

Área temática: Finanças

Análise de Risco de Crédito com o Uso de Modelos de Redes Neurais e Algoritmos Genéticos

AUTORES

MARIA APARECIDA GOUVÊA

Universidade de São Paulo
magouvea@usp.br

ERIC BACCONI GONÇALVES

Universidade de São Paulo
ebacconi@uol.com.br

Resumo

A tomada de decisões de concessão de crédito baseia-se fundamentalmente na avaliação do risco de inadimplência dos potenciais contratantes de produtos de crédito. Com o avanço tecnológico, modelos estatísticos foram desenvolvidos para dar sustentação à análise de solicitações de crédito, que há algumas décadas era realizada muitas vezes de forma qualitativa. Neste trabalho são apresentados, em um primeiro momento, conceitos de crédito e risco. Posteriormente, a partir de uma amostra de dados, fornecida por uma grande instituição financeira brasileira, estão desenvolvidos dois modelos, aplicando-se duas técnicas para a classificação de clientes: Redes Neurais e Algoritmos Genéticos. Em uma etapa final, são avaliadas e comparadas a qualidade e performance dos modelos desenvolvidos, onde é apontado qual o modelo que melhor se ajusta aos dados. Os resultados obtidos pelos modelos de rede neural e algoritmo genético são satisfatórios, sendo o primeiro modelo superior. Este trabalho ilustra os procedimentos a serem adotados por uma empresa para identificar o melhor modelo de concessão de crédito, a partir do qual é possível direcionar a estratégia da instituição, podendo-se aumentar a eficiência do seu negócio.

Palavras-chave: risco de crédito, rede neural, algoritmo genético.

Abstract

The taking of decisions of concession of credit is based basically on the evaluation of the risk of insolvency of the contracting potentials of credit products. With the technological advance, statistical models had been developed to give sustentation to the analysis of credit requests, that it has some decades was carried through many times of qualitative form. In this work, at a first moment, concepts of credit and risk are presented. Later, from a data sample, that was supplied by a great Brazilian financial institution, two models are developed, applying two techniques for the classification of customers: Neural Networks and Genetic Algorithms. In a final stage, the quality and the performance of these models are evaluated and compared and the model that best fits the data is pointed. The results obtained by neural network and genetic algorithm are good, but the first one is better. This work illustrates the procedures to be adopted by a company to identify the best model of concession of credit, from which it is possible to direct the strategy of the institution, being able itself to increase the efficiency of its business.

Keywords: credit risk, neural network, genetic algorithm.

1. Introdução

Com a estabilidade da moeda, atingida no Plano Real em 1994, os empréstimos financeiros passaram a ser um bom negócio para os bancos que já não obtinham os vultuosos lucros que provinham da desvalorização da moeda (ROSA, 2000, p. 1). Percebeu-se a necessidade de se aumentarem as alternativas de investimento para substituir a rentabilidade do período de inflação. Desde então as instituições têm se preocupado em aumentar suas carteiras de crédito. Entretanto, o empréstimo não poderia ser oferecido indiscriminadamente a todos aqueles clientes que o solicitassem, sendo necessárias formas de avaliar o candidato ao crédito.

Há alguns anos ao fazer uma solicitação de crédito, o cliente preenchia uma proposta que seria avaliada por um ou mais analistas que apresentavam um parecer em relação ao pedido (SEMOLINI, 2002, p. 103). Apesar de eficaz, este processo era lento, por não permitir a análise de muitos pedidos. Os modelos de análise para concessão de crédito começaram a ser adotados nas instituições financeiras com o objetivo de acelerar a avaliação das propostas.

Os modelos de análise para concessão de crédito, conhecidos como modelos de *credit scoring* baseiam-se em dados históricos da base de clientes existentes para avaliar se um futuro cliente terá mais chances de ser bom ou mau pagador. Os modelos de *credit scoring* são implantados nos sistemas das instituições, permitindo que a avaliação de crédito seja on-line.

2. Objetivos do estudo

Pretende-se com este trabalho:

- Desenvolver dois modelos de *credit scoring*, mediante o uso de duas técnicas estatísticas: Redes Neurais e Algoritmos Genéticos;
- Comparar os modelos em termos de indicadores de qualidade de ajuste e previsão.

3. Fundamentação teórica

Nesse capítulo serão apresentados conceitos teóricos que darão sustentação ao desenvolvimento do tema deste trabalho.

3.1. Crédito ao Consumidor

A expressão crédito ao consumidor pode ser entendida como uma forma de comércio onde uma pessoa física obtém dinheiro, bens ou serviços e compromete-se a pagar por isso futuramente, acrescendo ao valor original um prêmio (juros) (SANTOS, 2000, p. 15).

Atualmente, grandes varejistas impulsionam suas vendas, fornecendo crédito. Empresas automobilísticas, bancos e outros segmentos utilizam as linhas de crédito ao consumidor como uma alternativa a mais para obter lucros.

Entretanto, existe um fator associado ao crédito ao consumidor que é decisivo na decisão de disponibilizar ou não o crédito: o risco.

3.2. Risco de Crédito

O risco de crédito é a mais antiga forma de risco no mercado financeiro (FIGUEIREDO, 2001, p. 9). É consequência de uma transação financeira contratada entre um fornecedor de fundos (doador do crédito) e um usuário (tomador do crédito). Antes de qualquer sofisticação, produto da engenharia financeira, o puro ato de emprestar uma quantia a alguém traz embutida em si a probabilidade de ela não ser recebida, a incerteza em relação ao retorno. Isto é, na essência, o risco de crédito, e que se pode definir como: o risco de uma contraparte, em um acordo de concessão de crédito, não honrar seu compromisso.

Segundo Caouette *et al* (2000, p. 1), “se crédito pode ser definido como a expectativa de recebimento de uma soma em dinheiro em um prazo determinado, então Risco de Crédito é a chance que esta expectativa não se concretize”.

A atividade de concessão de crédito é função básica dos bancos; o risco de crédito toma papel relevante na composição dos riscos de uma instituição e pode ser encontrado tanto em operações onde existe liberação de dinheiro para os clientes como naquelas onde há apenas a possibilidade do uso, os limites pré-concedidos. Os principais tipos de crédito de um banco são: empréstimos, financiamentos, descontos de títulos, adiantamento a depositantes, adiantamento de câmbio, operações de arrendamento mercantil (*leasing*), avais e fianças etc. Nessas operações, o risco pode se apresentar sob diversas formas; conhecê-las conceitualmente ajuda a direcionar o gerenciamento e a mitigação.

No universo do crédito ao consumidor, a promessa de pagamento futuro envolve a idéia de risco. Todo crédito ao consumidor envolve risco, pois nunca existe a certeza do pagamento (LEWIS, 1992, p. 2). Cabe à análise de crédito estimar o risco envolvido.

O risco máximo que a instituição pode aceitar depende da política adotada pela empresa. O risco apresentado pelo solicitante é de extrema importância no processo de concessão de crédito, devendo ser considerados vários quesitos na sua avaliação.

3.3. Avaliação do risco de crédito

O ponto principal para a concessão de crédito é a avaliação do risco. Se o risco for mal avaliado a empresa certamente irá perder dinheiro, quer seja pelo aceite de clientes que irão gerar prejuízos ao negócio, quer seja pela recusa de clientes bons que gerariam lucros ao negócio. Empresas que têm uma avaliação melhor que as concorrentes na concessão de crédito levam vantagem em relação às demais, por ficarem menos vulneráveis às conseqüências decorrentes de decisões equivocadas no fornecimento de crédito.

A avaliação do risco de um potencial cliente pode ser feita de duas maneiras:

1. Por julgamento, uma forma mais subjetiva que envolve uma análise mais qualitativa;
2. Por classificação do tomador via modelos de avaliação, envolvendo uma análise mais quantitativa.

Atualmente, praticamente todas as grandes empresas que trabalham com concessão de crédito utilizam as duas formas combinadas.

Na avaliação do risco de crédito por meio de classificação do tomador é que são utilizados os modelos chamados *credit scoring*, que permitem uma mensuração do risco do tomador de crédito, auxiliando na tomada de decisão (concessão ou não do crédito).

3.4. Modelos de *credit scoring*

Há sete passos a serem seguidos para se construir um modelo de *credit scoring*, a saber:

1. Levantamento de uma base histórica de clientes

A suposição básica para se construir um modelo de avaliação de crédito é que os clientes têm o mesmo padrão de comportamento ao longo do tempo; portanto, com base em informações passadas são construídos os modelos. A disponibilidade e qualidade da base de dados são fundamentais para o sucesso do modelo (TREVISANI *et al*, 2004).

2. Classificação dos clientes de acordo com o padrão de comportamento e definição da variável resposta

Além de clientes bons e maus, também existem os clientes excluídos, que possuem características peculiares e não devem ser considerados (por exemplo, trabalha na instituição) e os clientes indeterminados, que estão na fronteira entre serem bons ou maus, não existindo uma posição clara para eles. Na prática, as instituições consideram apenas os clientes bons e maus para fazer o modelo devido à maior facilidade de trabalhar com modelos de resposta binária. Esta tendência também é observada nos trabalhos acadêmicos (ROSA, 2000; OHTOSHI, 2003; SEMOLINI, 2002; HAND; HENLEY, 1997; entre outros).

3. Seleção de amostra aleatória representativa da base histórica

É importante que as amostras de bons e maus clientes tenham o mesmo tamanho para se evitar qualquer possível viés devido à diferença de tamanhos. Não existe um número fixo para a amostra; entretanto, Lewis (1992, p. 31) sugere uma amostra de 1.500 clientes bons e 1.500 clientes maus para serem propiciados resultados robustos. Costuma-se trabalhar com três amostras: para construção do modelo, para validação do modelo e para teste do modelo.

4. Análise descritiva e preparação dos dados

Consiste em analisar segundo critérios estatísticos cada variável a ser utilizada no modelo.

5. Escolha e aplicação das técnicas a serem utilizadas para a construção do modelo

Neste trabalho serão utilizadas Redes Neurais e Algoritmos Genéticos. Hand e Henley (1997) destacam ainda Regressão Logística, Análise Discriminante, Regressão Linear e Árvores de Decisão, como métodos utilizados na prática. Recentemente alguns estudiosos também têm utilizado Análise de Sobrevivência (HARRISON; ANSELL, 2002; ANDREEVA, 2003). Não existe um método claramente melhor que os demais, tudo dependendo de como a técnica escolhida se ajusta aos dados.

6. Definição dos critérios de comparação dos modelos

Será definida a medida de comparação dos modelos, normalmente pelo índice de acertos e a estatística de Kolmogorov-Smirnov (KS).

7. Seleção e Implantação do melhor modelo

Por meio dos critérios previamente definidos, o melhor modelo é escolhido. Com isso deve-se programar a implantação do modelo. A instituição deve adequar seus sistemas para receber o algoritmo final e programar a utilização do mesmo junto às demais áreas envolvidas.

3.5. Redes neurais artificiais

Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento por intermédio de experiências.

3.5.1. Conceitos

Um modelo de rede neural artificial processa certas características e produz respostas similarmente ao cérebro humano. Redes neurais artificiais são desenvolvidas por meio de modelos matemáticos, onde as seguintes suposições são feitas (FAUSETT, 1994, p. 3):

1. O processamento das informações ocorre dentro dos chamados neurônios;
2. Os estímulos são transmitidos pelos neurônios por meio de conexões;
3. Cada conexão tem associada a si um peso, que, numa rede neural padrão, multiplica-se ao estímulo recebido;
4. Cada neurônio contribui para a função de ativação (geralmente não linear) para determinar o estímulo de saída (resposta da rede).

O modelo pioneiro de McCulloch e Pitts de 1943, para uma unidade de processamento (neurônio), pode ser resumido em:

- Sinais são apresentados à entrada;
- Cada sinal é multiplicado por um peso que indica sua influência na saída da unidade;
- É feita a soma ponderada dos sinais que produz um nível de atividade.

Se este nível excede um limite, a unidade produz uma saída.

No esquema, têm-se p sinais de entrada X_1, X_2, \dots, X_p e pesos correspondentes W_1, W_2, \dots, W_p e seja k o limite.

Neste modelo o nível de atividade é dado por:

$$a = \sum_{i=1}^p W_i X_i$$

A saída y é dada por: $y = 1$, se $a \geq k$; $y = 0$, se $a < k$

Na definição de um modelo de redes neurais três características devem ser observadas: a forma que a rede tem, chamada arquitetura; o método para determinação dos pesos, chamado algoritmo de aprendizado; e a função de ativação.

3.5.1.1 Arquitetura

Arquitetura refere-se ao formato da rede. Toda rede é dividida em camadas, usualmente classificadas em três grupos:

- Camada de Entrada: onde os padrões são apresentados à rede;
- Camadas Intermediárias ou Ocultas: onde é feita a maior parte do processamento, por meio das conexões ponderadas; podem ser consideradas como extratoras de características;
- Camada de Saída: onde o resultado final é concluído e apresentado.

Existem três tipos principais de arquitetura (HAYKIN, 1999, p. 46-48): redes *feedforward* com uma única camada, redes *feedforward* com múltiplas camadas, e redes recorrentes.

Redes *feedforward* com uma única camada: são o caso mais simples de rede, existindo apenas uma camada de entrada e uma camada de saída. As redes são alimentadas adiante, ou seja, apenas a camada de entrada fornece informações para a camada de saída. Algumas das redes que utilizam essa arquitetura são: Rede de Hebb, *perceptron*, ADALINE, entre outras.

Redes *feedforward* com múltiplas camadas: são aquelas que possuem uma ou mais camadas intermediárias. A saída de cada camada é utilizada como entrada para a próxima camada. Da mesma forma que a arquitetura anterior, este tipo de rede caracteriza-se apenas por alimentação adiante. As redes *multilayer perceptron* (MLP), MADALINE e de função de base radial são algumas das redes que utilizam esta arquitetura.

Redes Recorrentes: neste tipo de rede, a camada de saída possui ao menos uma ligação que realimenta a rede. As redes chamadas de BAM (*Bidirecional Associative Memory*) e ART1 e ART2 (*Adaptive Resonance Theory*) são redes recorrentes.

3.5.1.2. Processo de Aprendizado

A propriedade mais importante das redes neurais é a habilidade de “aprender” de acordo com o ambiente e com isso melhorar seu desempenho (CASTRO JR., 2003, p. 92). Esse aprendizado é realizado, ajustando-se os pesos por meio de um processo iterativo. O objetivo do processo é a obtenção de um algoritmo de aprendizado que permita uma solução generalizada para certa classe de problema.

Denomina-se algoritmo de aprendizado um conjunto de regras bem definidas para a solução de um problema de aprendizado. Existem muitos tipos de algoritmos específicos para determinados modelos de redes neurais. Estes algoritmos diferem entre si principalmente pelo modo como os pesos são modificados.

Existem basicamente três tipos de aprendizado:

1. Aprendizado Supervisionado: neste tipo de aprendizado, é indicada para a rede qual a resposta esperada. Trata-se do exemplo deste trabalho onde a priori já se sabe se o cliente é bom ou mau;

2. Aprendizado Não Supervisionado: neste tipo de aprendizado, a rede deve basear-se apenas nos estímulos recebidos; a rede deve aprender a agrupar os estímulos;
3. Aprendizado por Reforço: neste tipo de aprendizado, o comportamento da rede é avaliado por um crítico externo.

Cada tipo de aprendizado possui vários algoritmos possíveis de serem utilizados.

3.5.1.3. Funções de Ativação

Cada neurônio contribui para o estímulo de saída. A função de ativação desempenha o papel de restringir a amplitude de saída de um neurônio, em geral [0,1] ou [-1,1] (HAYKIN, 1999, p. 37). Alguns exemplos de funções de ativação utilizadas são:

- Função Limiar: $f(x) = 1$ se $x < k$ e 0, caso contrário
- Função Logística: $f(x) = \frac{1}{1 + e^{(-\alpha x)}}$
- Função Tangente Hiperbólica: $f(x) = \tanh(x)$

3.6. Algoritmos genéticos

Os algoritmos genéticos são uma família de modelos computacionais inspirados na evolução, que incorporam uma solução potencial para um problema específico numa estrutura semelhante à de um cromossomo e aplicam operadores de seleção, cruzamento (*cross-over*) e mutação a essas estruturas de forma a preservar informações críticas relativas à solução do problema.

3.6.1. Conceitos

O algoritmo é composto por uma população, que é representada por cromossomos, que nada mais são do que diversas soluções possíveis para o problema proposto. As soluções que são selecionadas para dar forma a soluções novas (a partir de um cruzamento) são selecionadas de acordo com a aptidão (*fitness*) dos cromossomos pais. Assim, quanto mais apropriado é o cromossomo, maior a possibilidade de ele se reproduzir. Esse processo é repetido até que a regra de parada seja satisfeita, ou seja, encontrar uma solução muito próxima da desejada. Todo algoritmo genético passa pelas seguintes fases:

Início: primeiramente é gerada uma população formada por um conjunto aleatório de indivíduos (cromossomos) que podem ser vistos como possíveis soluções do problema.

Função de Aptidão (*Fitness*): uma função de aptidão é definida para avaliar a “qualidade” de cada um dos cromossomos.

Seleção: de acordo com os resultados da função de aptidão, uma porcentagem dos mais adaptados é mantida, enquanto os outros são descartados (Darwinismo).

Cruzamento (*Cross-Over*): escolhem-se dois pais e baseando-se neles é gerado um “filho” baseado num critério específico de cruzamento. O mesmo critério é efetuado com o outro cromossomo, e o material dos dois cromossomos é trocado. Se nenhum cruzamento for executado, a prole é uma cópia exata dos pais.

Mutação: atribui-se à população um percentual de mutação. A mutação é a alteração de algum dos genes do cromossomo. O intuito da mutação é evitar que a população convirja para um máximo local. Assim, caso esta convergência ocorra, a mutação garante que a população irá “saltar” o ponto de mínimo local, tentando alcançar outros pontos de máximo.

Verificação do critério de parada: criada uma nova geração, verifica-se o critério de parada pré-estabelecido e retorna-se para a fase da função de aptidão, caso este critério não esteja satisfeito.

3.7. Critérios de avaliação de performance

Para avaliar a performance do modelo foram selecionadas duas amostras, uma de validação e outra de teste de mesmo tamanho (3000 clientes considerados bons e 3000 considerados maus para cada uma das duas). Os critérios que serão utilizados são apresentados a seguir.

3.7.1. Taxa de Acerto

Mede-se a taxa de acerto por meio da divisão do total de clientes classificados corretamente, pela quantidade de clientes que fizeram parte do modelo. Idem para a taxa de acertos dos bons e maus clientes.

Em algumas situações, é mais importante identificar um cliente bom do que um cliente mau (ou vice-versa); nesses casos, é comum dar-se um peso para a taxa de acertos mais adequada e calcular-se uma média ponderada da taxa de acertos.

Neste trabalho, como não há informações a priori sobre o que seria mais atrativo para a instituição financeira, utilizar-se-á o produto entre as taxas de acerto de bons e maus clientes como um indicador de acerto (Ia). Esse indicador privilegiará os modelos que tenham altos índices de acerto para os dois tipos de clientes. Quanto maior o indicador, melhor o modelo.

3.7.2. Teste de Kolmogorov-Smirnov

O outro critério bastante utilizado na prática (PICININI *et al*, 2003; OOGHE *et al*, 2001; Pereira, 2004) a ser abordado neste trabalho é o teste de Kolmogorov-Smirnov (KS).

O teste de KS é uma técnica não paramétrica para determinar se duas amostras foram extraídas da mesma população (SIEGEL, 1975, p. 144). Este teste se baseia na distribuição acumulada dos escores dos clientes considerados como bons e maus.

Para se verificar se as amostras têm a mesma distribuição, há tabelas que são consultadas de acordo com o nível de significância e tamanho da amostra (ver SIEGEL, 1975, p. 309-310). Neste trabalho, como as amostras são grandes, a tendência é que todos os modelos rejeitem a hipótese de igualdade nas distribuições. Será considerado melhor modelo àquele que possuir o maior valor no teste, pois este resultado indica uma separação maior entre bons e maus.

4. Aspectos metodológicos

4.1. Descrição do estudo

Uma instituição financeira deseja conceder empréstimos a seus clientes e, para isso, necessita de uma ferramenta que avalie o grau de risco associado a cada empréstimo para auxiliar o processo de tomada de decisão. Para viabilizar este projeto, foram disponibilizadas informações do histórico de clientes que contrataram um crédito pessoal.

4.2. O produto de crédito em estudo

O produto em estudo é o crédito pessoal. O crédito pessoal é uma operação de crédito ao consumidor rápida e prática. Não é preciso declarar a finalidade que será dada ao empréstimo, o qual é concedido de acordo com a capacidade de crédito do solicitante.

Outra característica do produto é a não exigência de bens como garantia de pagamento.

Considerou-se a modalidade com juros pré-fixados com prazos variando de 1 a 12 meses.

4.3. Os dados

Para a realização do estudo foram selecionados aleatoriamente, a partir do universo de clientes do banco em estudo, 10.000 contratos de crédito tidos como bons e 10.000 considerados maus, realizados no período de agosto de 2002 a fevereiro de 2003, sendo que todos estes contratos já venceram. Trata-se de uma base de dados histórica com informações

mensais de utilização do produto. A partir desta estrutura pode-se acompanhar o andamento do contrato e precisar em que momento o cliente deixou de pagar uma ou mais parcelas.

No trabalho a amostra é dividida em três sub-amostras: uma para construção do modelo, 8.000 dados (sendo 4.000 bons e 4.000 maus); a segunda para validação do modelo construído, 6.000 dados (sendo 3.000 bons e 3.000 maus) e a terceira também com 6.000 (com a mesma divisão equitativa) para testar o modelo obtido. Cada sub-amostra tem a sua função específica (ARMINGER *et al*, 1997, p. 294). A sub-amostra de construção do modelo é usada para estimação dos parâmetros do modelo, a sub-amostra de teste tem como função verificar o poder de predição dos modelos construídos, e a sub-amostra de validação, particularmente numa rede neural, tem a função de validar os parâmetros, evitando o “superajuste” (*overfitting*) do modelo. No modelo de algoritmo genético a amostra de validação terá o mesmo papel da amostra de teste, ou seja, avaliar a predição do modelo.

4.4. As variáveis

Foram consideradas as seguintes variáveis cadastrais (perfil do cliente) e de utilização e restrição (restrições de crédito e apontamentos sobre outras operações de crédito do cliente): sexo, estado civil, fone residencial, fone comercial, tempo no emprego atual, salário do cliente, quantidade de parcelas a serem quitadas, primeira aquisição, tempo na residência atual, valor da parcela, valor total do empréstimo, tipo de crédito, idade, CEP residencial, CEP comercial, código da profissão, nome da profissão, salário do cônjuge.

Para o produto em estudo, clientes com 60 ou mais dias de atraso foram considerados Maus (inadimplentes) e clientes com no máximo 20 dias de atraso como Bons. Os clientes denominados Indeterminados representam um grupo cujo comportamento de crédito não é claro para indicá-los como bons ou maus pagadores. Na prática, estes clientes são analisados separadamente pelo analista de crédito com base em análise qualitativa.

5. Aplicação

Nesta seção serão vistos os métodos de tratamento das variáveis, a aplicação das duas técnicas estudadas e os seus resultados, comparando-se o desempenho destas. Para a análise descritiva e categorização dos dados foi utilizado o *software* SPSS for Windows v.11.0, para a seleção das amostras e aplicação da rede neural foi utilizado o *software* Enterprise Miner v.4.1; para o algoritmo genético foi utilizado um programa desenvolvido pelos autores em *Visual Basic*.

5.1. Tratamento das variáveis

Inicialmente, as variáveis quantitativas foram categorizadas.

Para a categorização das variáveis contínuas, inicialmente foram identificados os decis destas variáveis. Partindo-se dos decis, o passo seguinte foi analisá-los de acordo com a variável resposta. Foi calculada a distribuição de bons e maus clientes por decil e em seguida calculada a razão entre bons e maus, o chamado risco relativo (RR).

Grupos que apresentaram risco relativo (RR) semelhante foram reagrupados a fim de se diminuir o número de categorias por variável.

Também para as variáveis qualitativas foi calculado o risco relativo para se diminuir o número de categorias, quando possível. Conforme Pereira (2004, p. 49), existem duas razões para se fazer uma “nova categorização” das variáveis qualitativas. O primeiro é evitar categorias com um número muito pequeno de observações, o que pode levar a estimativas pouco robustas dos parâmetros. O segundo é a eliminação de parâmetros do modelo; se duas categorias apresentam risco próximo, é razoável agrupá-las numa única classe.

O RR, além de auxiliar no agrupamento das categorias, ajuda a entender se a categoria em questão está mais ligada a clientes bons ou ruins. Esse método de agrupamento de categorias é explicado por Hand e Henley (1997, p. 527).

Ao trabalhar-se com as variáveis disponibilizadas, os seguintes cuidados foram tomados:

- As variáveis sexo, primeira aquisição e tipo de crédito não foram recodificadas por já se tratarem de variáveis binárias;
- A variável profissão foi agrupada conforme a similaridade da natureza das ocupações;
- Telefone comercial e residencial ficaram na forma binária como posse ou não;
- As variáveis CEP comercial e CEP residencial foram agrupadas de acordo com os três primeiros dígitos; depois, foi calculado o risco relativo de cada faixa e houve o reagrupamento de acordo com risco relativo semelhante, procedimento idêntico ao adotado por Rosa (2000, p. 17), que é explicado por Hand e Henley (1997, p. 527);
- A variável salário do cônjuge foi descartada da análise devido a *missings*;
- Foram criadas duas novas variáveis, percentual do valor do empréstimo sobre o salário e percentual do valor da parcela sobre o salário. Ambas variáveis quantitativas, escala razão que foram categorizadas em faixas da mesma forma que as demais.

A Tabela 1 apresenta as variáveis utilizadas.

Tabela 1 - Variáveis categorizadas

Variável	Categoria	Nome da variável
Sexo	Masculino Feminino	V_SEXO_M V_SEXO_F
Estado civil	Casado Solteiro Outros	V_EST_C V_EST_S V_EST_O
Fone residencial	Sim Não	V_FN_R_S V_FN_R_N
Fone comercial	Sim Não	V_FN_C_S V_FN_C_N
Tempo no emprego atual	Até 24 meses, 25 a 72, 73 a 127, Acima de 127	V_TP_E1 a V_TP_E4
Salário	Até R 650, + 650 a 950, +950 a 1575, +1575 a 2015, +2015 a 3000, Acima de R 3000	V_SAL_F1 a V_SAL_F6
Quantidade de parcelas	Até 4, 5 a 6, 7 a 9, 10 a 12	V_Q_PC_1 a V_Q_PC_4
Primeira aquisição	Sim Não	V_PR_AQ_S V_PR_AQ_N
Tempo na residência atual	Até 12 meses, 13 a 24, 25 a 120, Acima de 120	V_TP_R1 a V_TP_R4
Valor da parcela	Até R 125, +125 a 160, +160 a 260, Acima de R 260	V_VL_PR1 a V_VL_PR4
Valor total do empréstimo	Até R 300, +300 a 400, +400 a 500, +500 a 800, +800 a 1800, Acima de R 1800	V_VL_EM1 a V_VL_EM6
Tipo de crédito	Carnê Cheque	V_CRE_CN V_CRE_CH
Idade	Até 25 anos, 26 a 40, 41 a 58, Acima de 58 anos	V_IIDADE1 a V_IIDADE4
Faixa de CEP residencial	1 2 3 4 5	V_CEP_F1 a V_CEP_F5
Faixa de CEP comercial	1 2 3 4 5	V_CEC_F1 a V_CEC_F5
Código de profissão	1 2 3 4 5 6 7	V_COD_P1 a V_COD_P7
% Valor da parcela/Salário	Até 10%, 10.1% a 13.5%, 13.6% a 16.5%, 16.6% a 22.5%, Acima de 22.5%	V_FX_P1 a V_FX_P5
% Valor do Emprést/Salário	Até 28%, 28.1% a 47.5%, 47.6% a 65%, Acima de 65%	V_FX_E1 a V_FX_E4
Tipo de cliente	1 = Bom 0 = Mau	TIPO

5.2. Rede neural

Neste trabalho, será utilizada uma rede com aprendizado supervisionado, pois já se conhece previamente se o cliente em questão é bom ou mau. Segundo Potts (1998, p. 44), a estrutura de rede neural mais utilizado para este tipo de problema é *multilayer perceptron* (MLP), que se trata de uma rede com arquitetura *feedforward* com múltiplas camadas. A literatura consultada (ARMINGER *et al*, 1997; ARRAES *et al*, 1999; ZERBINI, 2000; CASTRO JR., 2003; OHTOSHI, 2003) comprova esta afirmação. Neste estudo será adotada uma rede MLP.

As redes MLP podem ser treinadas utilizando-se os seguintes algoritmos: Gradiente Descendente Conjugado, Levenberg-Marquardt, *Back propagation*, *Quick propagation* ou Delta-bar-Delta. O mais comum (CASTRO JR., 2003, p. 142) é o algoritmo *Back propagation*, que será detalhado posteriormente.

O modelo implementado tem uma camada de neurônios de entrada; um único neurônio camada de saída, que corresponde ao resultado se o cliente é bom ou mau na classificação da rede e uma camada intermediária com três neurônios, pois foi a rede que apresentou melhores resultados, tanto no quesito de maior percentual de acertos, quanto no quesito de redução do erro médio. Redes que possuíam um, dois ou quatro neurônios, também foram testadas.

Cada neurônio da camada escondida é um elemento de processamento que recebe n entradas ponderadas por pesos W_i . A soma ponderada das entradas é transformada por meio de uma função de ativação não linear $f(.)$.

A função de ativação utilizada neste estudo será a função logística, $\frac{1}{1 + e^{(-g)}}$, onde

$g = \sum_{i=1}^p W_i X_i$ é a soma ponderada das entradas do neurônio.

O treinamento da rede consiste em encontrar os pesos W_i que minimizam uma função de erro. Foi utilizado o algoritmo *Back propagation*. Primeiro, um padrão é apresentado à camada de entrada da rede. A atividade resultante flui através da rede, até que a resposta seja produzida pela camada de saída. No segundo passo, a saída obtida é comparada à desejada para esse padrão particular. Se esta não estiver correta, o erro é calculado. O erro é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas são modificados. Esse processo se repete até se atingir o critério de parada. O erro médio foi o critério de parada adotado. Esse erro é calculado pelo módulo da diferença entre o valor que a rede localizou e o esperado.

Na amostra de validação o erro foi um pouco maior (0,62 x 0,58), visto que o modelo é ajustado com base na primeira amostra.

Inicialmente, a má classificação é de 50%, pois a alocação de um indivíduo como bom ou mau cliente é aleatória; com o aumento das iterações, é atingido o melhor resultado de 30,6% de erro para a amostra de treino e 32,3% para a amostra de validação.

Na Tabela 2, estão algumas das estatísticas da rede adotada.

Tabela 2 - Estatísticas da Rede Neural adotada

Estatísticas obtidas	Treino	Validação
Classificação incorreta de casos	0.306	0.323
Erro médio	0.576	0.619
Erro quadrático médio	0.197	0.211
Graus de liberdade do modelo	220	
Graus de liberdade do erro	7780	
Graus de liberdade total	8000	

5.3. Algoritmos genéticos

O algoritmo genético foi utilizado para encontrar uma equação discriminante que permita pontuar os clientes e, posteriormente, separar os bons e maus clientes de acordo com o escore obtido. Esse caminho foi adotado por Kishore *et al* (2000) e Picinini *et al* (2003).

O algoritmo implementado foi similar ao apresentado em Picinini *et al* (2003). Cada uma das 72 categorias de variável recebeu um peso aleatório inicial. A esses coeficientes foi introduzida uma constante aditiva. O valor de escore do cliente é dado por:

$$S_j = \sum_{i=1}^{72} w_i (p_{ij}), \text{ onde}$$

S_j = Escore obtido pelo cliente j

w_i = Peso relativo à categoria i

p_{ij} = indicador binário igual a 1, se o cliente j possui a categoria i e 0, caso contrário.

Para se definir se o cliente é bom ou mau foi utilizada a seguinte regra:

Se $S_j \geq 0$, o cliente é considerado bom

Se $S_j < 0$, o cliente é considerado mau

Assim sendo, o problema que o algoritmo deve solucionar é encontrar o vetor $W=[w_1, w_2, \dots, w_{72}]$ que resulte em um critério de classificação com uma boa taxa de acertos na predição do desempenho de pagamento do crédito.

Seguindo as fases de um algoritmo genético, têm-se:

- *Início*: foi gerada uma população de 200 indivíduos, com cada cromossomo contendo 73 genes. O peso inicial w_i de cada um dos genes foi gerado aleatoriamente no intervalo $[-1, 1]$ (Picinini *et al*, 2003, p. 464).
- *Função de Aptidão (Fitness)*: cada cliente foi associado ao cálculo de um escore e classificado como bom ou mau. Comparando-se com a informação já conhecida a priori sobre a natureza do cliente, pode-se calcular a precisão de cada cromossomo. O indicador de acertos (Ia) será a função de aptidão, ou seja, quanto maior o indicador, melhor será o cromossomo.
- *Seleção*: neste trabalho foi utilizado um elitismo de 10%, ou seja, para cada nova geração, os vinte melhores cromossomos são mantidos, enquanto os outros cento e oitenta são formados por meio de cruzamento e mutação.
- *Cruzamento (Cross-Over)*: para a escolha dos pais para o cruzamento, foi utilizado o método conhecido como roleta (*roulette wheel*) para seleção dentre os vinte cromossomos que foram mantidos (CHEN; HUANG, 2003, p. 436-437). Neste método, cada indivíduo recebe uma probabilidade de ser sorteado de acordo com seu valor de função de aptidão; para o processo de troca de material genético, foi utilizado um método conhecido como cruzamento uniforme (PAPPA, 2002, p. 22). Neste tipo de cruzamento, cada gene do cromossomo filho é escolhido aleatoriamente entre os genes de um dos pais, enquanto o segundo filho recebe os genes complementares do segundo pai.
- *Mutação*: no processo de mutação, cada gene do cromossomo é avaliado independentemente. Cada gene de cada cromossomo tem probabilidade de 0,5% de sofrer mutação. Sempre que um gene for escolhido para a mutação, a alteração genética é realizada, adicionando-se um pequeno valor escalar k neste gene.
- *Verificação do critério de parada*: foi definido um número máximo de gerações igual a 600. Após as seiscentas iterações, o cromossomo com maior aptidão será a solução.

O algoritmo foi executado três vezes. Seguem os resultados do algoritmo com o maior Ia.

As variáveis com peso muito pequeno foram descartadas. Após uma análise de sensibilidade, decidiu-se considerar as variáveis com peso superior a 0,10 ou inferior a -0,10, exceto para a constante, que se mostrou importante mesmo abaixo do ponto de corte.

O peso das variáveis é apresentado na Tabela 3. Os pesos negativo e positivo indicam maior relação da variável com os clientes considerados, respectivamente, maus e bons. A variável com maior peso negativo foi TcredCarnê e com maior peso positivo foi Nparcelasfx1.

Tabela 3 – Pesos finais das variáveis

Pesos negativos		Pesos positivos	
Variável	Peso	Variável	Peso
Fonecomerc_N	-0.98	Nparcelasfx1	1.42
TcredCarnê	-0.98	Sexofem	0.97
Idade2	-0.98	Profiss7	0.95
Sálclifx1	-0.95	Empr/Sal3	0.95
Profiss2	-0.91	Casado	0.93
Nparcelasfx4	-0.88	Idade4	0.89
Sálclifx4	-0.87	Nparcelasfx2	0.88
Parc/Sal3	-0.8	Parc/Sal5	0.88
CEPres2	-0.79	Valortotfx1	0.83
Valortotfx5	-0.76	TcredCheque	0.81
Nparcelasfx3	-0.65	Tresidfx4	0.75
Sálclifx3	-0.61	Valortotfx2	0.59
Valortotfx4	-0.59	Outros	0.58
CEPcom2	-0.59	CEPres5	0.57
Profiss4	-0.56	Tempregofx4	0.56
Tempregofx1	-0.55	Parc/Sal1	0.55
Foneres_S	-0.54	Sálclifx6	0.47
Idade1	-0.54	Profiss3	0.45
CEPcom3	-0.5	Valorparc4	0.41
Tempregofx2	-0.45	Tempregofx3	0.39
Parc/Sal2	-0.45	Tresidfx3	0.39
CEPres4	-0.44	Valorparc2	0.34
Empr/Sal1	-0.42	Profiss9	0.33
Empr/Sal4	-0.39	Sexomasc	0.29
Valortotfx6	-0.28	Valortotfx3	0.25
CEPres3	-0.28	1aquis_N	0.24
1aquis_S	-0.28	Tresidfx1	0.19
CEPres1	-0.23	Solteiro	0.14
CEPcom1	-0.22	CEPcom4	0.13
CEPcom5	-0.21	Profiss1	0.13
Tresidfx2	-0.14		
Sálclifx2	-0.12		
Profiss8	-0.12		
Constant	-0.08		

5.4. Avaliação da performance dos modelos

Após obtidos os modelos, foram escoradas as três amostras e calculados o Ia e o KS para cada um dos modelos. Os resultados são apresentados nas tabelas 4 e 5.

Tabela 4 - Resultados de classificação

Rede neural										
Treinamento				Validação			Teste			
Predito →				Predito →			Predito →			
	Mau	Bom	% Acerto	Mau	Bom	% Acerto	Mau	Bom	% Acerto	
Observ ↓	Mau	2979	1021	74.5	2236	764	74.5	2255	745	75.2
	Bom	1430	2570	64.3	1177	1823	60.8	1193	1807	60.2
	Total	4409	3591	69.4	3413	2587	67.7	3448	2552	67.7
Algoritmo genético										
Treinamento				Validação			Teste			
Predito →				Predito →			Predito →			
	Mau	Bom	% Acerto	Mau	Bom	% Acerto	Mau	Bom	% Acerto	
Observ ↓	Mau	2692	1308	67.3	1946	1054	64.9	2063	937	68.8
	Bom	1284	2716	67.9	1043	1957	65.2	1073	1927	64.2
	Total	3976	4024	67.6	2989	3011	65.1	3136	2864	66.5

Os dois modelos apresentaram bons resultados de classificação, pois, segundo Picinini *et al* (2003, p. 465): “Modelos de *credit scoring* com taxas de acerto acima de 65% são considerados bons por especialistas”.

Os percentuais de acerto foram similares. A maior taxa de acerto foi para os clientes maus. A Tabela 5, a seguir, apresenta os resultados dos critérios Ia e KS.

Tabela 5 - Índices de comparação

Ia	Amostra		
	Treinamento	Validação	Teste
Rede neural	47.9	45.3	45.3
Algoritmo genético	45.7	42.3	44.2
KS	Amostra		
	Treinamento	Validação	Teste
Rede neural	39	35	35
Algoritmo genético	34	30	32

Os valores KS podem ser considerados bons. Picinini *et al* (2003, p. 465) explicam: “O teste de Kolmogorov-Smirnov (KS) é utilizado no mercado financeiro como um dos indicadores de eficiência de modelos de *credit scoring*, sendo que o mercado considera um bom modelo àquele que apresente um valor de KS igual ou superior a 30”.

Na escolha do modelo mais adequado, sob o prisma dos indicadores Ia e KS, foi eleito o modelo de **rede neural**, pois apresentou melhores resultados em todos os critérios.

6. Conclusões e recomendações

O objetivo deste estudo foi desenvolver modelos de predição de *credit scoring* com base em dados de uma grande instituição financeira usando Redes Neurais e Algoritmos Genéticos.

No desenvolvimento dos modelos de avaliação de crédito precauções na amostragem, definição clara nos critérios de classificação de clientes e tratamento das variáveis antes da modelagem foram cuidados tomados, visando otimizar resultados e minimizar erros.

Os dois modelos apresentaram resultados satisfatórios para a base de dados em questão. O modelo de rede neural apresentou resultados superiores.

A Tabela 6 evidencia a grande variedade de resultados obtidos em outros trabalhos.

Tabela 6 - Precisão da classificação dos modelos construídos (literatura pesquisada)

	Regressão logística	Árvores de classificação	de Redes neurais	Algorit genét.	Análise discriminate
Fritz e Hosemann (2000)		79.5	81.6	82.4	82.7
Arraes <i>et al.</i> (1999)	84.8		85.4		
Chen <i>et al.</i> (2002)			91.9	92.9	
Nanda e Pendharkar (2001)				65.0	62.5
Ohtoshi (2003)	83.5	73.9	85.0		
Picinini <i>et al.</i> (2003)	63.5		64.4	67.5	
Arminger <i>et al.</i> (1997)	67.6	66.4	65.2		
Huang <i>et al.</i> (2004)	77.0		80.0		
Semolini (2002)	68.3		67.4		
Rosa (2000)	70.4	66.6			

Observe-se que as redes neurais apresentam uma grande gama de estruturas e variações que podem ser exploradas em outros trabalhos. Técnicas novas em problema de risco de crédito, como análise de sobrevivência, também merecem atenção em estudos futuros.

Referências bibliográficas

- ANDREEVA, G. (2003) *European generic scoring models using logistic regression and survival analysis*. Bath: Young OR Conference.
- ARMINGER, G., ENACHE, D. & BONNE, T. (1997) Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Trees and Feedforward Networks. *Computational Statistics*, 12(2): 293-310. Berlim: Springer-Verlag.
- ARRAES, D., SEMOLINI, R. & PICININI, R. (1999) *Arquiteturas de Redes Neurais Aplicadas a Data Mining no Mercado Financeiro*. Uma Aplicação para a Geração de Credit Ratings. São José dos Campos: IV Congresso Brasileiro de Redes Neurais.
- CAOUILLE, J., ALTMANO, E. & NARAYANAN, P. (2000) *Gestão do Risco de Crédito*. Rio de Janeiro: Qualitymark.
- CASTRO JR., F. H. F. (2003) *Previsão de Insolvência de Empresas Brasileiras Usando Análise Discriminante, Regressão Logística e Redes Neurais*. Dissertação de Mestrado. Departamento de Administração Universidade de São Paulo. FEA/USP.
- CHEN, M.-C.; HUANG, S.-H (2003) Credit scoring and rejected instances reassigning through evolutionary computation techniques, *Expert Systems with Applications*, v. 24, n. 4, p. 433-441 St. Louis :Elsevier Science.
- CHEN, M.-C., HUANG, S.-H. & CHEN, C.-M. (2002) *Credit Classification Analysis through the Genetic Programming Approach*. Taipei: Proceedings of the 2002 International Conference in Information Management. Tamkang University.
- FAUSETT, L. (1994) *Fundamentals of Neural Networks*. Englewood-Cliffs: Prentice-Hall.
- FIGUEIREDO, R. P. (2001) *Gestão de Riscos Operacionais em Instituições Financeiras – Uma Abordagem Qualitativa*. Dissertação de Mestrado. Belém: Universidade da Amazônia UNAMA.

- FRITZ, S. & HOSEMANN, D. (2000) Restructuring the Credit Process: Behaviour Scoring for German Corporates. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 9(1): 9-21. Nottingham: John Wiley & Sons.
- HAND, D. J. & HENLEY, W. E. (1997) Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of Royal Statistical Society: Series A* (160): 523-541. London: Royal Statistical Society.
- HARRISON, T. & ANSELL, J. (2002) Customer retention in the insurance industry: using survival analysis to predict cross-selling opportunities. *Journal of Financial Services Marketing*, 6(3): 229-239. London: Henry Stewart Publications.
- HAYKIN, S. (1999) *Redes Neurais Princípios e Prática*. Porto Alegre: Bookman.
- HUANG, Z., CHEN, H., HSU, C.-J., CHEN, W. & WU, S. (2004) Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4): 543-558. St. Louis: Elsevier Science.
- KISHORE, J. K.; PATNAIK, L. M.; MANI, V.; AGRAWAL, V. K. (2000) Application of genetic programming for multicategory pattern classification. *IEEE Transactions on evolutionary Computation*, v. 4, n. 3, p. 242-257, Birmingham: IEEE Computational Intelligence Society.
- LEWIS, E. M. (1992) *An Introduction to Credit Scoring*. San Rafael: Fair Isaac and Co., Inc.
- NANDA, S. & PENDHARKAR, P. (2001) Linear models for minimizing misclassification costs in bankruptcy prediction. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 10(3): 155-168. Nottingham: John Wiley & Sons.
- OHTOSHI, C. (2003) *Uma Comparação de Regressão Logística, Árvores de Classificação e Redes Neurais: Analisando Dados de Crédito*. Dissertação de Mestrado. Departamento de Estatística. Universidade de São Paulo. IME/USP.
- OOGHE, H., CAMERLYNCK, J. & BALCAEN, S. (2001) *The Ooghe-Joos-De Vos Failure Prediction Models: A Cross-Industry Validation*. Working paper. Department of Corporate Finance. University of Ghent.
- PAPPA, G. L. (2002) *Seleção de Atributos Utilizando Algoritmos Genéticos Multiobjetivos*. Dissertação de Mestrado. Departamento de Informática. Pontifícia Universidade do Paraná.
- PEREIRA, G. H. A. (2004) *Modelos de risco de crédito de clientes: Uma aplicação a dados reais*. Dissertação de Mestrado. Departamento de Estatística. Universidade de São Paulo. IME/USP.
- PICININI, R., OLIVEIRA, G. M. B. & MONTEIRO, L. H. A. (2003) *Mineração de Critério de Credit Scoring Utilizando Algoritmos Genéticos*. Bauru: VI Simpósio Brasileiro de Automação Inteligente: 463-466.
- ROSA, P. T. M. (2000) *Modelos de Credit Scoring: Regressão Logística, CHAID e REAL*. Dissertação de Mestrado. Departamento de Estatística. Universidade de São Paulo. IME/USP.
- SANTOS, J. O. (2000) *Análise de Crédito: Empresas e Pessoas Físicas*. São Paulo: Atlas.
- SEMOLINI, R. (2002) *Support Vector Machines, Inferência Transdutiva e o Problema de Classificação*. Dissertação de Mestrado. Departamento de Engenharia Elétrica. Universidade Estadual de Campinas. FEEC/UNICAMP.
- SIEGEL, S. (1975) *Estatística Não-Paramétrica para as Ciências do Comportamento*. São Paulo: McGraw-Hill.
- TREVISANI, A. T., GONÇALVES, E. B., D'EMÍDIO, M. & HUMES, L. L. (2004) *Qualidade de Dados – Desafio Crítico para o Sucesso do Business Intelligence*. Itajaí: XVIII Congresso Latino Americano de Estratégia.
- ZERBINI, M. B. A. A. (2000) *Três Ensaios sobre Crédito*. Tese de Doutorado. Departamento de Economia. Universidade de São Paulo. FEA/USP.