um conjunto de dados com sinais de entrada e os respectivos sinais de saídas desejados, denominado conjunto de treinamento. Durante o processo de aprendizagem o conjunto de treinamento é aplicado repetidas vezes até que o critério de parada seja satisfeito. A apresentação das amostras a cada ciclo de treinamento deve ser feita de forma aleatória para dar uma característica estocástica na busca do melhor conjunto de pesos. Existem vários algoritmos propostos para a utilização prática do método de *backpropagation*, entre eles Método de Newtom Modificado, Método de Levenberg-Marquardt e o Método do Gradiente Escalonado Conjugado.

Após a aplicação do conjunto de treinamento e encontrados os pesos é utilizado um outro conjunto de dados denominado de validação para verificar o ajuste da rede em indivíduos que não foram utilizados no treinamento Desta forma é possível avaliar a capacidade de generalização da rede verificando se não houve um super ajuste aos dados o que pode fazer a rede muito boa para o conjunto de treinamento e muito ruim para outro indivíduo.

O processo de busca dos pesos de uma rede não tem garantia de convergência para o mínimo global assim como o critério de parada não é claro. existem critérios que olham para o vetor de erro ou para o vetor gradiente, assim como para o tempo decorrido no treinamento. A escolha do critério depende do problema em estudo e da técnica de construção do algoritmo de *backpropagation*.

3.2.2 Exemplo

Os modelos de crédito tem se utilizado de outras técnicas para a construção de modelos para contorna obstáculos que as técnicas tradicionais não tem tido sucesso, tais como:

- Baixa taxa de clientes considerados maus,
- Pequenas amostras, no caso de nichos de mercados muito específicos.

Neste exemplo será construído um modelo utilizando os mesmos dados apresentados no exemplo de regressão logística para exemplificar as metodologia de construção e não os resultados no sentido de qual é o melhor modelo.

A idéia é a mesma, relacionar o sexo dos clientes com o fato dele pagar ou não o crédito concedido, porém em redes neurais não estamos preocupados em identificarmos relações ou explicar comportamentos, queremos apenas encontrar qual o melhor mapeamento entre a entrada (Sexo) e a Saída (pagou ou não).

Neste sentido utilizaremos uma rede do tipo MLP para fazer este mapeamento com uma camada escondida de 3 neurônios. A escolha da quantidade e neurônios depende do nível de ruído da base. Como neste exemplo é fictício o número 3 foi escolhido por conveniência. As funções de ativação

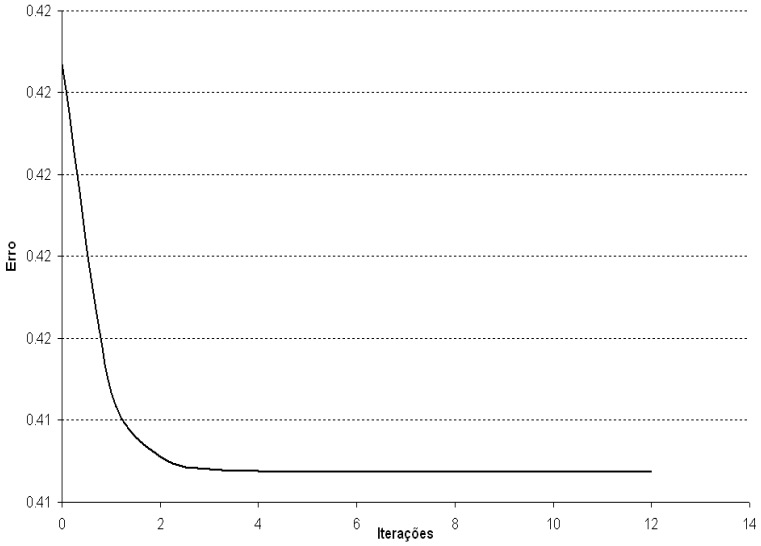

Fig. 3.6: Curva de Treinamento Redes Neurais

dos neurônios da camada escondida é a tangente hiperbólica e da camada de saída é uma combinação linear.

Para este estudo foi necessária a utilização de um software para se encontrar os pesos, assim foi escolhido o software SAS com *Enteprise Miner*. Este é um software muito conhecido para análise de grandes bancos de dados e construção de modelos de *Scoring*.

Para este conjunto de dados o gráfico de erro de treinamento é dado pela figura 3.6 na qual pode se ver que em 4 interações o software foi capaz de encontrar um solução.

A tabela 3.3 mostra e estimativa dos pesos em cada neurônio. Nesta tabela a primeira coluna mostra de onde está vindo a informação e a segunda aonde ela está entrando, se no neurônio 1 (H11) ou no 2 (H12) e assim por diante.

Neste exemplo, os parâmetros de quantidade de neurônios e funções de ativação foram escolhido de maneira didática, porém na prática é necessária se testar várias configurações destes parâmetros e escolher o melhor modelo. O critério de melhor modelo pode ser variável, nas seção 3.5 serão discutidos dois critérios paraestáa seleção.

Origem Sinal	Destino Sinal	Parâmetro
X	H11	-0.037
X	H12	-0.068
X	H13	-0.244
BIAS	H11	-0.230
BIAS	H12	0.058
BIAS	H13	0.497
H11	Y1	-0.799
H12	Y1	-10.641
H13	Y1	-0.255
BIAS	Y1	-1.641

Tab. 3.3: Estimativa dos Pesos dos Neurônios na Rede

Outro passo importante de se utilizar na construção de redes neurais é uma amostra de validação, pois ela evita super ajustes, que é quando o modelo se adapta aos dados da amostra de construção porém para qualquer outra amostra ele não funciona.

Em comparação a regressão logística a método de redes neurais oferece a vantagem de não considerar pré-supostos para achar os parâmetros o que possibilita a utilização em amostras pequenas ou ainda em mapeamentos nos quais não podemos considerar condições para a distribuição da variável resposta.

3.3 Métodos Naturais

Na busca por novas metodologias que possam ser utilizadas na construção de modelos de risco de crédito uma das mais usuais é a de Computação Natural. Computação natural é uma área que observa as teorias de evolução dos organismos vivos e tenta criar conceitos similares aplicados ao aprendizado de máquina. Neste sentido apareceram as técnicas de Algoritmos Genéticos e as Estratégias Evolutivas(EE).

No contexto de previsão de risco a computação natural pode ser aplicada como um método para se otimizar outras funções que não a taxa de risco propriamente dita. Assim podemos abordar perguntas como: o resultado de uma instituição financeira é medido em função do seu lucro ou tamanho de

ativo, porque não criar modelos para medir diretamente o resultado de um empréstimo a um dado cliente ou grupo?

A grande dificuldade dos modelos que tentam prever diretamente a rentabilidade (ou resultado) é que a variável resposta não tem uma distribuição conhecida, condição necessária para se utilizar modelos paramétricos, ou ainda o objetivo não é simplesmente criar um modelo que retorne a melhor previsão de resultado para cada cliente, e sim um modelo que retorne o melhor resultado para p grupo de clientes.

Em fim, estes modelos são necessários quando o objetivo a ser otimizado não se encaixa no conceito de minimizar o erro da estimativa utilizado tanto nos modelos parâmetricos quanto em Redes Neurais. A Seguir apresentaremos algumas destas técnicas.

3.3.1 Algoritmos Genéticos

Algoritmos Genéticos (AGs) são algoritmos estocásticos de busca baseados em idéias evolutivas de genética e seleção natural. Eles combinam processos naturais necessários à evolução, especialmente aqueles estabelecidos por Charles Darwin de sobrevivência do mais apto com troca de informação estruturada, porém randômica, para formar um algoritmo de busca com a habilidade inovadora da busca humana.

Apesar de já estarem sendo desenvolvidos desde 1962, a primeira conquista significativa em Algoritmos Genéticos foi feita em 1975 com a publicação de *Adaptation in Natural and Artificial System* por John Holland [10], seus colegas e seus alunos na Universidade de Michigan. Desde então veêm sendo largamente estudados, experimentados e aplicados em vários campos da engenharia, da biologia e das ciências naturais.

Os AGs recebem como entrada uma população de indivíduos em representação genotípica (geração inicial) e uma função(*fitness*) que avalia a adequação relativa de cada indivíduo. Os indivíduos são codificados como listas ordenadas (cromossomos) onde cada atributo equivale a um gene e seu valor a um alelo. O tamanho da lista está relacionado ao número de atributos necessários para descre-

ver o indivíduo. À geração inicial, os AGs empregam operadores de *crossover* e mutação para gerar novos indivíduos. Ele usa vários critérios de seleção de modo a escolher os melhores indivíduos para a reprodução. O quão bom é cada indivíduo é determinado pela função objetivo.

O emprego mais comum dado aos AGs é em problemas de otimização, onde o problema a ser resolvido faz o papel do ambiente e cada indivíduo da população é associado a uma solução candidata. Assim, um indivíduo estará mais apto ao ambiente quando ele corresponder a uma solução mais eficaz para o problema. Com a evolução, espera-se que a cada geração obtenha-se soluções candidatas mais e mais eficazes, contudo sem a garantia de se chegar à solução ótima no final do processo evolutivo. É importante enfatizar que algoritmo genético é uma alternativa de abordagem de problemas classificada como método fraco, concebido para resolver problemas genéricos em mundos não-lineares e não-estacionários e não garantem eficiência total na obtenção da solução. Geralmente garantem uma boa aproximação para a solução. Desse modo, devem ser considerados apenas quando métodos fortes, que operam em mundos lineares, contínuos, diferenciáveis e/ou estacionários não se aplicam ou falham.

O primeiro passo na implementação de um algoritmo genético é a geração de uma população inicial. Para os algoritmos genéticos canônicos cada membro desta população é representado por uma lista binária de comprimento l

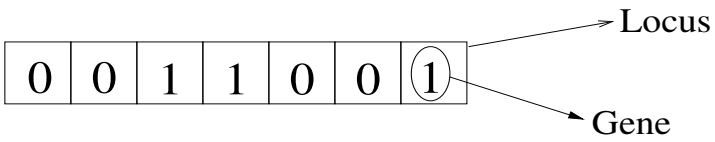

Fig. 3.7: Codificação Algoritmo Genetico

Cada seqüência é freqüentemente referida como um genótipo ou, alternativamente, como um cromossomo onde os *bits* são os alelos. A execução de um algoritmo genético pode ser dividida em dois estágios. Inicia-se com a população corrente, em seguida seleciona-se indivíduos para formar uma população intermediária em uma etapa chamada seleção. Logo após, a re-combinação (*crossover*) e a mutação são aplicadas à população intermediária gerando novos indivíduos que serão inseridos na

população de acordo com seus respectivos graus de adaptação. É gerada assim uma nova população. O processo que se estende desde o início da geração da população inicial até a criação de uma nova população constitui uma geração na execução do algoritmo genético.

Então os principais processos da execução de um algoritmo genético é a o processo de seleção e o processo de geração da nova população. O processo de seleção é feito através da avaliação de cada indivíduo pelo *Fitness*, assim os indivíduos mais adaptados são escolhidos.

O segundo processo é a combinação dos indivíduos da população para a geração de novos indivíduos. A combinação pode ser feita de por operadores de mutação ou reprodução. O primeiro altera o valor do gene de 0 para 1 ou vice versa de acordo com a probabilidade de mutação. O segundo seleciona dois indivíduos da população original e seleciona um ponto de separação na qual os genes anteriores do indivíduo 1 são alocados ao novo primeiro indivíduo e o restante dos genes são coletados do indivíduos 2 original, gerando um indivíduo que é a combinação dos originais.

O algoritmo genérico tem o problema de tratar com codificação binária dos números e também a posição dos genes influencia na probabilidade de encontrar o melhor resultado.

3.3.2 Estratégias Evolutivas

Outra classe de algoritmos inspirados nas teorias da evolução são as Estratégias Evolutivas (EE). Elas formam um conjunto de algoritmos desenvolvidos para a otimização de parâmetros em funções reais, mais apropriados a modelos de risco de crédito. Eles foram criados por RECHENBERG (1965) [11], SCHWEFEL (1965) [12] e Bienert na Universidade Técnica de Berlim por volta de 1964.

As estratégias evolutivas utilizam além da idéia de evoluir os valores que otimizam a função de avaliação a de evoluir também os parâmetros de controle da estratégia. Assim a estratégia pode se adaptar de acordo com a região do espaço de busca a qual ela se encontra, por exemplo, numa região com uma derivada grande o algoritmo pode dar passos longos e evoluir mais rápido, já quando está chegando perto da solução pode evoluir com passos menores para explorar melhor a região da solução.

Nas EE um indivíduo é composto pela estrutura de dados x e os conjuntos θ e σ que são os conjuntos de parâmetros de estratégia. Após a inicialização dos parâmetros, os vetores θ e σ são atualizados e após a sua atualização o vetor x é atualizado e aplicada a seleção para escolher os mais adaptados de acordo com a função de avaliação.

No contexto de modelos de risco de crédito a grande vantagem das Estratégias Evolutivas é a de poder utilizar funções de avaliação não usuais aos modelos de otimização do erro entre a saída e a entrada. Por exemplo podemos utilizar:

- A estatística de Kolmogorov-Smirnov(KS),
- A rentabilidade de cada um dos clientes
- A rentabilidade de um grupo de clientes,
- Em quanto tempo o cliente deixará de pagar o empréstimo.

As Estratégias Evolutivas utilizam mecanismos para que a partir de uma população inicial sejam criados filhos e de acordo com a avaliação da função escolhida os indivíduos são selecionados. Assim é necessário definir quais os métodos a serem utilizados para gerar os filhos e qual o método de seleção será utilizado a partir da função de avaliação.

Considerando μ como sendo o número de indivíduos na população e λ como sendo o número de filhos gerados a cada geração existem alguns tipos de estratégias:

- (1+1) -EE : Nesta estratégia a população tem apenas um indivíduo e apenas o operador de mutação é aplicado. O indivíduo é representado pelo vetor (x, σ). Neste caso é utilizada a distribuição gaussiana com variância σ para gerar uma variação no valor original, este novo indivíduo é aceito se e somente se ele tiver um valor de avaliação melhor que o seu pai. Está é denominada EE de dois membros pois o filho compete com o pai. Nesta estratégia o desvio padrão permanece inalterado durante o processo, ou é utilizado a regra proposta por SCHWEFEL (1981). É possível provar que se o problemas for regular ele tem convergência garantida,

porém nada é dito sobre o tempo ou a taxa de convergência. Assim o processo pode demorar muito para atingir a melhor solução.

- (μ+1) - EE : está é a primeira estratégia multimembros, onde existem μ, e cada pai gera um filho através do processo de mutação. Com a introdução de μ pais é possível a reprodução. Então são selecionados dois pais e a combinação entre eles é feita de forma linear. Após aplicados os operadores de mutação e combinação os μ melhores, de acordo com a função de avaliação, são selecionados.

- (μ+λ) - EE e (μ,λ) - EE : Essas são as estratégias multimembros. Nestas estratégias o desvio padrão também é incorporado como sendo parte do código genético, também sendo possível aplicar mutação e re-combinação. Na estratégia μ+λ) os μ pais geram λ filhos e todos eles são levados para seleção, permanecendo os μ indivíduos com melhor valor de avaliação. Isso pode ser um problema quando se tem uma superfície dinâmica ou com muito ruído. Para evitar isso foi proposto a estratégia (μ,λ) onde somente os λ filhos são aptos a permanecer para a próxima geração, criando assim um fator de esquecimento.

Para realizar estas estratégias existem vários operadores de combinação tanto para os parâmetros de busca como para os parâmetros de controle como o desvio padrão. Essas estratégias de atuação variam desde a forma de atualização até ao cálculo da distribuição utilizada. Outro ponto importante é o tamanho da população, pois ela determina a quantidade de variabilidade que teremos na população. Existem propostas de estratégias que controlam o tamanho da população de acordo com a variabilidade e a velocidade de evolução na direção da solução.

Um fator importante das estratégias evolutivas é o controle da população, pois é necessário se manter uma diversidade mínima entre os indivíduos para que eles não encontrem um mínimo ou máximo local e não sejam capazes de sair desta região do espaço. É sempre recomendável utilizar a introdução de geração de indivíduos aleatórios para aumentar a chance de encontrar o máximo/mínimo da função de avaliação.

Nas técnicas de computação natural é sempre importante lembrar que não existe garantia de encontrar o máximo da função de avaliação, pois ele sempre utiliza uma busca estocástica com tempo finito de busca o que não tem probabilidade 1 de encontrar o resultado.

Como os modelos de crédito são funções reais que visam avaliar o risco do cliente não pagar podemos utilizar as EE para encontrar os melhores parâmetros ou transformações que aplicados as informações de entrada podemos melhor avaliar o risco do cliente. É importante que não fiquemos apenas na busca de parâmetros lineares, pois os métodos atuais são mais eficientes em encontrar as melhores soluções para estes parâmetros. É importante sugerir novas funções de avaliação e novas funções para relacionar as informações de entrada e saída dos modelos.

3.3.3 Exemplo

A implementação de modelos que utilizam técnicas de Estratégia Evolutivas tem a dificuldade de ainda existirem poucos softwares comerciais com tais rotinas implementadas. Assim, para criarmos um exemplo de utilização desta técnica foi necessário a implementação destes algoritmos em Matlab.

Os dados utilizados foram os mesmos apresentados na Regressão Logística e em Redes Neurais. A estratégia utilizada foi (μ+λ) - EE, a onde foram utilizados μ pais que geram λ filhos e os mais adaptados serão escolhidos.

As técnicas de mutação utilizadas foram:

- Mutação Gausiana, na qual é gerado aleatoriamente um valor com distribuição normal com média 0 e variância 1.
- Geração aleatória de indivíduos, para manter a diversidade.

A ordem de funcionamento do programa está descrito na Figura 3.8. Neste fluxograma temos desde a inicialização da população, necessária a primeira geração de filhos até a avaliação do critério de parada, que muitas vezes é o tempo de processamento ou o número de gerações.

Outro ponto importante a ser escolhido é o critério os indivíduos da população de parâmetros serão avaliados. Neste caso, para ilustrar a utilização de outras respostas o critério para escolha

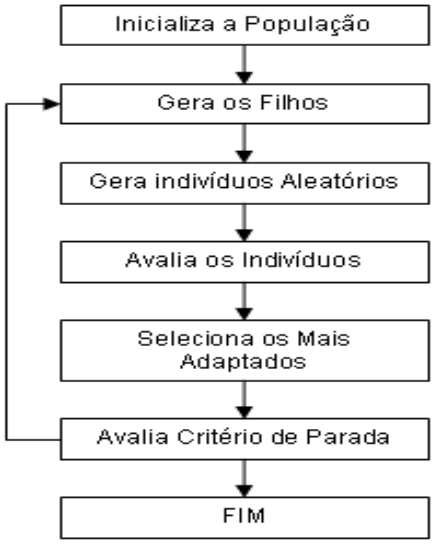

Fig. 3.8: Esquema de Funcionamento da EE

dos parâmetros no conjunto de dados será a estatística de Kolmogorov-Smirnov(KS). Assim a cada geração os indivíduos serão avaliados de acordo com o seu desempenho em gerar a estatística para o conjunto de dados.

Neste exemplo os parâmetros iniciais de busca variam entre -2 e 2, os dados utilizados tem 11 variáveis de entrada e são geradas 2000 gerações.

O Código a Seguir mostra a utilização da Técnica.

A avaliação do melhor indivíduo a cada geração é apresentado na figura 3.9. Nota se pelo gráfico que o algoritmo tem uma evolução maior no inicio e depois fica algumas gerações parado no mínimo local, porém ele ainda em no final variabilidade na população para sair deste mínimo e evoluir. Um dos grandes questionamentos aqui é quando parar, muitas vezes a escolha é pelo tempo devido ao alto custo de processamento.

```
ndimensao=11,
mu=5,
xmim=-2,
xmax=2,
totalGeracoes=2000,
   [X] = rand(mu,ndimensao).*(xmax-xmim)-(xmax-xmim)./2;
   contGeracao = 1;
   while 1
      [linhaTotalClone, colunaTotalClone] = size(X);
      Xfilho=X+(randn(linhaTotalClone, colunaTotalClone));
      [Xfilho]=[X;Xfilho;rand(mu,ndimensao).*(xmax-xmim)-(xmax-xmim)./2];
      fitnessTotal = fitness(Xfilho,mau,bom);
      [linhaTotal, colunaTotal] = size(Xfilho);
      if contGeracao < totalGeracoes
         fitnessOrdenado = sort(fitnessTotal);
         for contIndividuo=1:mu
            for contIndividuo2=1:linhaTotal
            if fitnessOrdenado(contIndividuo)==fitnessTotal(contIndividuo2)
               newX(contIndividuo,:) = Xfilho(contIndividuo2,:);
            end
         end
      end
      X = newX;
      fitMax(contGeracao,1) = fitnessOrdenado(1);
      melhorx(contGeracao,:)=X(1,:);
      else
         break
      end
      [contGeracao,1/fitMax(contGeracao)]
      contGeracao = contGeracao+1;
   end
```

3.4 Seleção de Modelos

Existem centenas de trabalhos que se preocupam com a estimativa dos parâmetros dos modelos e suas precisões associadas, porém existem poucos trabalhos que se preocupam em definir critérios de seleção dos vastos modelos propostos atualmente. Nesta seção será tratado alguns destes critérios.

Para o contesto de seleção de modelos nós consideramos que existe um conjunto de dados e que a inferência estatística é feita baseada em modelos. Classicamente é considerado que existe um único

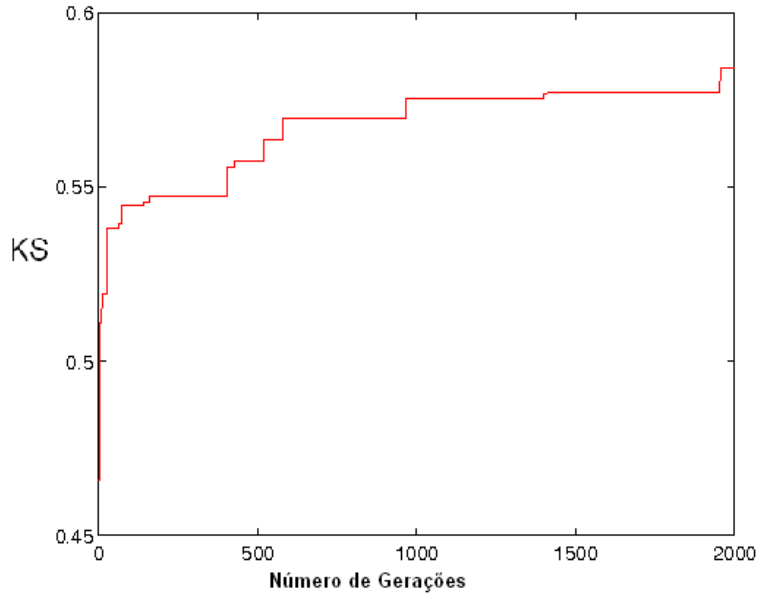

Fig. 3.9: Esquema de Funcionamento da EE do exemplo

modelo correto ou que existe o melhor modelo para ser selecionado, assim os parâmetros do modelo são estimados a partir de uma amostra e o erro do modelo é calculado. Deste modo a incerteza de seleção do modelo é ignorada, afinal o melhor modelo já foi encontrado. Este método é plausível pois usualmente se considera que o modelo verdadeiro está entre todos os modelos testados.

O critério de seleção de modelos deve seguir alguns princípios:

- ser estimado a partir dos dados para cada modelo ajustado,
- deve se resumir a um número,
- deve permitir calcular a incerteza de se estar escolhendo o melhor modelo.

Os critérios de seleção de modelo se utilizam de duas correntes teóricas da estimativa dos parâmetros: a estimativa por máxima verossimilhança e a estimativa segundo os critérios de Bayes.

Duas abordagens para construção de critérios satisfazem essas exigências: seleção baseada na teoria da perda de informação de Kullback-Leibler(k-l) e a seleção de modelos baseada nos fatores

de Bayes. O Critério de Informação de Akaike(AIC) representa a primeira abordagem e o Critério de Informação de Bayes (BIC) representa uma aproximação para a segunda abordagem, pois o cálculo exato dos fatores de Bayes pode ser muito complexo. A idéia é apresentar conceitos teóricos de seleção de modelos e fazer uma comparação entre AIC e BIC.

3.4.1 AIC

O critério de Akaike é um medida de bondade de ajuste do modelo [1] apoiada no conceito da teoria da informação. O AIC fornece uma definição matemática do princípio da parcimônia na construção de modelos, ou seja, quando o erro quadrático médio de dois modelos é o mesmo o melhor modelo é o com o menor número de parâmetros. Assim o AIC é um método prático para fazer o balanceamento entre a complexidade do modelo e o quão bem o modelo se ajusta aos dados.

Pesquisadores da Teoria da Informação não acreditam na noção de modelos verdadeiros. Modelos, por definição, são somente aproximações de uma realidade ou verdade desconhecida, não existe modelo capaz de refletir perfeitamente a realidade. George Box fez a famosa afirmação "Todos os modelos estão errados, porém alguns são úteis". Além disso a escolha do melhor modelo para análises de dados depende do tamanho da amostra, pequenos efeitos somente podem ser detectados com amostras grandes, pois a quantidade de informações em amostras grandes é muito maior. Em alguns campos o tamanho dos conjuntos de dados são muitos grandes(terabytes) possibilitando a construção de aplicações muito mais parametrizadas e estruturadas do que em campos com poucos dados. É importante ressaltar que a teoria da informação se suporta no paradigma de que os dados foram coletados adequadamente.

A inferência suportada por modelo é guiada por três princípios:

- *Simplicidade e Parcimônia* - A seleção de modelo é um balanço entre variância e erro de classificação e este é o principio estatístico da parcimônia. Inferência em modelos com poucos parâmetros pode ser viesada, enquanto em modelos com muitos parâmetros podem ser pobres

[1] Entendemos por bondade de ajuste "O quanto melhor o modelo se ajusta aos dados"

em precisão ou em identificar os efeitos que são de fato relevantes. Estas considerações clamam por um balanço entre super e sub ajuste de modelos.

- *Múltiplas Hipóteses* - Aqui não existe hipótese nula, mas várias hipóteses (modelos a serem selecionados) bem justificadas que estão sendo comparadas. São coletados dados reais que são analisados sendo esperado que eles tragam maior suporte a algumas hipóteses e menor suporte a outras hipóteses. Assim em algum momento no tempo devemos ter ainda alguns modelos sobre consideração. O número de modelos em consideração deve ser mantido baixo pois uma análise sobre centenas de modelos não é justificada especialmente em situações com amostras pequenas.

- *Força da Evidência* - O poder do teste é uma parte muito importante da teoria estatística. Testes de hipótese onde o mais importante (o que se quer testar)estáa na hipótese nula tem um poder muito baixo, pois é necessário muita evidência para se dizer que ela não é válida. Assim esses testes tem resultados superficiais, o que faz que no caso de seleção de modelos eles se tornem inúteis.

Em 1951, S. Kullback and R. A. Leibler [13] publicaram o famoso artigo que quantificava o significado de **Informação** . O resultado, chamado Informação de K-L, é uma quantidade fundamental que tem raízes no conceito de entropia. Considerando que f representa toda a realidade ou verdade, f não tem parâmetros. Será usado g para denotar o modelo de aproximação. A informação de K-L $I(f, g)$ é a informação perdida quando se utiliza o modelo g para aproximar f, este é definido como:

$$I(f, g) = \int f(x) log\left(\frac{f(x)}{g(x/\theta)}\right) dx \tag{3.19}$$

A informação K-l é caracterizada como sendo a distância entre a realidade e o modelo, então quanto menor está distância melhor o modelo. Este critério não pode ser usado diretamente pois seria necessário conhecer toda a realidade e os parâmetros θ do modelo de aproximação. Na prática os parâmetros dos modelos são estimados e isso inclui, na maioria dos casos, uma grande incerteza

sobre os parâmetros.

A informação de K-L pode ser expressa como:

$$I(f,g) = \int f(x)log(f(x))dx - \int f(x)log(g(x/\theta))dx \tag{3.20}$$

ou

$$I(f,g) = E_f[f(x)log(f(x))] - E_f[f(x)log(g(x/\theta))] \tag{3.21}$$

onde as esperanças são tomadas com relação a verdade. A quantidade $E_f[f(x)log(f(x))]$ é uma constante C. Então:

$$I(f,g) = C - E_f[f(x)log(g(x/\theta))] \tag{3.22}$$

está constante C não depende dos dados, então somente parte da informação K-L precisar ser estimada a cada modelo.

Akaike mostrou que a parte critica de se usar um rigoroso critério de seleção de modelos baseado na informação de K-L era estimar:

$$E_y E_x[log(g(x/\widehat{\theta}(y)))] \tag{3.23}$$

onde o produto é $E_f[f(x)log(g(x/\theta))]$ com θ substituído pelo estimador de máxima verossimilhança de θ assumindo g e os dados y. Enquanto somente y denota os dados é conveniente considerar que x e y são amostras independentes e identicamente distribuídos. Então está esperança é o alvo em todos os critérios de seleção de modelos baseado na informação de K-L

Akaike(1974) encontrou uma relação formal entre a informação de K-L e a teoria de Máxima Verossimilhança. Ele mostrou que o máximo do log da verossimilhança é um estimador viesado para $E_y E_x[log(g(x/\widehat{\theta}(y)))]$, porém este viés é aproximadamente k, o número de parâmetros no modelo g. Este é um resultado assintótico muito importante pois um estimador não viesado para

$E_y E_x[log(g(x/\widehat{\theta}(y)))]$ em grandes amostras é:

$$log(L(\widehat{\theta}/dados)) - K = C - \widehat{E}_{\widehat{\theta}}[I(f, \widehat{g})] \tag{3.24}$$

onde $\widehat{g} = g\left(\cdot/\widehat{\theta}\right)$

Está descoberta permite combinar as técnicas de estimação dos parâmetros e seleção de modelos sobre o mesmo padrão. Akaike encontrou a esperança do estimador, corrigido pelo viés assintótico, da informação K-L baseada na máxima verossimilhança. Assim o critério de informação de Akaike é:

$$AIC = -2log(L(\widehat{\theta}/dados)) - 2k \tag{3.25}$$

No caso de estimativa com quadrados mínimos com distribuição dos erros normal AIC pode ser expresso como

$$AIC = -2log(\widehat{\sigma}^2) - 2k \tag{3.26}$$

onde

$$\widehat{\sigma}^2 = \frac{\sum(\hat{\epsilon_i})^2}{n} \tag{3.27}$$

e $\hat{\epsilon_i}$ é a estimativa dos resíduos do modelos ajustado. Neste caso *k* precisa ser o total de parâmetros no modelos, incluindo o intercepto e σ^2. Então o AIC é facilmente calculado para ambos os métodos de estimativa de parâmetros.

Considerando um conjunto de modelos candidatos, e sendo todos eles bem suportados por suas teorias, o AIC é calculado para cada modelo. Usando AIC os modelos podem ser ordenados do melhor para o pior baseado somente nos dados disponíveis. Este é um conceito simples suportado por uma teoria bem fundamentada. É importante lembrar a estimativa AIC é assintótica e sendo assim é necessário uma amostra grande para convergir.

Num nível conceitual um modelo construído com dados de boa qualidade permitem a separação

entre informação e ruído, onde informação se refere a estrutura dos dados, e ruído se refere a variância não explicada. Nós sempre queremos modelos que minimizem a perda de informação e separe corretamente o ruído.

A teoria de informação é um método simples e relativamente fácil de empregar em um grande número de situações reais e disciplinas cientificas. Os métodos de seleção de modelos não deve ser usados indiscriminadamente. Um bom conjunto de modelos é essencial, pois envolve profissionais preparados para construir tais modelos e integração entre ciência e a pratica.

3.4.2 BIC

Schwarz em 1978 encontrou o critério de Informação Bayesiano(BIC) como sendo:

$$BIC = -2log(L) + kLog(n) \tag{3.28}$$

onde o melhor modelo tem o menor valor do critério.

A literatura de seleção de modelos, como um todo, é confusa com relação ao BIC nos seguintes aspectos:

1. Para encontrar o BIC é considerado a existência do modelo verdadeiro, ou melhor, que o modelo verdadeiro necessita estar entre os modelos a serem selecionados?

2. O que a probabilidade do modelo significa?

3. O modelo com probabilidade 1 necessariamente é o modelo verdadeiro?

Matematicamente, para uma amostra iid (independente e identicamente distribuída) e um conjunto fixo de modelos, existe um modelo com probabilidade posterior p_t tal que $n \to \infty$, $p_t \to 1$ e as probabilidades de todos os outros modelos tendem a zero. Neste sentido claramente existe um modelo alvo que o BIC procura.

A resposta às perguntas 1 e 3 são simples: Não.

O BIC pode ser encontrado sem considerar que o modelo verdadeiro esteja no conjunto a ser escolhido. Assim ao utlizar o BIC o conjunto não precisa conter o modelo verdadeiro. Por outro lado, a convergência da probabilidade para 1 não significa que o modelo seja o verdadeiro.

A resposta para a segunda pergunta envolve caracterizar o modelo alvo para o qual BIC converge. Este pode ser caracterizado em termos da discrepância de Kullback-Leibler e K(número de parâmetros). Para um modelo g_r a distância K-l do verdadeiro modelo é denotada por $I(f, g_t r)$. Geralmente $g_r \equiv g_r(x/\theta)$ denotaria uma família de modelos paramétricos, mas é utilizado g_r para denotar uma família específica de modelo com $\theta_0 = \Theta$. Para a família $g_r(x/\theta)$ quando $n \to \infty$ o estimador de Mínimos Quadrados e o estimado Baysiano pontual convergem para θ_0, assim assintoticamente podemos caracterizar o modelo representado por $g_r(x/\theta)$. Neste ponto também temos o conjunto de distâncias de Kullback-Leibler.

Podemos assumir, sem perda de generalidade, que os modelos estão ordenados do pior para o melhor, então :

$$I(f, g_1) \geq I(f, g_2) \geq I(f, g_3) \geq \cdots I(f, g_{R-1}) \geq I(f, g_R)$$

Vamos definir Q como sendo a fronteira do conjunto de modelos, e que Q possa considerar o valor R no qual o melhor modelo é único. Quando Q assume mais de um valor, todos os modelos que ele abrange tem o mesmo valor de $I(f, g_t) = I(f, g_{t+1}) = \cdots = I(f, g_r)$, mas mesmo assim consideramos que estes modelos estão ordenados.

Então o melhor modelo no conjunto a ser escolhido é o mais parcimoniso, ou seja o com menos parametros, que neste caso é g_t. Neste cenário a seleção do modelo por BIC converge para 1 e a probabilidade ser selecionado p_t também converge para 1. No entanto se $I(f, g_t) > 0$ o modelo selecionado não é igual ao modelo verdadeiro f, assim o modelo escolhido é chamado de modelo quase-verdadeiro.

A probabilidade Bayesiana a posteriore p_t é a inferência de que o modelo é um modelo quase verdadeiro no conjunto de todos os modelos. Para uma amostra grande o modelo g_t é o melhor modelo

a ser usado para inferência entre todos os modelos em questão. Para amostras pequenas o modelo selecionado pelo BIC poder ser muito mais parsimonioso do que o modelo verdadeiro. A preocupação com o tamanho da amostra é que o modelo selecionado pelo BIC pode ser sub-ajustado, pois o BIC se aproxima, quando n aumenta, do BIC verdadeiro por baixo o que para amostras pequenas pode fazer o valor ser bem diferente do verdaeiro.

Na realidade apenas pode se afirmar que o BIC é assintoticamente consistente para o modelo quase verdadeiro e que para amostras pequenas o estimador é bem viesado. Também do ponto de vista da inferência a probabilidade ser 1 não justifica afirmar que é o modelo verdadeiro.

3.4.3 Exemplo Seleção de Modelo para a Idade

Neste exemplo será utilizado redes neurais do tipo MLP, redes neurais com definições automáticas e regressão logística para estimar a idade dos clientes a partir do número de Cadastro de Pessoas Físicas (CPF). Para avaliar a performance do método de seleção de modelos eles serão calculados utilizando diferentes tamanhos de amostras.

No Brasil o Cadastro de Pessoas Físicas (CPF) é um número atribuído pela Receita Federal do Brasil para que ela possa fazer o acompanhamento do contribuinte. Este número é seqüencial, assim as pessoas mais velhas têm a tendência de ter um número menor que uma pessoa mais nova. Isso é válido atualmente pois no passado, muitas pessoas de idades variadas provavelmente solicitaram o CPF ao mesmo tempo causando uma grande distorção nesta relação. Estas distorções também devem ser observadas em alguns estados mais rurais, pois nestes lugares as pessoas costumam solicitar o cadastro mais tarde.

Encontrar a idade do cliente apenas utilizando a informação do CPF é uma necessidade das instituições financeiras, pois conhecer o cliente com o menor número de informações possíveis pode trazer vantagens competitivas.

Como entrada, ou variáveis dependentes, foram usadas as informações retiradas do número do cpf:

- Quantidade de Dígitos do CPF, exceto os dígitos verificadores,
- Último Digito antes do digito verificador (Estado),
- Os 5 primeiros dígitos do CPF antes do digito verificador

O último digito do CPF identifica em qual estado o CPF foi tirado conforme lista abaixo

- 0 Rio Grande do Sul
- 1 Distrito Federal, Goiás, Mato Grosso, Mato Grosso do Sul e Tocantins
- 2 Amazonas, Pará, Roraima, Amapá, Acre e Rondônia
- 3 Ceará, Maranhão e Piauí
- 4 Paraíba, Pernambuco, Alagoas e Rio Grande do Norte
- 5 Bahia e Sergipe
- 6 Minas Gerais
- 7 Rio de Janeiro e Espírito Santo
- 8 São Paulo
- 9 Paraná e Santa Catarina

A base de dados tem 300 Mil clientes que foram divididos em 60% para Treinamento, 20% para Validação e 20% para Teste. A idade dos clientes foi calculado a partir da data de nascimento com base em outubro de 2006. O software utilizado foi o SAS com Enterprise Miner 5.2. Este software permite a construção e comparação de Redes MLP, Regressão Linear e uma rede neural automaticamente construída. está rede auto construída seleciona automaticamente o número de neurônios, a função de ativação.

O algoritmo utilizado pelo Auto Neural determina a melhor configuração para a rede neural através de uma busca limitada, na qual um nó é adicionado por vez e os pesos são treinados. Para o treinamento seguinte a contribuição dos neurônios anteriores é retirada. O algoritmo pode funcionar adicionando neurônios em uma única camada ou adicionando camadas com tamanho uniforme. O software tem vários parâmetros de controle ente eles: o número mínimo de neurônios, método de parada (tempo, número de interações, validação cruzada, etc), como os neurônios devem ser adicionados, etc. Além disso o algoritmo testa qual a melhor função de ativação para cada treinamento. As funções possíveis são:

Função	**Intervalo**	**Função de g**
Identidade	$(-\infty, +\infty)$	g
Exponencial	$(0, +\infty)$	$exp\,(g)$
Recíproca	$(0, +\infty)$	$1/g$
Quadrado	$(0, +\infty)$	g^2
Logistic	$(0, 1)$	$\frac{1}{1+e^{-g}}$
Softmax	$(0, 1)$	$\frac{e^{-g}}{\sum Exponenciais}$
Seno	$[-1, 1]$	$sin(g)$
Coseno	$[-1, 1]$	$cos(g)$
Tan Hiper	$(-1, 1)$	$tanh(g)$

Para analisar o comportamento do critério de seleção dos modelos de acordo com a variação do tamanho da amostra os modelos foram construídos com tamanhos de amostra diferentes e calculando assim o valor do critério para casa situação é possível avaliar qual seria a decisão para cada tamanho de amostra.

As figuras 3.10 e 3.11 mostram a comparação entre os valores dos critérios para cada tipo de modelo e tamanho de amostra. O melhor modelo é o com o menor valor do critério, assim para cada tamanho de amostra foi marcado qual seria o melhor modelo.

Para efeito de comparação consideramos que a melhor decisão é a com o maior tamanho de

Critério de Akaike				
	Tamanho da Amostra			
	300.000	3.000	300	60
Auto Neural	1.531.220,2	15.406,4	1.600,4	330,8
Rede MLP	1.560.318,5	15.486,2	1.606,6	329,1
Regressão	1.631.389,2	16.348,0	1.645,6	329,5

Fig. 3.10: Tabela de Comparação para o Critério de Akaike

Critério de Schwarz				
	Tamanho da Amostra			
	300.000	3.000	300	60
Auto Neural	1.531.825,1	15.916,9	1.707,9	360,1
Rede MLP	1.560.774,8	15.744,4	1.765,8	406,6
Regressão	1.631.516,6	16.414,1	1.649,3	331,6

Fig. 3.11: Tabela de Comparação para o Critério de Schwarz

amostra existente. Assim ao compararmos as decisões tomadas pelo critério de Akaike observa se que apenas com tamanho de amostra igual a 60 a decisão pelo melhor modelo foi diferente da tomada com o maior tamanho de amostra, mostrando que o critério não é fortemente influenciado pelo tamanho de amostra.

Para o BIC pode se verificar que ele é muito mais influenciado pelo tamanho da amostra pois já com amostras de tamanho três mil a escolha do modelo é diferente da com amostra de trezentos mil. Como sugere a teoria o BIC, para amostras pequenas ele é viesado e seleciona um modelo mais parcimonioso que os modelo verdadeiro, neste caso o BIC está selecionando a regressão logistica que é sabidamente um modelo mais parcimonioso que as redes neurais.

Nota-se também que no caso da maior amostra a escolha do melhor modelo feita através do AIC e do BIC é concordante, já no caso dos outros tamanhos de amostra elas são discordantes o que pode levar a tomar decisões equivocadas. Neste este o Akaike se mostrou um critério mais estável e menos dependente do tamanho da amostra, indicando que ele pode ser um critério mais estável e que seleciona mesmo para amostras relativamente pequenas, um modelos mais próximo do modelo

verdadeiro.

Capítulo 4

Ensemble

4.1 Introdução

O empilhamento (*Ensemble*) de modelos utilizando redes neurais, que foi proposto por L.K HANSEN e P. SALAMON [14], é um método de modelagem que visa minimizar a taxa do erro de generalização de modelos com uma ou mais variáveis a serem preditas por treinarem várias redes neurais e combina las. O empilhamento funciona reduzindo o viês que um modelos tem com a sua amostra de desenvolvimento. Essa redução funciona generalizando em um segundo espaço no qual as entradas são as saídas dos modelos construídos na primeira fase. Quando usado com modelos que tem várias respostas o empilhamento pode ser visto como um sofisticado sistema.

Os métodos de *Ensemble* vêm sendo aplicados em redes neurais utilizando vários métodos para combinar os classificadores. Podemos expandir este conceito para utilização de outros métodos de combinar as variáveis originais. Podem ser utilizados técnicas lineares ou de aprendizado de máquina para se produzir melhores classificadores.

A seguir apresentamos alguns métodos conhecidos de combinação de classificadores e depois será discutido um método proposta para dados de risco de crédito.

4.2 Revisão dos Métodos de *Ensemble*

A combinação de classificadores tem sido utilizada por várias razões diferentes, ou se tem várias amostras, diferentes características a serem modeladas ou ainda diferentes resultados de treinamentos para uma determinada técnica. Em resumo estas situações geram vários modelos que precisam ser combinados com a esperança de gerar uma maior precisão no resultado. Estas situações podem acontecer de várias maneiras:

- O projetista do modelo tem várias amostras retiradas em diferentes contextos e representações. Por exemplo, dados em diferentes meses do ano em uma população que tenham sazonalidades.
- Algumas vezes o projetista tem acesso aos vários modelos construídos para diferentes situações e ele precisa construir um terceiro uso dos modelos originais. Por exemplo, ele tem um modelo para ofertar limites de cheque especial e outro para financiar veículos e ele precisa construir um para ofertar produtos parcelados com garantia.
- Vários treinamentos de uma rede neural para o mesmo conjunto de dados, porque é necessário escolher o melhor se é possível combinar e utilizar todos.
- Em amostras muito grandes muitas vezes é impraticável treinar a rede com todo o conjunto de dados, então é possível se ter várias amostras dos conjuntos de dados resultando assim em vários modelos.

Existem várias maneiras de se utilizar o método de *Ensemble* e vários autores tem proposto sugestões de como combinar os modelos e até agora não existem um método definitivo. Uma típica combinação de modelos consiste em ter um conjunto de modelos (ou classificadores) e combinar o seus resultados individuais. O modelo, como os modelos interagem e quando cada um deve ser chamado é determinado pela arquitetura de combinação.

Existem várias formas de combinação, Duinand Mao Jain [15] listou dezoito estratégias para se fazer a combinação dos modelos e as classificou em três categorias:

- Paralela,
- Serial,
- Hierárquica.

Na arquitetura paralela todos os modelos são convocados ao mesmo tempo e seus resultados são combinados por uma regra. Na arquitetura serial os modelos são chamados serialmente e os resultados são combinados chamando mais modelos até parar. Usualmente se utilizam modelos mais baratos computacionalmente e menos precisos no início, para depois se utilizar modelos mais caros computacionalmente e mais precisos. Na arquitetura hierárquica os modelos estão dispostos como numa árvore de decisão e de acordo com a resposta de um modelo um determinado tipo de modelo ou outro tipo é chamando. Isso permite usar modelos diferentes em diferentes regiões do espaço de busca da solução.

Cada uma destas estratégias tem suas vantagens e desvantagens. Abaixo listamos algumas das técnicas retiradas do artigo de Duinand Mao Jain:

- Votação,
- soma, média, mediana,
- produto, minímo e máxímo,
- *ensemble* generalizado,
- erro adaptativo,
- empilhamento,
- regressão logística,
- arvore neural,
- redução do conjunto de respostas,

- *Bagging*,
- *Boosting*.

Entre as técnicas citadas as três primeiras são de aplicação direta, sendo as mais primitivas em termos de relacionamento dos modelos. As técnicas mais conhecidas para *ensemble* são o *Bagging*, *Boosting* e o empilhamento (*Stacking*).As três técnicas tem modos diferentes e iremos contemplar uma idéia geral de utilização de cada uma delas.

• *Bagging*

O nome *bagging* vêem de *Bootstrap aggregation*, foi o primeiro método efetivo de *ensemble* e é um dos mais simples métodos. O algoritmo foi originalmente desenvolvido para modelos de classificação mas é muito usado em árvores de classificação. Os passos do algoritmo são:

1. Gerar N amostras com reposição do conjunto original de dados (Amostras *Bootstrap*),
2. Construir modelos para cada uma das amostras,
3. Combinar as saídas por média ou votação.

Para a metodologia de *Bagging* ter bons resultados comparado com as metodologias de um só modelo, é necessário que as amostras geradas ou o tipo de modelo seja muito instável podendo assim gerar amostras muito diferentes. Em ambos os casos os modelos gerados seriam diferentes o suficiente para justificar uma combinação das saídas destes.

• *Boosting*

O método de *boosting* pode ser visto como a criação de um modelo médio. Ele é um dos métodos de *ensemble* mais utilizados para se combinar modelos. Os passos para construção do algoritmos são:

1. Criar um modelo inicial, que seja um pouco melhor que uma seleção aleatória,

2. Selecionar as amostras para as quais este modelo erra a classificação,

3. Construir outro modelos para estas amostras,

4. Ponderar de acordo com a taxa de acerto dos modelos,

5. Combinar as saídas dos modelos pela média ou votação considerando os pesos.

Assim é possível utilizar modelos simples que tenham baixa capacidade de precisão, porém podem ser calculados com pouco esforço computacional, para criar modelos com alta capacidade de precisão. Isso na prática facilita a implementação exigindo menos dos sistemas de cálculo.

• *Stacking*

O empilhamento de modelos é um método pouco usado apesar de oferecer a possibilidade de se combinar diferentes tipos de metodologia, ao contrário dos dois anteriores, de construção de *ensembles*. Os passos para a utilização da técnica são:

1. Separar o conjunto de dados em dois conjuntos disjuntos.

2. Treinar várias tipos de modelos no primeiro conjunto.

3. Testar estas técnicas no segundo conjunto.

4. Usar as predições feitas no passo anterior como entradas de um outro modelo e estimar o último modelo.

Pode se notar que o modelo no passo 4 é uma combinação de modelos que pode ser feita usando algumas das técnicas anteriores ou até mesmo todas e construindo um *bagging* no final.

As técnicas de *ensemble* tem se desenvolvido muito nos últimos anos, porém ainda temos muito pouca confirmação teórica para os resultados encontrados na prática. As provas que existem ainda são para os métodos mais simples com aplicações diretas de outras técnicas como *Bootstrap*.

4.3 Método de Ensemble Proposto

O uso de *Ensemble* é motivado pelo fato de termos uma melhor precisão dos modelos o que tem sido mostrado por vários pesquisadores, [16], [17] e [18]. Ao invés de escolher a melhor técnica entre regressão logística e RNAs será utilizado uma modificação no método de construção de modelo de crédito sugerido pela estratégia 3 adotada na dissertação de mestrado de Gustavo Henrique [19].

A estratégia 3 sugerida pela tese são ajustados simultaneamente modelos para cada produto ou conta dos clientes. Após criados estes modelos por contas eles são combinados.

Neste método, as informações são agrupadas por produto e os modelos são construídos em paralelo, assim ambas as técnicas são utilizadas para se construir um modelo final que possa absorver todas as características boas de ambas as técnicas. Na classificação adotada anteriormente este modelo se encaixa na característica de paralelo na metodologia de empilhamento.

O uso do *Ensemble* também facilita o uso de RNAs em dados de *Behaviour Scoring* pois estes tipos de dados tem algumas características que dificultam a aplicação direta da rede neural:

- Alto volume de dados. As bases podem passar dos 10 milhões de registros.

- Grande quantidade de informações a serem avaliadas. Normalmente instituições financeiras guardam um registro de todas as atividades dos clientes, assim pode se utilizar mais de mil características de comportamento para cada cliente e dependendo da criação delas pode chegar a até dez ou quinze mil.

- Devido a característica das instituições financeiras de criarem conglomerados atualmente elas oferecem uma grande quantidade de diferentes produtos para diferentes públicos ou não , podendo assim o cliente ter uma grande combinação de produtos. Então como criar modelos para tender a todos? É melhor fazer modelo para cada combinação?

- Como incluir informações macro econômicas para os modelos se adaptarem a oscilações de mercado?

O método utilizado é uma variação do método de empilhamento de modelos. Neste caso, diferente do empilhamento tradicional, é feita uma sugestão de como separar os dados, não com amostra de indivíduos disjuntos mas sim com separação das informações de entradas sendo disjuntas. Então podemos citar os passos para se descrever o método como sendo:

1. Separar o conjunto de dados em vários conjuntos de acordo com as características das informações de entrada.

2. Construir modelos para cada um dos conjuntos de informações de entrada, podendo as saídas serem de acordo com cada conjunto de dados. Até mesmo o conjunto de dados pode ser outro.

3. Unir as saídas dos modelos em um único conjunto de dados,

4. Usando as predições feitas no passo anterior como entradas de um outro modelo, estimar o modelo final com a saída desejada como resposta.

Com estes passos é possível se construir um modelo em situações que as informações de entrada do modelo podem ser agrupadas por afinidade ou condição externa ao modelo,como por exemplo:

- Modelo para reconhecimento de fala em situações que pode ter vários microfones em posições diferentes. Cada microfone pode representar um conjunto e variáveis para modelos separados, sendo unidos no final para gerar uma única resposta.

- Modelos para previsão do tempo. Será que as informação de como está a umidade do ar do região deveriam estar junto com informação de chegada de uma frente fria. Porque não fazer modelos separados e somente unir no final?

Este também é o caso de produtos de crédito. Muitas vezes é necessário se ter uma avaliação de risco global do cliente, porém as informações de entrada estão fortemente agrupadas de acordo com os produtos de crédito que ele possui, e cada um dos produtos tem diferentes características como forma de parcelamento, se é parcelado, percentual pago e outras. A seguir detalhamos a utilização destes passos especificamente em modelos de risco de crédito.

No caso de risco de credito o método proposto se inicia com o planejamento do modelo. Neste fase é preciso definir:

- seleção do publico alvo (amostragem) e
- quais produtos irão participar do modelo final,
- qual a resposta desejada do modelo final,

Está fase é muito importante pois determina toda a construção dos modelos, de quantos serão feitos e qual o nível de abrangência destes. Aqui devemos nos preocupar com:

- qual população será utilizada em cada modelo, por exemplo : para um produto que tem poucos clientes pode se decidir utilizar somente os que tem o produto, pois assim será maximizada a utilizacão de informações a respeito deste produto.
- a resposta utilizada em cada modelo será se o cliente não pagou aquele produto específico ou se ele não pagou nenhum dos produtos? Deve se levar em consideração a resposta desejada do modelo final.
- Qual o tipo de modelo utilizado em cada uma dos produtos, modelo linear, não linear, RNA, etc.

Após selecionado os produtos e amostras, as variáveis explicativas devem ser agrupadas por sua origem, ou seja, se a informação é do produto então ela será utilizada no conjunto das informações do modelo para o produto 1. Este tipo de agrupamento tem por objetivo utilizar ao máximo as informações de comportamento de cada produto evitando que alguma informação fortemente correlacionada com uma informação de outro produto acabe por reduzir sua importância ficando assim fora do modelo final. Não pode se esquecer, nesta fase, das informações de caráter geral tais como:

- informações do cadastro do cliente,

- informações de *Bureau* externo,
- informações positivas.

Estas informações devem ser agrupadas em um conjunto separado como se fosse um produto, gerando assim um modelo a parte. Isso é sugerido porque estas informações são dis-associadas de qualquer produto e ao mesmo tempo estão influenciando em todos eles. Assim se elas formarem um produto em separado poderão no modelo final influenciar a todos.

Com as variáveis explicativas agrupas por produto é construído um modelo para cada grupo de variável tendo como resposta se o cliente é bom ou mau pagador no produto em que as variáveis estão mais relacionadas. Neste modelo é utilizada uma técnica de seleção de variáveis para se selecionar as informações mais relevantes para explicar o mau cliente. É recomendado que se utilize alguma seleção de variável para o modelo final seja mais parcimonioso.

Aqui também são feitos todos os testes de estabilidade e qualidade do ajuste do modelo, como se o modelo fosse independente, ou seja, cada modelo tem que ter um bom desempenho individualmente. As variáveis explicativas podem ser trabalhadas criando se *dummys* para as informações categóricas, agrupando os clientes com taxa de bom/mau próximas e as contínuas devem ser avaliadas para se encontrar a melhor transformação, que tenha uma relação linear com a probabilidade de mau cliente, exigência no caso de se utilizar a técnica de regressão logística.

Com os modelos por grupo de produtos prontos, ou seja, já com as melhores variáveis explicativas selecionadas e os parâmetros ajustados, iremos utilizar o valor estimado de probabilidade de sinistro de cada modelo em cada produto e combina-los utilizando um modelo (este modelo pode ser do mesmo tipo dos anteriores ou ainda um outro) se o cliente for mau em qualquer um dos produtos então ele é mau em todos os produtos. Este modelo combinado tem a função de agrupar as respostas dos modelos iniciais de cada produto para assim fornecer uma probabilidade de o cliente ser mau em todos os produtos de crédito que ele possuir gerando assim uma visão consolidada do cliente.

A figura 4.1 mostra o funcionamento da metodologia de uma forma gráfica. Nela pode se ver cada uma das informações (Var1.1, Var @.1, etc) de entrada conectada a um produto em específico,

elas podem se conectar diretamente ou através de interações, se tiver sentido (Multiplicação das informações de entrada). A saída destes modelos é conectada como entrada do modelo final que combina todas elas gerando uma resposta única.

Neste caso, o método de *Ensemble* pode ser considerado paralelo, pois temos vários modelos sendo calculados desta forma e hierárquico, pois no final temos um modelo que depende dos anteriores para ser calculado.

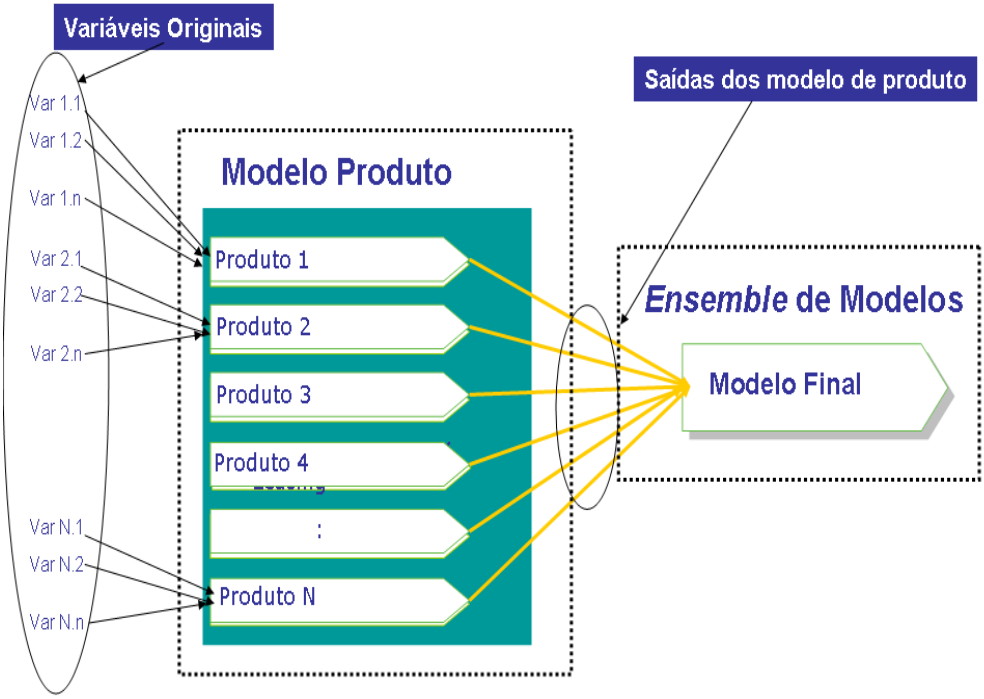

Fig. 4.1: Esquema de Montagem do Modelos

Neste tipo de configuração de construção dos modelos pode se citar várias vantagens, entre elas:

1. Os modelos por produto podem ser construídos individualmente, podendo até serem construídos em paralelo, reduzindo assim o tempo de construção do modelo.

2. Cada modelo individual tem um número reduzido de variáveis proporcionando maior agilidade na escolha das informações realmente relevantes.

3. Oferece técnicas com método objetivo para selecionar as variáveis originais como a regressão logística.

4. O número reduzido de informações no modelo final permite a aplicação de uma Rede Neural para assim utilizarmos a capacidade de aproximações não lineares oferecidos pela rede.

5. O modelo final tem a capacidade de absorver características não lineares entre os modelos e a probabilidade do cliente ser mau pagador.

6. Possibilidade de controlar onde devem ser inseridos os modelos mais complexos como modelos não lineares evitando assim tempo excessivo de processamento.

7. Facilidade de troca ou inclusão de novos produtos, fato comum no mercado financeiro pois a todo momento novos produtos são criados ou características dos existentes são alteradas. Neste caso não precisamos analisar todas as informações novamente apenas as que serão substituídas e rever o modelo combinado.

8. Tempo total gasto para se construir o modelo é menor do que se fosse aplicada a rede neural diretamente por causa do tamanho das bases.

9. Garante no modelo final a representatividade de todas as informações dos produtos, pois produtos que tem poucas observações tendem a ter suas características resumidas a um indicador se o cliente tem ou não o produto.

10. Maior estabilidade ao longo do tempo. Modelos de Behaviour perdem a eficiência com a sua utilização pois o perfil de cliente mau pagador altera. Com mais informações representadas no modelo este perfil tem uma abrangência maior o que permite uma maior estabilidade.

11. está configuração também facilita o uso dos modelos porque com bases segmentadas de colunas pode se ter uma melhor performance no cálculo dos modelos mensalmente, apesar de se ter mais modelos a serem calculados. Só para se ter uma idéia hoje grandes instituições financeiras podem levar até 2 semanas para calcular todos os modelos.

Para avaliar a eficácia do método proposto iremos aplica lo em um conjunto de dados reais. Para avaliar os modelos combinados iremos utilizar o Critério de Akaike e o Critério de Informação de Bayes.

4.4 Teste Método *Ensemble*

Para avaliar as metodologia de construção de modelos apresentadas neste trabalho iremos utilizar dados reais de construção de modelo de behaviour. Os dados são reais porém os valores foram alterados para que não seja possível identificação de pessoas ou mesmo a utilização destes dados por outras instituições. A base contem um milhão e meio de clientes com informações de movimentação da conta corrente, de pagamento de produtos parcelados, pagamentos de produtos com garantias. Como é comum, nas aplicações práticas, em modelos de *behaviours* foram observados 4 meses de histórico para se avaliar as informações do cliente e o sinistro foi avaliado 12 meses após o mês de referência.

O primeiro passo é definir os produtos serem abordados. Na base em questão iremos trabalhar com:

- Cheque Especial,
- Parcelados sem Garantia,
- Parcelados com Garantia,
- Informações de Bureal Externos,
- Informações Cadastrais,

O próximo passo foi agrupar as informações de acordo com o produto. Assim as informações de movimentação de conta corrente foram utilizadas no modelo de produto rotativo, as informações de pagamentos de parcelados foram utilizadas nos produtos parcelados e as informações dos produtos com garantia foram utilizadas nos respectivos produtos. A baixo listamos as principais informações usadas por tipo de produto.

- **Cheque Especial**
  - Saldos de conta
  - Tipos de pagamentos efetuados
  - Giro financeiro
  - Índice de Utilização
- **Parcelados Sem Garantia**
  - Tempo de contrato
  - Percentual pago do Contrato
  - Forma de pagamento (Boleto ou débito em conta)
- **Parcelados Com Garantia**
  - Tempo de contrato
  - Valor do Bem
  - Tipo de Garania
- **Informações Negativas de Mercado**
  - Tipo de negativação
  - Giro financeiro
- **Informações Cadastrais**

- Idade
- Sexo
- Quantidade de Dependentes

Neste exemplo os modelos foram construídos cada um com o seu público, ou seja o público utilizado no modelo para cheque especial é somente quem tem o produto e assim em todos os outros produtos. O modelo final tem como público qualquer cliente que esteja em um dos modelos anteriores. É importante lembrar que modelos que não têm produto associado também foram construídos com todo o público.

Após separadas as informações foram construídos modelos de behaviour para cada produto separadamente considerando como mau cliente se este atrasou mais de 90 dias no produto em um período de 12 meses. Os modelos individuais foram construídos através da técnica de Regressão Logística , utilizando o método de stepwise para selecionar as variáveis originais mais importantes. A base de dados recebido já está com as saídas destes modelos prontas, ou seja, não é objetivo aqui discutir a construção destes modelos, mas sim como combiná-los.

Como os modelos foram construídos cada um com o seu público, podemos também avaliar o desempenho deste com relação ao seu público original. Como desempenho, neste caso, iremos utilizar a estatística de Kolmogorov-Smirnov conhecida como KS. Como conhecida, está estatística está definida entre 0 e 1 e quanto mais próximo de 1 melhor o modelo.

As fíguras 4.2 a 4.6 mostram qual é o desempenho dos modelos individualmente para se prever se o cliente vai deixar de pagar qualquer um dos produtos analisados nos próximos 12 meses. Nelas temos o valor do KS para cada modelo e também um gráfico que mostra o percentual de maus pagadores em faixas de 10% da populaçào do modelo. As faixas de maiores valores são as que tem menores taxas des de maus clientes.

Estes gráficos com as taxas ajudam a identificar a qualidade do modelo, pois pelas definições de Basiléia adotadas na Europa (as regras no Brasil ainda não estão totalmente divulgadas pelo Bacen, porem acredita se que devem seguir as regras da Europa) o modelo não pode ter concentração e as

taxas de maus clientes devem ser decrescentes.

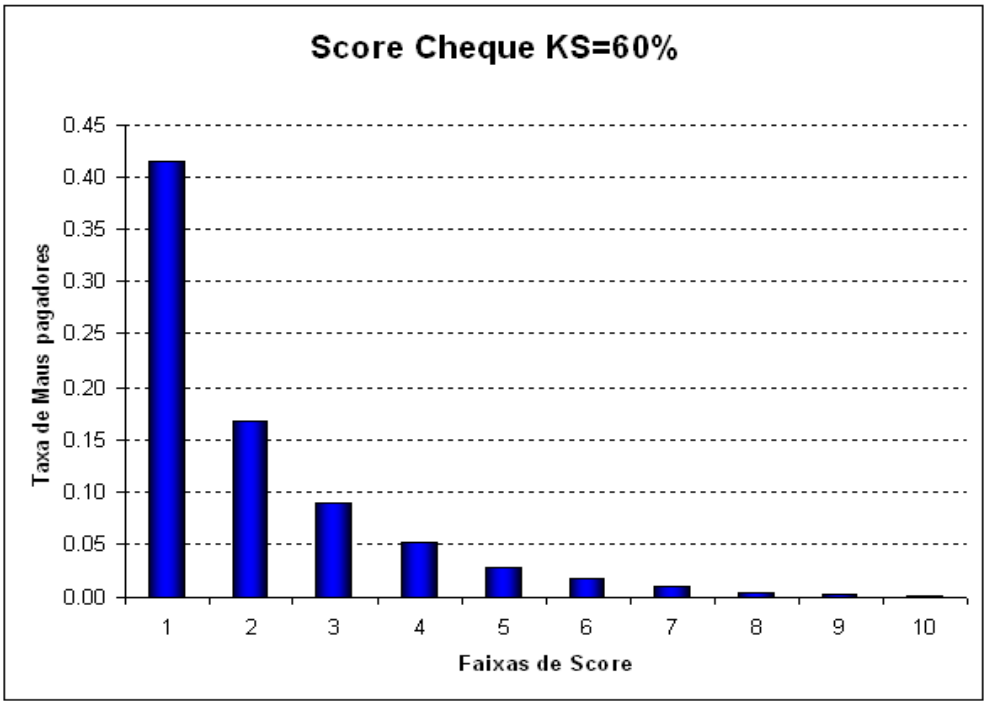

Fig. 4.2: Desempenho modelo de Cheque Especial

Em alguns destes modelos pode se ver que eles não conseguem manter a taxa de maus clientes em ordem decrescente, indicando que sozinhos não tem um bom desempenho. O modelo de informações de mercado também apresenta concentrações de pontuação com o mostra a figura 4.7, pode se notar que ele tem faixas com mais de 20% de clientes não sendo possivel inclusive criar 10 faixas, foram possíveis somente oito faixas.

Então seguindo a metodologia proposta pelo Gustavo na sua dissertação estes modelos individuais serão combinados, criando assim o modelos final.

Assim com os modelos individuais prontos, pode se partir para a construção de *Ensemble*. A entrada do modelo de *Ensemble* são as probabilidades estimadas pelos modelos logísticos individuais mais um indicador se o cliente tem aquele produto ou não. Este indicador é importante pois ele ajuda a equalizar a distribuição das taxas de maus clientes na população.

As interações também foram levadas em consideração, ou seja, foi avaliado o quando o risco de

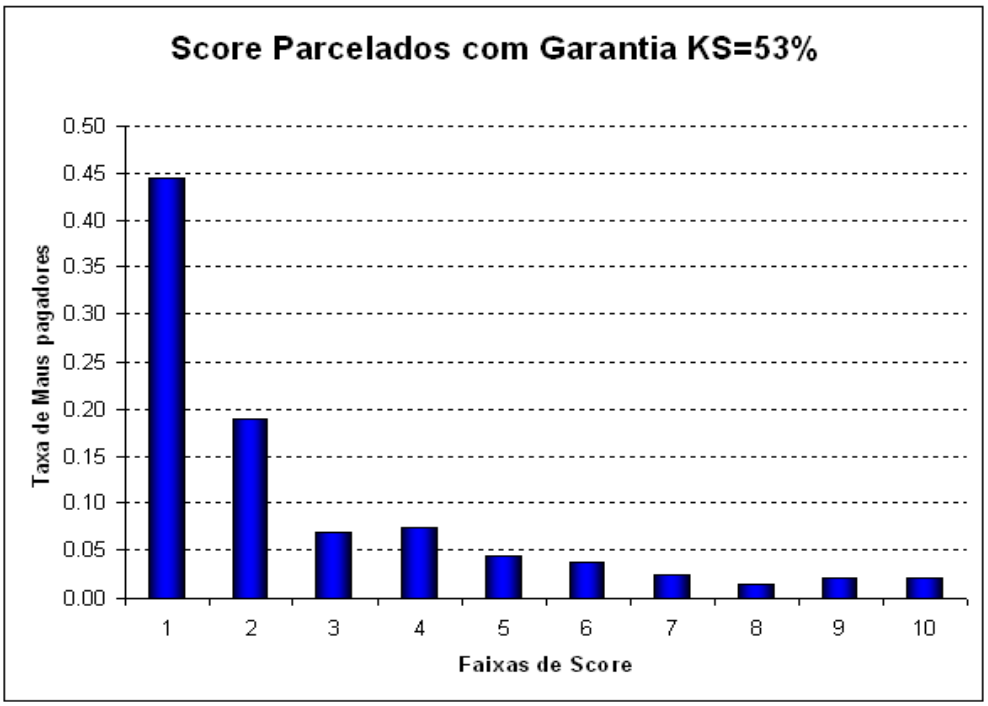

Fig. 4.3: Desempenho modelo de Produtos Parcelados com Garantia

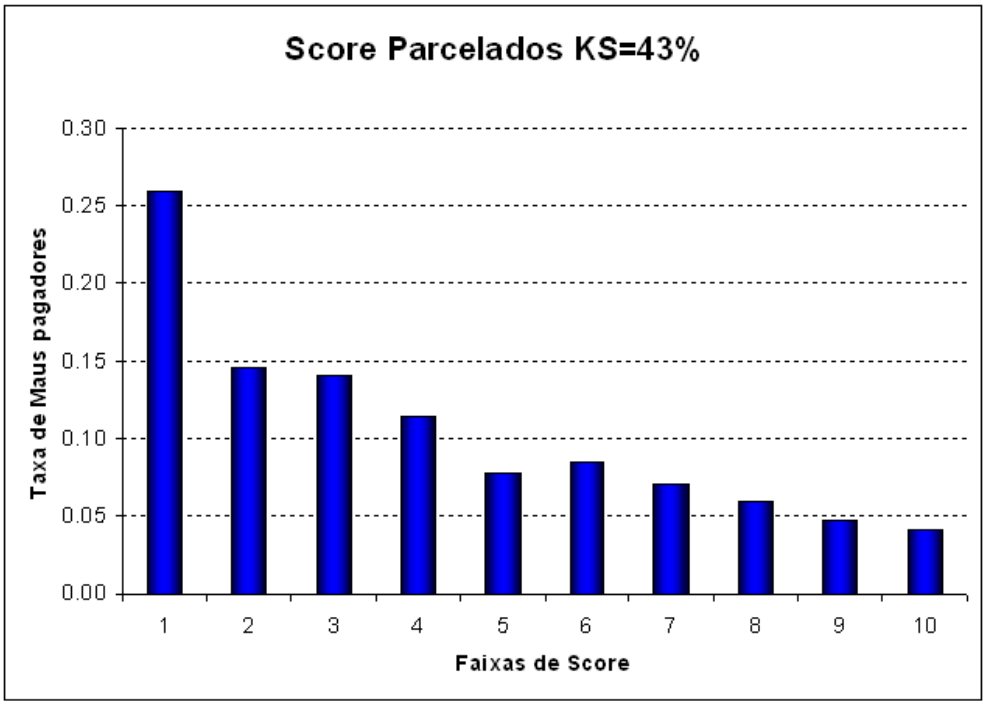

Fig. 4.4: Desempenho modelo de Produtos Parcelados sem Garantia

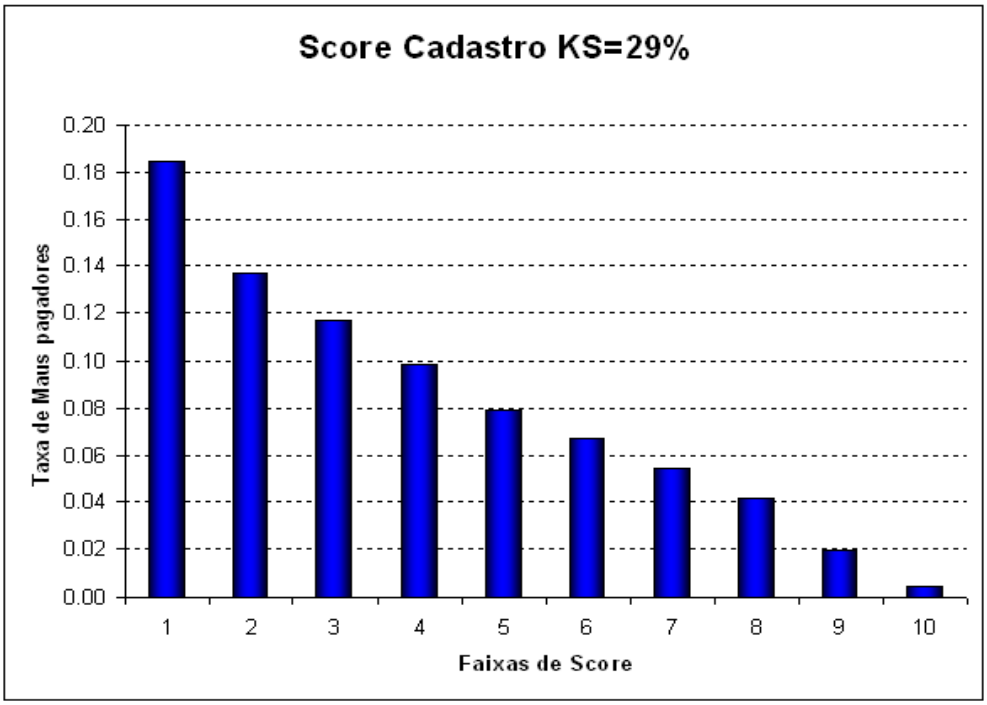

Fig. 4.5: Desempenho modelo para Informações Cadastrais

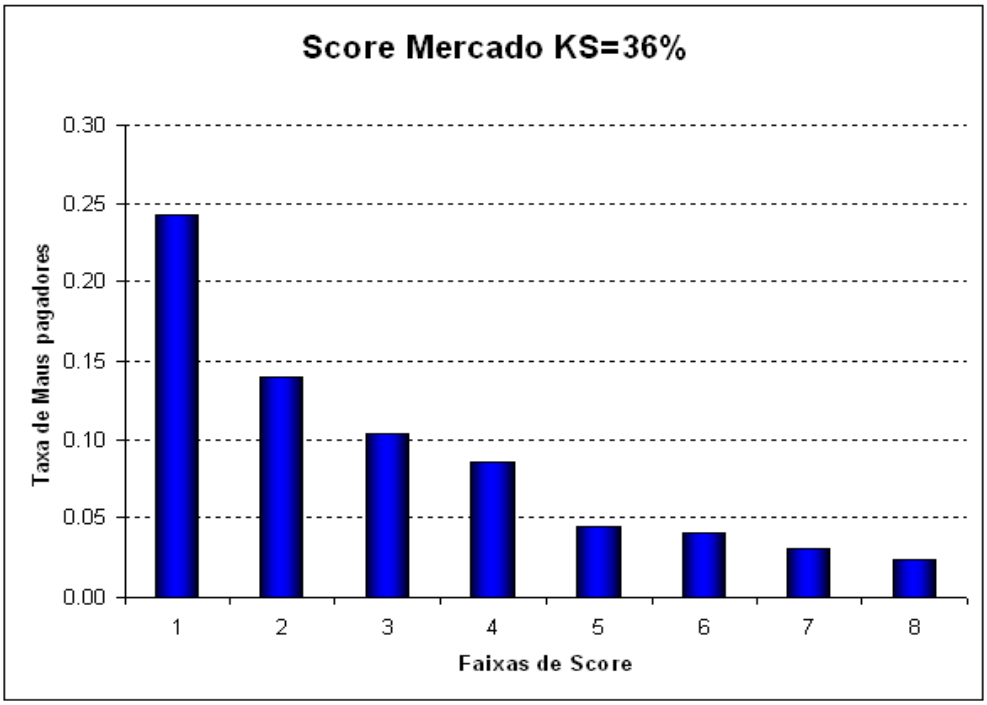

Fig. 4.6: Desempenho modelo para Informação Negativas de Mercado

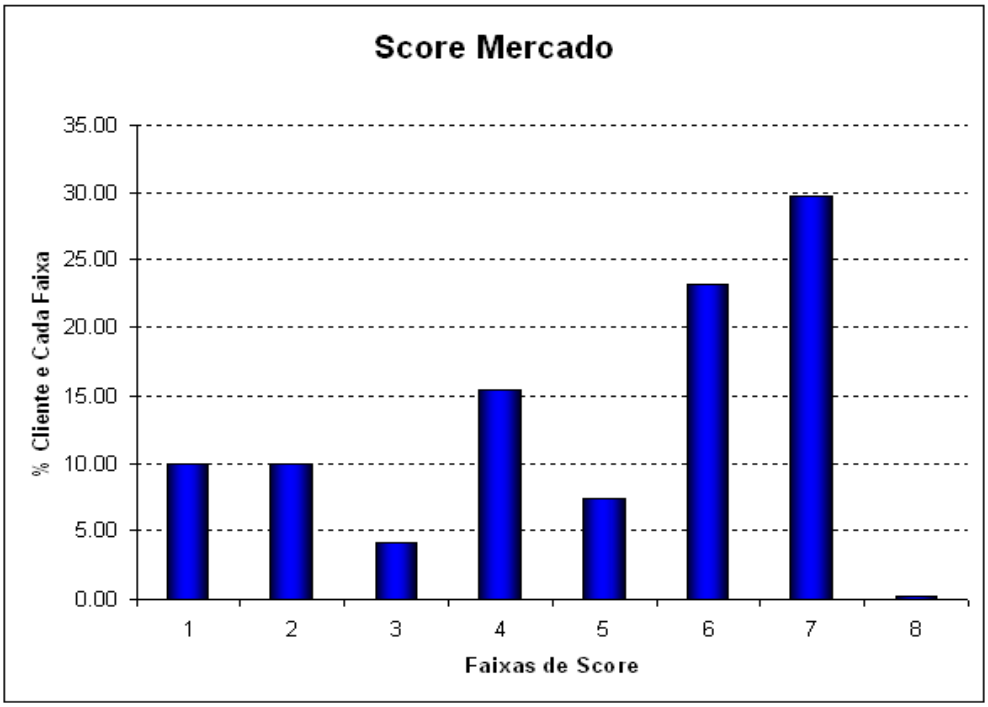

Fig. 4.7: Distribuição modelo para Informação Negativas de Mercado

um produto afeta o outro produto. Para isso as probabilidades multiplicadas dois a dois também foram analisadas.

Para efeito de comparação utilizamos três métodos de *Ensemble*:

1. Rede Neural do Tipo MLP,
2. Regressão Logística,
3. Regressão linear
4. Estratégia Evolutiva

Os três primeiros modelos utilizados são técnicas mais conhecidas para a solução deste tipo de problema e já tratadas neste trabalho. Todas elas podem ser utilizadas para se construir um *Ensemble*. A quarta técnica utilizada para se estimar os parâmetros é a estratégia evolutiva. A idéia de utilizar

uma técnica que possa trabalhar com poucas amostras ou até mesmo conseguir resultados melhores que as técnicas habituais.

A estratégia evolutiva utilizada tem como principal operador evolutivo o método da Mutação Unidimencional e a estratégia é comandada pelo algoritmo Dopt-AiNet [20] . A figura 4.8 mostra o funcionamento deste algoritmo.

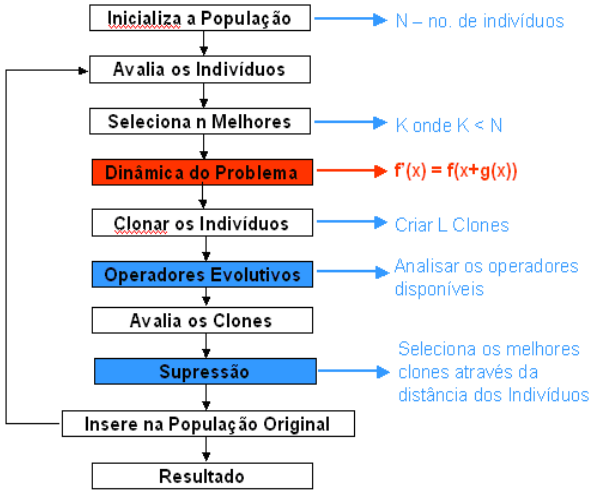

Fig. 4.8: Doptainet

O procedimento de mutação unidimensional gera um conjunto de direções que são praticamente impossíveis de serem obtidas utilizando uma distribuição gaussiana. Ele gera uma matriz D com elementos d_{ii} = 1 e d_{ij}= 0 para *i* diferente *j*, ou seja, fazendo com que seja possível percorrer uma dimensão de cada vez, dando o nome à mutação.

Também são utilizados os vetores direcionais 1 e -1 (compostos por todos os elementos iguais a 1 e -1, respectivamente) e pela matriz gaussiana de direção contendo N_c direções. Ou seja, a partir de agora serão gerados $N_c + n + 2$ clones para mutação, sendo n a dimensão do espaço de busca.

Determinadas as direções, as demais etapas ocorrem de forma similar à mutação tradicional. O pseudo código para a mutação unidimensional é apresentado a seguir:

Neste algoritmo, os parâmetros de entrada são a matriz de clones C e a função a ser otimizada f. A função a identidade(n) gera uma matriz identidade $n \times n$, e a *função aurea(c,d,f)* retorna o tamanho

```
[C] = Função multunidimensional(C,f)
   G = gaussiana();
   D = [identidade(n); 1; -1; G]
   Para cada vetor "c"da matriz C de clones faça,
      Para cada linha "d"da matriz D faça,
         alpha= aurea(c,d,f)
         c' = c + alpha *d
         Se f(c') é melhor do que f(c),
            c = c'
         Fim;
      Fim;
   Fim;
Fim;
```

do passo na direção *d* partindo do ponto *c*. Note que essa mutação já inclui a mutação gaussiana, tornando desnecessária a execução desta parte em separado.

O tamanho do passo que o clone vai dar em cada direção é determinado pela função áurea. Em algoritmos clássicos de otimização, como os métodos de gradiente e de Newton, dada a direção de maior decrescimento o tamanho desse passo é calculado através de uma busca unidimensional.

Um método que possui características de convergência global para funções unimodais e convexas, sem a necessidade de informações sobre a função, é o método de seção áurea [21]. Esse método é baseado nos algoritmos de enumeração para programação não-linear, nos quais são definidos intervalos fechados do espaço de busca cada vez menores em torno do ótimo global.

O ponto chave desse algoritmo é a forma como ele sub-divide o intervalo de busca utilizando um número conhecido como razão áurea. A razão áurea é um número encontrado em proporções de diversas figuras geométricas na natureza, como em espirais nas conchas de moluscos, o crescimento de uma concha sem alteração de seu formato, a razão entre a medida de nosso braço e o antebraço, e muitos outros. Essa medida tem sido utilizada de longa data em projetos de arquiteturas da Grécia antiga até pinturas feitas por Leonardo da Vinci. A razão áurea é denominada pela letra grega $\phi = \frac{(1+\sqrt{5})}{2}$ seu valor é o número irracional que vale aproximadamente 1,618.

O método da seção áurea funciona da seguinte forma: dado um intervalo fechado pré-estabelecido

[a,b], dois pontos dentro desse intervalo são gerados utilizando a razão áurea de forma que:

$$\alpha = a + (1 - |\phi - 1|)(b - a) e \alpha = a + (\phi - 1)(b - a) \tag{4.1}$$

O pseudo-algoritmo é dado abaixo:

```
[α] = Funcao aurea (x0,d,a,b,f);
   r = (5-1)/2;
   α = a + (1-r)*(b-a);
   β = a + r*(b-a);
   y1 = f(x0 + *d);
   y2 = f(x0 + *d);
   Enquanto |b-a|> ,
      Se y1>y2,
         a = α;
         β = α;
         y1 = y2;
         beta = a + r*(b-a);
         y2 = f (x0 + β*d);
      Senão,
         b = β;
         β = α ;
         y2 = y1;
         α = a + (1-r)*(b-a);
         y1 = f(x0 + α *d);
      Fim;
   Fim;
   α = (b+a)/2;
Fim
```

Os testes foram realizados no conjunto de dados disponível e no capítulo 5 serão apresentados e discutidos os resultados encontrados.

Capítulo 5

Resultados

Para comparar os modelos construídos no capitulo anterior sugere se [22] a utilização do critério de Akaike, Critério de informação de Bayes e a estatística de Kolmogorov-Smirnov(KS) . A tabela 5.1 mostra os resultados obtidos com os modelos propostos. No caso das Estratégias evolutivas também está sendo comparado o desempenho em diferentes tamanhos de amostras. Estes testes foram feito devido ao grande tempo necessário para processamento do algoritmo. Para algumas amostras foi executado o algoritmo mais de uma vês.

Modelo	KS	Akaike	Critério Baysiano
Regressão Linear	60.1%	22.354	26.731
Regressão Logistica	59.6%	22.350	26.726
Redes Neurais	60.6%	22.345	26.721
EE Amostra 300	42.2%	-	-
EE Amostra 1000	53.5%	(224.365)	(219.988)
EE Amostra 1000	57.6%	(1,270.990)	(1,266.610)
EE Amostra 5000	58.0%	(97.379)	(93.002)
EE Amostra 10000	52.8%	(110.889)	(106.513)
EE Amostra 10000	53.0%	(111.536)	(107.160)
EE Amostra 15000	54.3%	24.554	28.931
EE Amostra 50000	54.2%	(110.275)	(105.898)
EE Amostra 100000	56.4%	26.264	30.640

Fig. 5.1: Tabela de Resultados

Pode se verificar que o desempenho da regressão logística, linear e da rede neural são muito

parecidos, diferencia apenas na segunda casa decimal. Com esses desempenhos próximos qualquer uma das técnicas poderia ser escolhidas. Para ajudar na escolha também é possível olhar outros critérios gráficos para se escolher o melhor modelo.

Quanto às estratégias evolutivas elas não tem um bom desempenho, a que teve melhor AIC foi a com amostra de 100.000 mil indivíduos. Neste caso a escolha do critério de avaliação dos melhores indivíduos não foi o melhor pois ele não leva em consideração outros fatores como a necessidade de que o modelo tenha uma ordem decrescente das taxas de maus cliente. Obviamente estas características podem ser adicionadas a trabalhos futuros.

Outro fator encontrado durante a simulação é que como está-se trabalhando com amostras, o algoritmo usado nas EE é muito sensível a amostra escolhida, tentando ao máximo se adaptar a aquela amostra e quando o modelo é testado na população ele tem um desempenho bem inferior. Neste sentido quanto maior a amostra mais próximo da população fica o resultado e melhor a ordenação do modelo também.

No caso das estratégias evolutivas também apresentamos as figuras 5.2 a 5.8 quem contém a evolução do treinamento a cada geração. O critério escolhido para avaliar os indivíduos da população de parâmetros em cada geração é o mesmo utilizado no exemplo de EE, a estatística de Kolmogorov-Smirnov(KS). Assim a cada geração os indivíduos foram avaliados de acordo com o seu desempenho em gerar a estatística para o conjuntos de dados da amostra e para todo o conjunto de dados. As figuras mostram a avaliação do melhor indivíduo a cada geração.

A figura 5.2 mostra que existe uma grande diferença entre os valores encontrados pelo treinamento da estratégia na amostra com apenas 300 indivíduos e na população com um milhão e meio de indivíduos. Conforme o aumento do tamanho da população está diferença vai sendo reduzido até ser muito próxima como nas figuras 5.7 e 5.8 das amostras maiores.

Nestas figuras também nota se que o algoritmo tem uma evolução rápida no início e depois fica preso em um ponto de mínimo local tendo certa dificuldade em sair apesar de empregado todos os tipos de mutação citados no texto.

É necessário analisar também a capacidade do modelo de ordenar os indivíduos de acordo com a

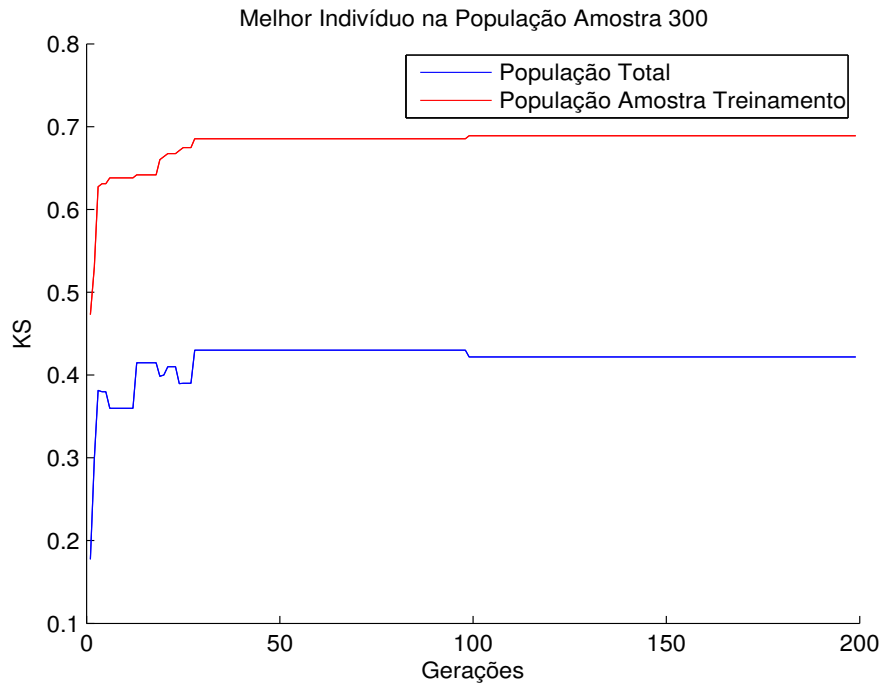

Fig. 5.2: Melhor Indivíduo por geração EE 300

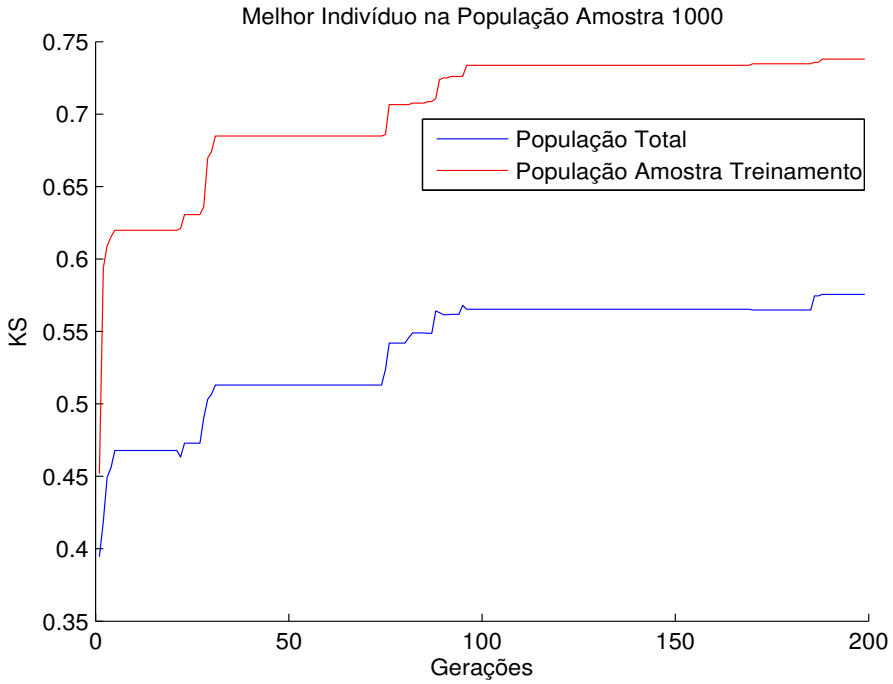

Fig. 5.3: Melhor Indivíduo por geração EE 1000

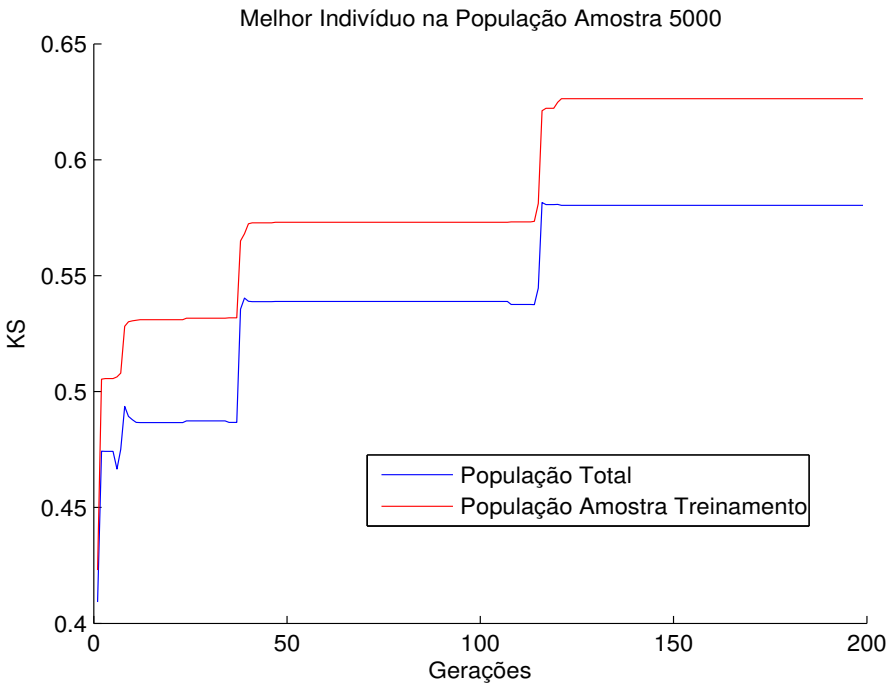

Fig. 5.4: Melhor Indivíduo por geração EE 5000

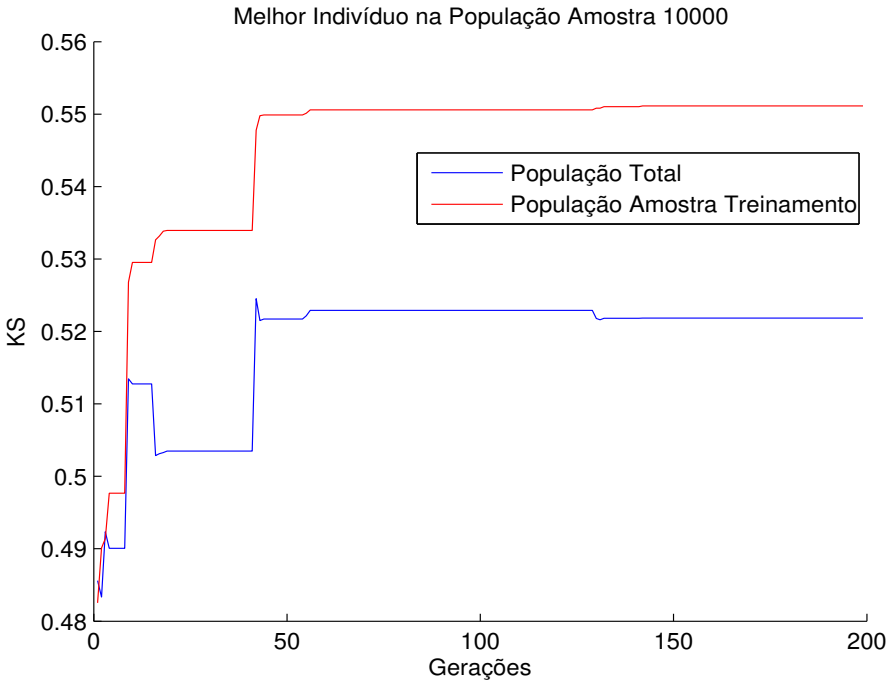

Fig. 5.5: Melhor Indivíduo por geração EE 10000

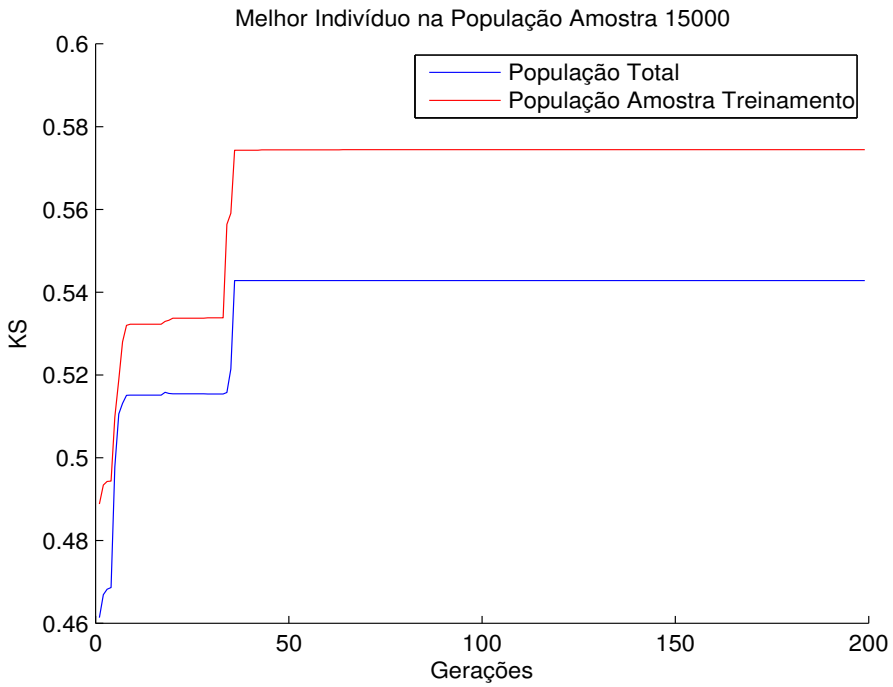

Fig. 5.6: Melhor Indivíduo por geração EE 15000

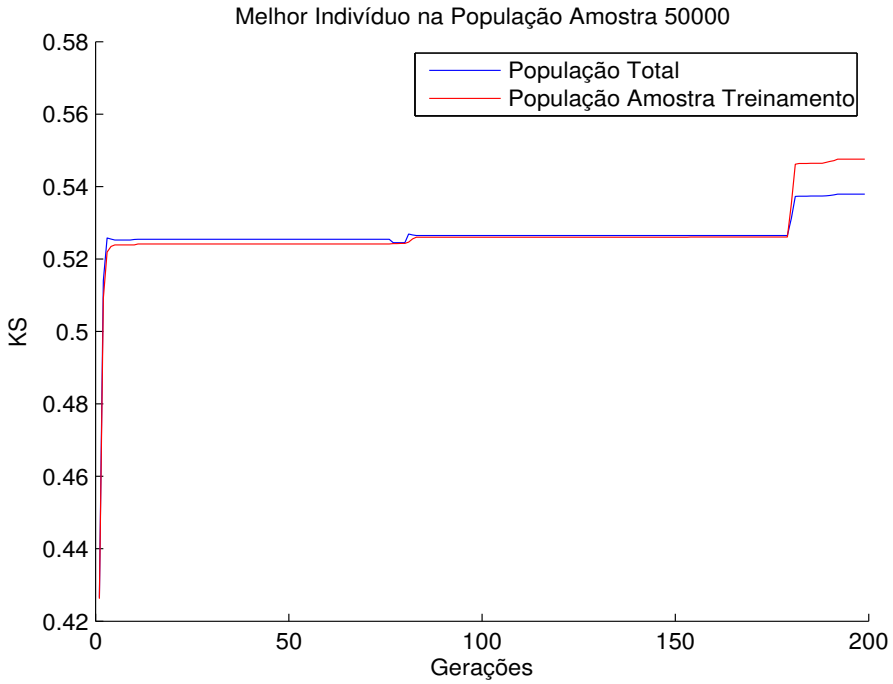

Fig. 5.7: Melhor Indivíduo por geração EE 50000

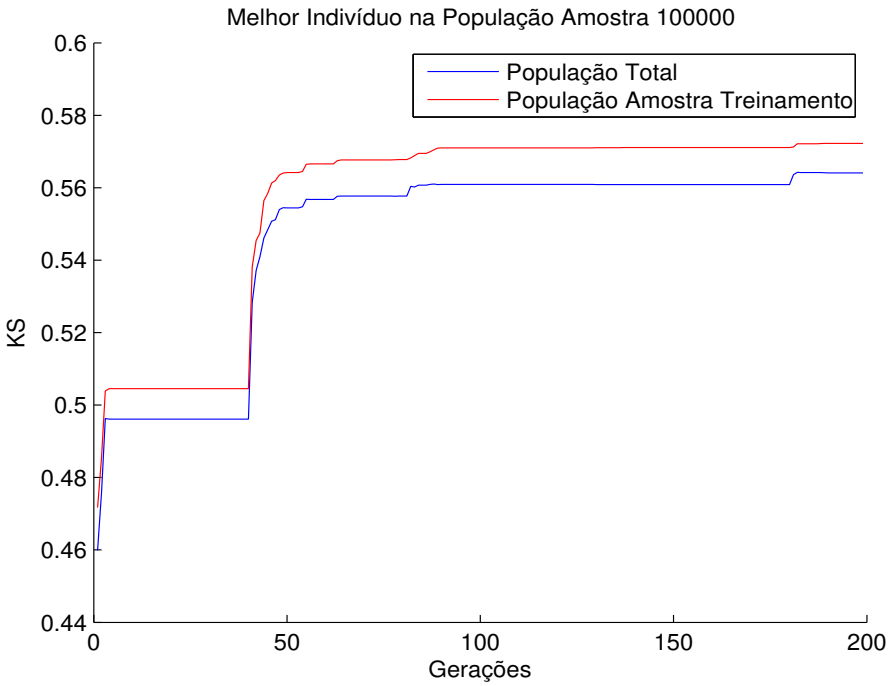

Fig. 5.8: Melhor Indivíduo por geração EE 100000

taxa de clientes não pagadores. As figuras 5.9, 5.10 e 5.11 mostram a pontuação dos modelos regressão logística, linear e redes neurais divididas em dez classes com tamanhos iguais (10% cada). Em todos os resultados estes três modelos parecem ter resultados muito parecidos, porém na figura 5.10 as taxas de maus clientes nas melhores faixas para regressão logística se mostram maiores que as do Regressão linear e da Rede Neural, indicando que os outros dois modelos podem ter um desempenho um pouco melhor.

Para metodologia de estratégia evolutiva também foram construídas as figuras 5.12 a 5.18 de taxa de maus clientes por faixa de pontuação do modelo (apesar do baixo desempenho dos indicadores), seguindo os mesmos critério de ter 10% dos clientes em cada faixa.

Nestas figuras pode se ver que os modelos apesar de terem um ks razoável não conseguem ordenar o população de forma apropriada. Nas figuras de modelos construídos com menor tamanho de amostra pode se perceber que essa capacidade é pior ainda, indicando que a técnica ou a métrica utilizada não foi capaz de produzir generalização suficiente que a partir de uma amostra pequena possa

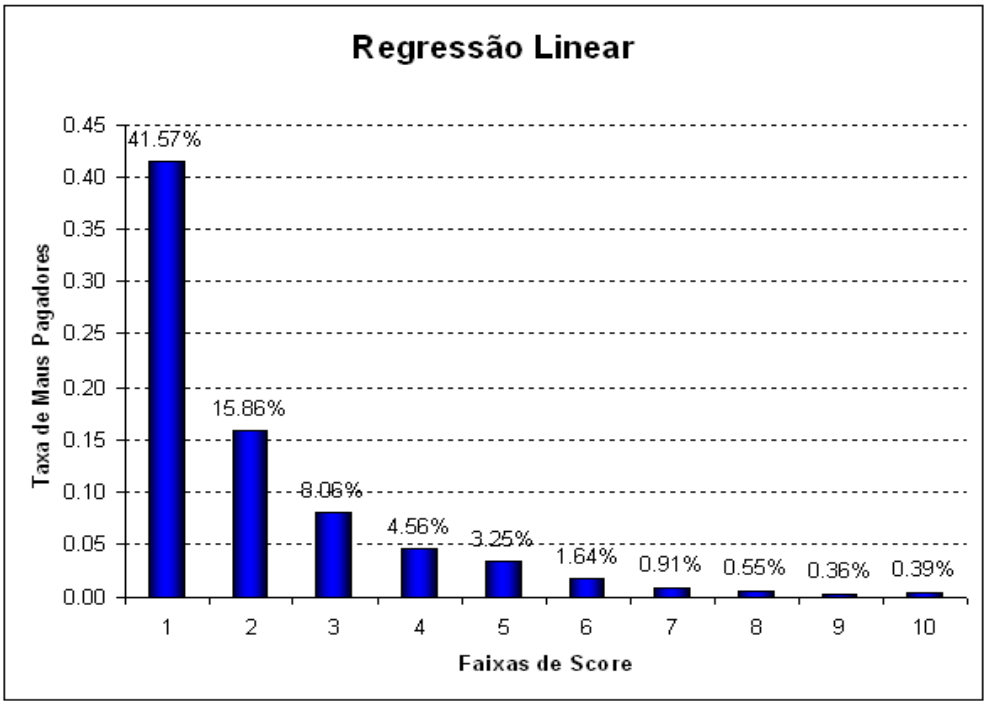

Fig. 5.9: Resultados

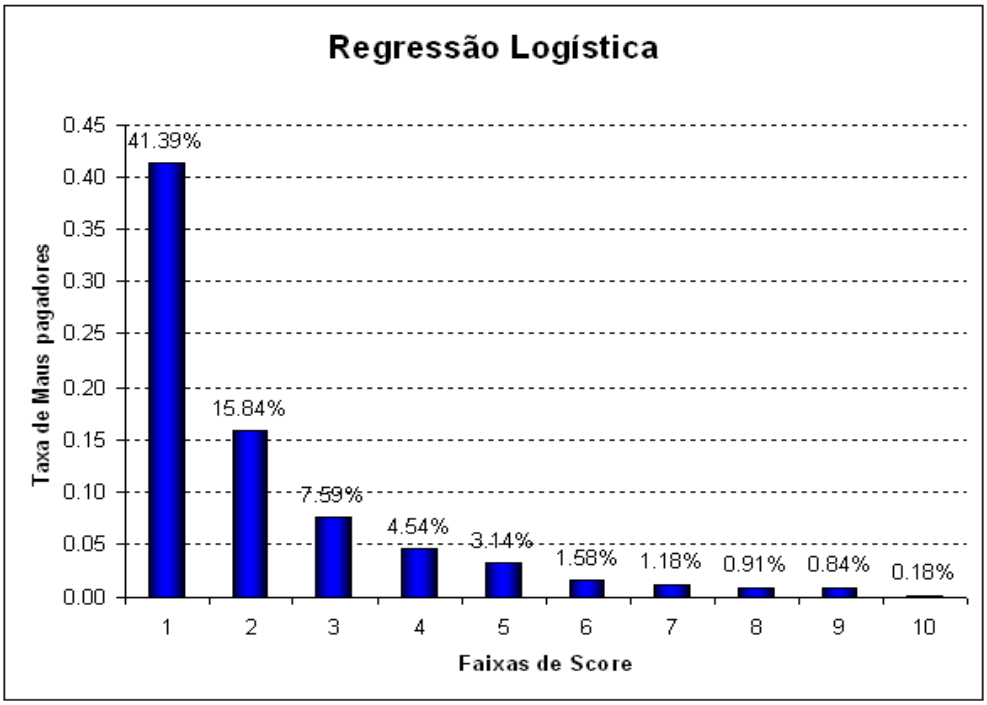

Fig. 5.10: Resultados

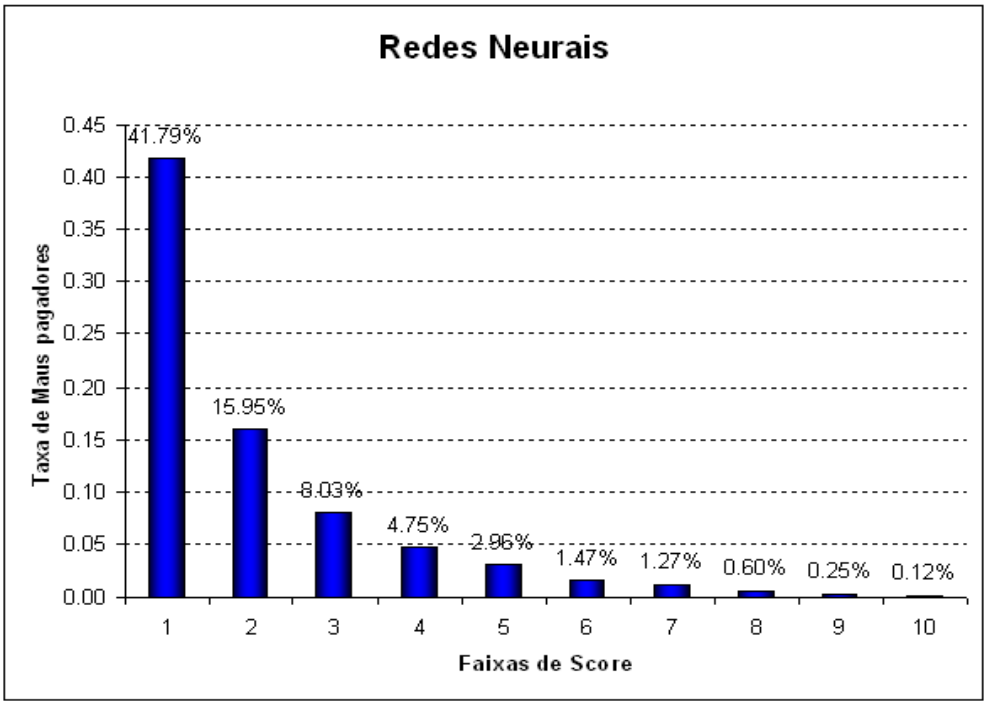

Fig. 5.11: Resultados

garantir para a população a estimativa de um modelo suficientemente genérico. Isso é ruim, pois os modelos de risco sempre são feitos com populações do passado e utilizados no presente, então uma boa capacidade de generalização é fundamental para este tipo de modelo.

Os resultados aqui mostrados para a técnica de estratégia evolutiva sugere que devem ser escolhidos outros critérios para de avaliar a evolução dos indivíduos a cada geração e conseqüentemente um critério de avaliação diferente. O critério utilizado não garante outras características essenciais aos modelos de crédito, como ordenação por exemplo. Aqui também fica a sugestão de que se deve tomar cuidado na escolha de outros critérios para serem utilizados para modelos de previsão de risco, critérios mal escolhidos podem gerar resultados insuficientes para a utilização do modelo.

Com a regulamentação por parte do Banco Central da utilização do acordo de Basiléia II os critérios para se utilizar os modelos devem ser mais rigorosos, assim como aconteceu nos países Europeus. Com critérios mais rigorosos novas técnicas sugeridas devem ser exaustivamente testadas para que não existam resultados inesperados na sua utilização.

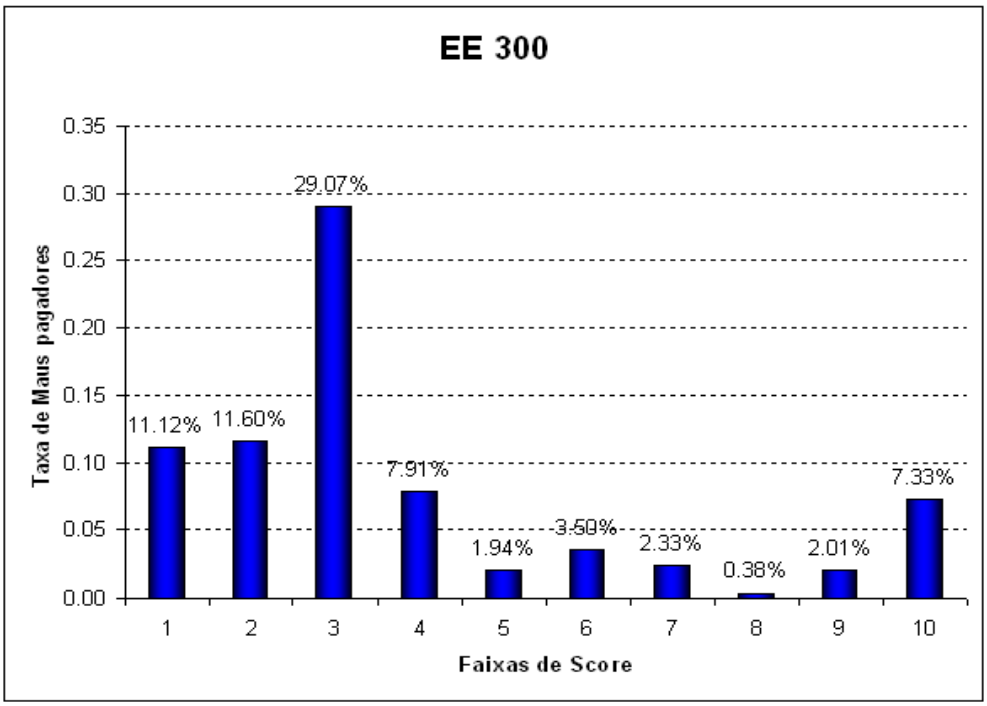

Fig. 5.12: Odernação pelas faixas de pontuação do modelo amostra 300

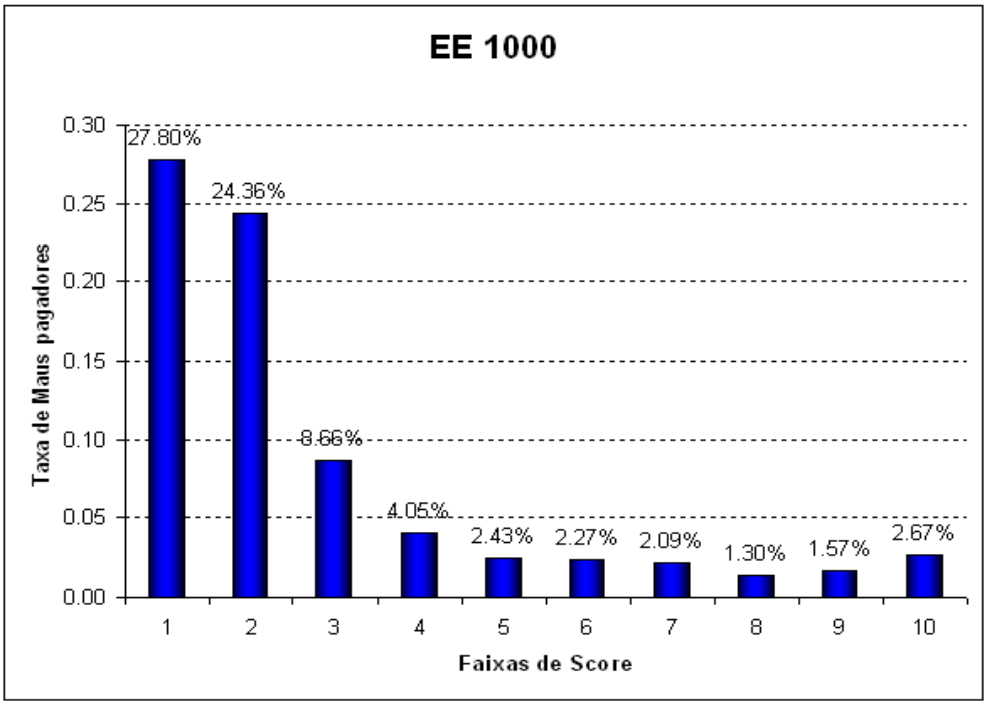

Fig. 5.13: Odernação pelas faixas de pontuação do modelo amostra 1000

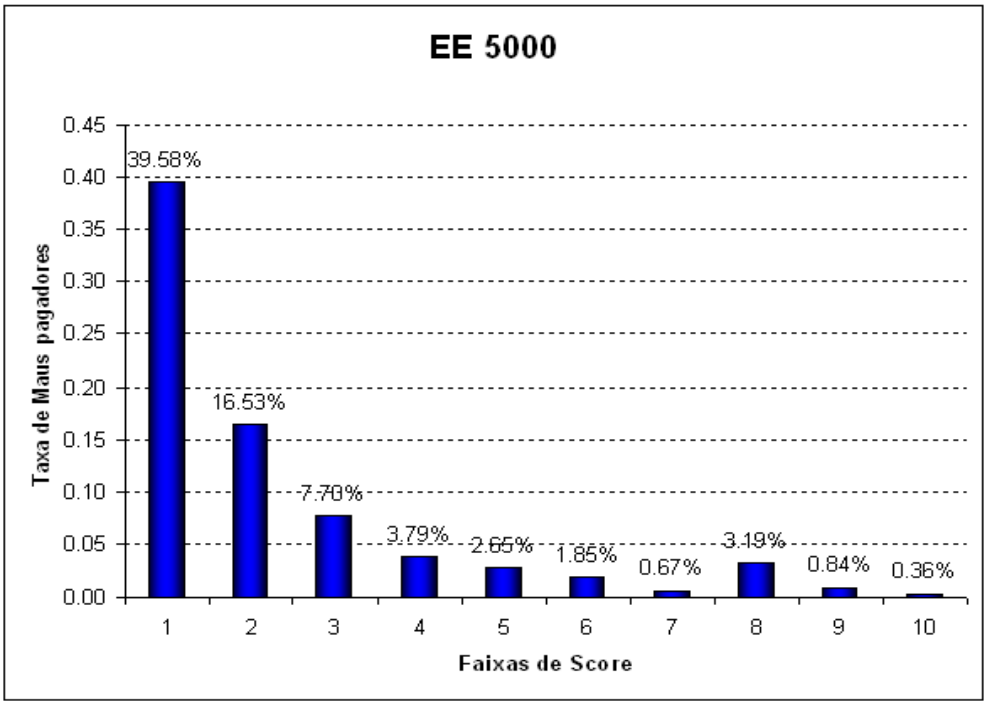

Fig. 5.14: Odernação pelas faixas de pontuação do modelo amostra 5000

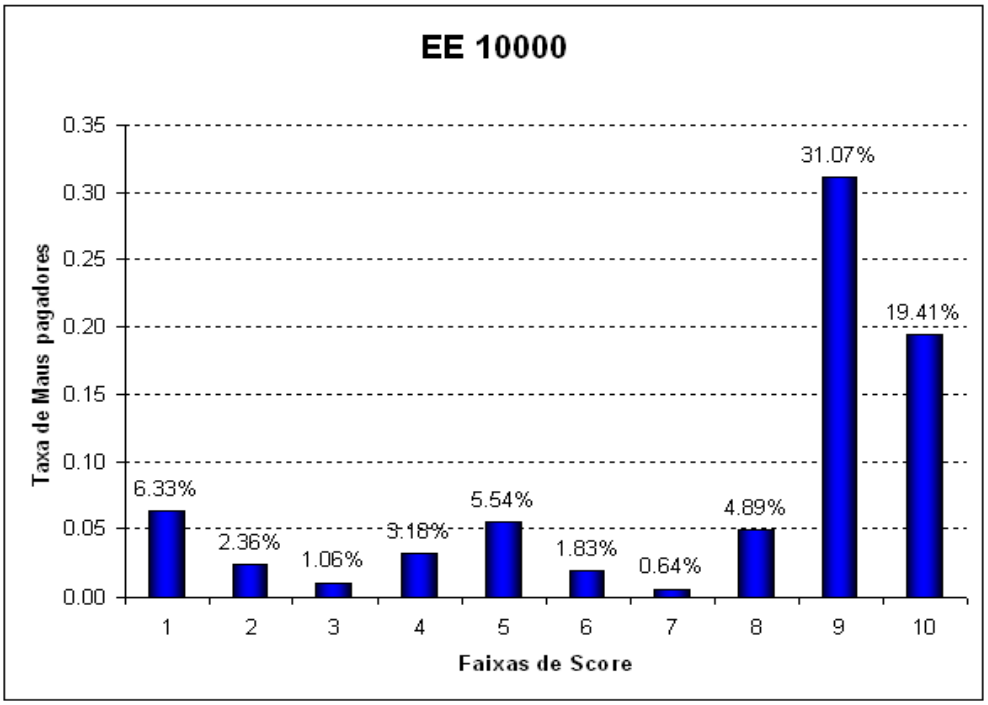

Fig. 5.15: Odernação pelas faixas de pontuação do modelo amostra 10000

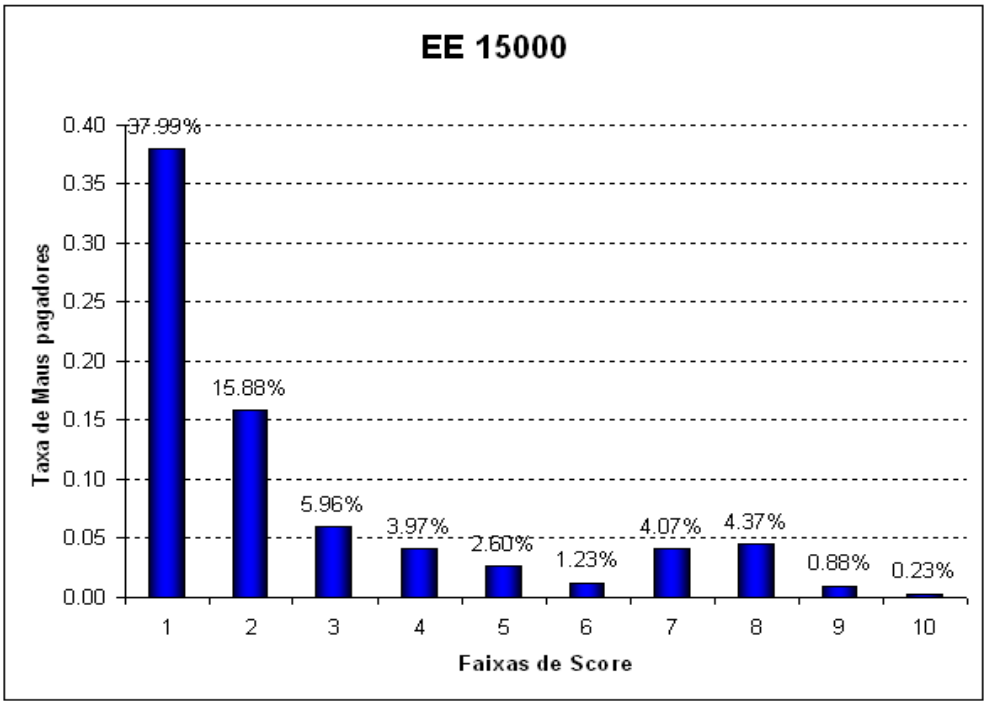

Fig. 5.16: Odernação pelas faixas de pontuação do modelo amostra 15000

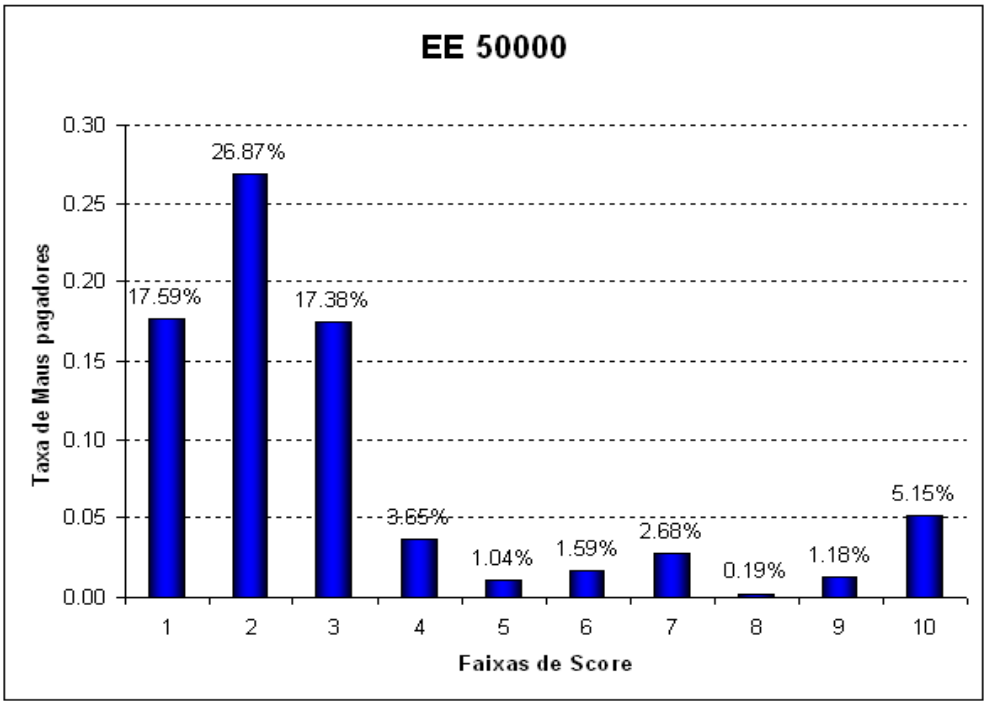

Fig. 5.17: Odernação pelas faixas de pontuação do modelo amostra 50000

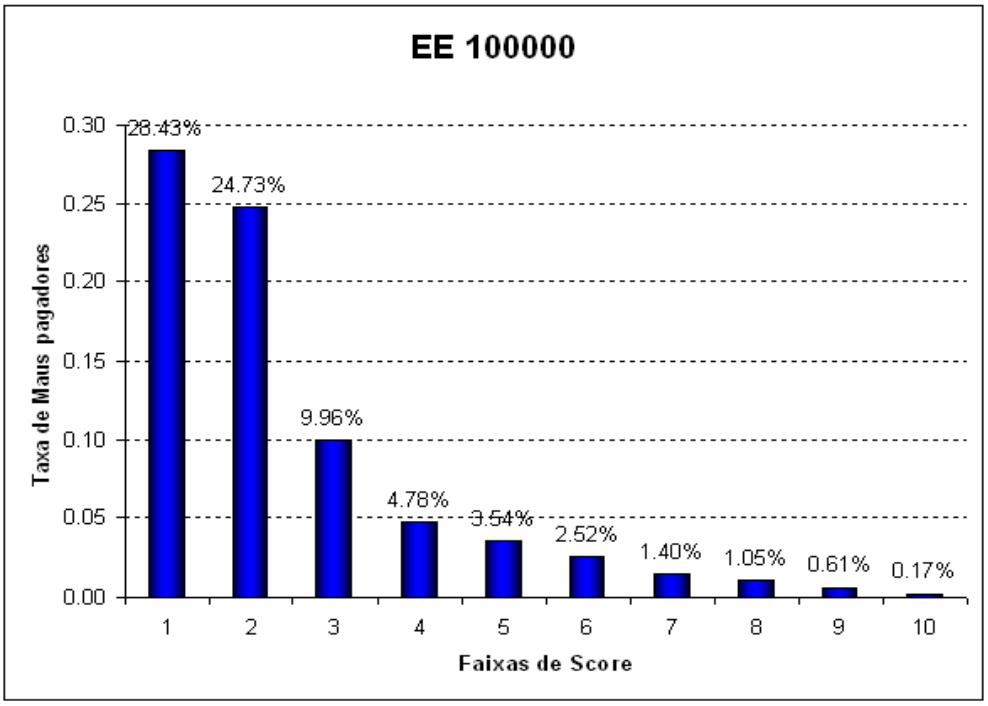

Fig. 5.18: Odernação pelas faixas de pontuação do modelo amostra 100000

Capítulo 6

Conclusões e Sugestões para Trabalhos Futuros

Neste trabalho foram apresentadas as principais técnicas atualmente utilizadas na construção de modelos de risco de crédito. Estes modelos tem sido amplamente utilizados pelas instituições financeiras para acompanhamento e controle dos riscos de suas carteiras de crédito. Nestas técnicas foram apresentados exemplos para entendimento da utilização e estimação dos parâmetros.

O entendimento do funcionamento das técnicas utilizadas para se obter o modelo é de extrema importância pois cada vez mais tem se exigido das instituições demonstrações de que os modelos funcionam, e isso passa por mostrar ao Banco Central o funcionamento das técnicas utilizadas e suas características. Neste sentido o trabalho contribui para reunir conceitos simples de utilização dos modelos e uma explicação de funcionamento das estimativas dos parâmetros.

Uma variação da técnica de *Ensemble* proposta tinha por objetivo construir modelos mais robustos e de melhores ou iguais em desempenho, porém pelos resultados apresentados as técnicas mais tradicionais ainda são as que tem melhor desempenho e capacidade de generalização.

As técnicas tradicionais de construção de modelos (Regressão logística, Linear e Redes Neurais) apresentaram os melhores resultados em termos de indicadores de KS, AIC e BIC e todos eles muito próximos. Os resultados apresentados mostram a capacidade do modelo de ordenar os clientes. A

regressão logística teve um resultado um pouco pior criando algumas classes nas melhores faixas com taxas maiores que a regressão linear e a rede MLP.

A Estratégia Evolutiva teve os piores resultados apresentados, tanto em indicadores quanto em questão de ordenação. Com o aumento do tamanho da amostra utilizada para construção dos modelos a técnica de EE começa a se aproximar das três principais, reduzindo também a diferença entre o treinamento da amostra e o resultado na população geral. Isso mostra que as simulações realizadas não apresentaram uma boa capacidade de generalização.

Em grande parte os resultados inferiores encontrados pela EE se devem ao indicador utilizado na avaliação da população de parâmetros, o KS. está estatística está apenas preocupada com o máximo da diferença entre a distribuição acumulada das populações de bom e mau, não levando em conta ordenação e a qualidade da previsão. Assim, uma escolha de outro indicador pode gerar resultados muitos diferentes, lembrando que umas das grandes vantagens desta técnica é que ela é aberta a escolha de indicadores.

No Brasil ainda não estão totalmente definidas as regras para adoção do Acordo de Basiléia II. Este acordo tem diretrizes principais, mas cada pais, através da instituição do governo que regula o sistema financeiro(Banco Central, no caso do Brasil que define as regras a serem adotadas). Após a crise econômica de 2008 os países desenvolvidos estão discutindo a regulamentação do setor algumas novas diretrizes devem ser criadas.

Como comentários pertinentes a este trabalho temos:

1. O modelo atualmente adotado pelo Brasil é muito rígido criando certas barreiras para o crescimento do crédito e conseqüentemente do país, porém no meio desta crise essas regras se mostraram capazes de ajudar o próprio setor a se manter e dar fôlego ao governo em criar medidas anti crise, como a redução do compulsório.

2. As regras adotadas no Brasil devem ser um meio termo entre o que Basiléia II direciona e as regras atuais. Um modelo entre a auto gestão dos riscos pregada por Basiléia e uma maior intervenção do Bacen nos recursos deixado pelos bancos como garantia. As instituições financeiras

devem ser acompanhadas de perto para evitar que corram riscos desnecessários com o objetivo de maiores lucros.

Como continuação deste trabalho tem se como sugestão os itens abaixo:

- Utilização de outras formas de avaliação dos indivíduos no algoritmo EE, como o próprio AIC e BIC.
- Utilização de modelos que permitam multi objetivos, por exemplo : Maximizar o KS e o AIC ou a ordenação.
- Utilização de métricas financeiras para avaliação dos modelos de EE, assim seria possível maximizar o retorno dos modelos.
- Avaliação dos modelos em uma outra amostra deslocada no tempo, pois os modelos são feitos com amostras de clientes no passado e aplicados no presente. Assim só depois de uma utilização do modelo é que é possível, na pratica, avaliar o seu desempenho.

Referências Bibliográficas

[1] Lyn C. Thomas. A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers. *International Journal of Forecasting*, 16, 149-172, 2000.

[2] Basiléia. *International Convergence of Capital Measurements and Capital Standards*. Bank International Settlements, www.bis.org, 2004.

[3] R. A. Fisher. *The Use of Multiple Measurements in Taxonomic Problems*. Annals of Eugenics, 7, 179-188 edition, 1936.

[4] David Durand. *Risk Elements in Consumer Installment Financing*. National Bureau of Economic Research, 1941.

[5] E. W. Forgy J. H. Myers. *The development o numerical credit evaluation systems*. Journal of American Statistics Association, 58, 799-806 edition, 1963.

[6] Ana Estela Antunes da Silva. *Uma Abordagem Multi-Objetivo e Multimodal Para Reconstrução de Árvores Filogenéticas*. Tese de Doutorado apresentada a Universidade Estadual de Campinas, 2006.

[7] W.S. MacCulloch and W. Pitts. *A Logical Calculus of the Ideas Immanent in Nervous Activity*. Bull. Math. Biophys, 5, pp. 115-133 edition, 1943.

[8] M.L. Minsky and S Papert. *Perceptrons*. MIT Press, 1969.

[9] Cybenko. *Approximation by Superpositions of a Sigmoidal Function*. Math. Control Signals Syst., 2 303-314 edition, 1989.

[10] J. Holland. *Adaptation in Natural an Artificial Systems*. University of Michigan Press, 1975.

[11] I RECHENBERG. *Cybernetic Solution Path Of An Experimental Problem Royal Aircraft Establishment Library Translation No 1122*. Farnborough, UK, 1965.

[12] H. P. SCHWEFEL. *Cybernetische Evolution als Strategie Der Experimentellen Forschung In Der Strömungstechnik*. Technical University of Berlin, 1965.

[13] S. Kullback and R. A. Leibler. *On Information and Sufficiency*. The Annals of MathematicalStatistics, The George Washington University and-washing, D. C, vol.22,no.1., pp.79-86 edition, 1951.

[14] L.K HANSEN and P. SALAMON. *Neural Network Ensembles*. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 12, no. io, edition, 1990.

[15] JAIN Anil K., Robert P. W. DUIN, and Jianchang MAO. *Statistical Pattern Recognition: A Review*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.22, pp.4-37 edition, 2000.

[16] V. Ganapathy Mohd. Hark Lye b. Abdullah. *Neural Network Ensemble For Financial Trend Prediction*. IEEE - TENCON, 2000.

[17] Scott Dellana David West and Jingxia Qian. *Neural Network Ensemble Strategies for Financial Decision Applications*. Computers & Operations Research, volume 32, 2543-2559 edition, 2005.

[18] Kagan Tumer and Joydeep Ghosh. *Error Correlation and Error Reduction in Ensemble Classifiers*. Connection Science, Special Issue On Combining Artificial Neural Networks: Ensemble Approaches, vol. 8, no. 3 & 4, pp 385-404 edition, 1996.

[19] Gustavo Henrique Araujo Pereira. *Modelo de Risco e Crédito de Clientes: Uma Aplicação a Dados Reais*. Dissertação apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo, 2004.

[20] F. O. de França, F. J. Von Zuben, and L.Nunes de Castro. *An Artificial Immune Network for Multimodal Function Optimization on Dynamic Environments*. Genetic And Evolutionary Computation Conference archive Proceedings of the 2005 conference on Genetic and evolutionary computation, pages: 289 - 296 edition, 2005.

[21] M S Bazaraa. *Linear Programming and Network Flows*. J. Willey and Sons, 1990.

[22] Kenneth P. Burnham. *Multimodel Inference: Understanding AIC and BIC in Model Selection*. ociological Methods & Research, vol. 33, no. 2, 261-304 edition, 2004.