Uma sistemática para construção e escolha de modelos de previsão de risco de crédito

Lisiane Priscila Roldão Selau
José Luis Duarte Ribeiro

Resumo

Com o aumento recente nos volumes de créditos a pessoas físicas e, por consequência, nos índices de inadimplência, as empresas estão buscando melhorar sua análise de crédito incorporando critérios objetivos. Técnicas multivariadas têm sido utilizadas para construir modelos de previsão de crédito que, baseados em informações cadastrais dos clientes, levam à criação de um padrão de comportamento em relação à inadimplência. O objetivo deste artigo é propor uma sistemática para construção de modelos de previsão de risco de crédito e avaliar seu desempenho usando três modelos específicos: análise discriminante, regressão logística e redes neurais. O método proposto (denominado Modelo PRC) é composto de seis etapas: (i) delimitação da população; (ii) seleção da amostra; (iii) análise preliminar; (iv) construção do modelo; (v) escolha do modelo; e (vi) passos para implantação. O Modelo PRC foi aplicado em uma amostra de 17.005 clientes de uma rede de farmácias com crediário próprio. Os resultados para este banco de dados específico apontam uma pequena superioridade do modelo de redes neurais em relação aos outros modelos, que pode ser atribuída a sua não linearidade em relação à combinação de variáveis.

Palavras-chave: *Análise de crédito. Análise discriminante. Regressão logística. Redes neurais.*

1 *Introdução*

O crédito no Brasil sempre foi escasso e, devido a políticas mal concebidas e ao processo inflacionário do passado, a cultura do crédito como instrumento para crescimento dos negócios ainda está em estágio inicial. Entretanto, em consequência da maior estabilidade da economia brasileira nos últimos anos, as empresas têm percebido o crédito como um gerador de riquezas e de novos negócios (GOLDBERG apud BUENO, 2003).

Segundo Pereira (2006), o crescimento vertiginoso do crédito nos últimos anos vem alertando os analistas do setor sobre um período de turbulência que está por vir. A Tabela 1 apresenta informações sobre o crescimento do crédito e da renda mensal dos brasileiros. Observa-se uma disparidade entre os números, sendo o aumento da concessão por crédito substancialmente maior que o aumento na renda média.

O crescimento da demanda por crédito à pessoa física no Brasil vem revolucionando esse mercado, fazendo com que as empresas do setor se moldem para ficar à altura das oportunidades. O fenômeno provocou uma reengenharia nos sistemas de crédito em relação à tecnologia na concessão (PEREIRA, 2006).

As empresas que concedem crédito estão apostando numa melhor análise de crédito, evitando trabalhar com clientes que ofereçam maior risco, diminuindo o índice de inadimplência. Por isso estão utilizando, além da experiência do analista, métodos e técnicas que auxiliam na tarefa de decidir se um cliente é merecedor de crédito. A gestão de risco passou a ocupar, nos últimos tempos, posição de destaque na administração financeira, especialmente em consequência da expansão do crédito, do crescimento do mercado e da globalização (BUENO, 2003).

Segundo Steiner et al. (1999), a correta decisão de concessão de crédito é essencial para a sobrevivência das instituições financeiras. Qualquer erro na decisão de conceder o crédito pode significar que, em uma única operação, haja a perda do ganho obtido em dezenas de outras transações bem sucedidas, já que o não recebimento representa a perda total do montante emprestado. Portanto, é importante prever e reduzir a inadimplência, pois os prejuízos com créditos

Tabela 1. Comparativo de indicadores de crédito e renda no Brasil.

	Out./04	**Out.**/05	**Crescimento**
Crédito pessoa física (R$ MM)	108,4	151,5	39,8%
Renda média mensal (R$)	949,24	966,1	1,8%

Fonte: Pereira (2006).

mal sucedidos deverão ser cobertos com a cobrança de altas taxas de juros em novas concessões.

Conforme afirma Schrickel (1997), a análise de risco envolve a habilidade de estabelecer uma regra de decisão para orientar a concessão de crédito, dentro de um cenário de incertezas e constantes mutações e informações incompletas. Esta habilidade depende da capacidade de analisar logicamente situações, muitas vezes complexas, e chegar a uma conclusão prática e factível de ser implementada.

Em muitas empresas, a avaliação da concessão de crédito é baseada em uma variedade de informações vindas de diversas fontes. Os gerentes analisam essas informações de maneira subjetiva e, muitas vezes, não conseguem explicar os processos de tomada de decisão, embora consigam apontar os fatores que influenciam as decisões. Além disso, estes ambientes são dinâmicos, com constantes alterações, e as decisões devem ser tomadas rapidamente (MENDES FILHO et al., 1996).

O uso dos modelos de previsão de risco é vital em alguns casos. Esses modelos, baseados em dados recentes de clientes com a empresa, geram uma pontuação para as características que levam à criação de um padrão de comportamento em relação à inadimplência. Segundo Guimarães e Chaves Neto (2002), quando a empresa tem à sua disposição uma regra de reconhecimento de padrões e classificação que indique previamente a chance de inadimplência de um futuro cliente, a decisão de concessão de crédito fica facilitada, podendo-se então utilizar argumentos quantitativos em substituição aos argumentos subjetivos e decidir com maior confiança.

O objetivo deste artigo é divulgar uma sistemática de construção de modelos de previsão de risco de crédito (Modelo PRC) para concessão de crédito a pessoas físicas, além de comparar desempenho, vantagens e desvantagens de três técnicas estatísticas multivariadas utilizadas para a construção do modelo: análise discriminante, regressão logística e redes neurais.

Este artigo está organizado em cinco seções. Após a introdução apresentada nesta seção inicial, a segunda seção traz a fundamentação teórica, na qual é exposto o referencial sobre modelos de previsão de risco de crédito e as três técnicas multivariadas utilizadas para sua construção. Na terceira seção é detalhada a sistemática proposta para a construção dos modelos PRC. Na quarta seção são apresentados os principais resultados da construção dos modelos e avaliação do desempenho das três técnicas em um banco de dados relacionados a clientes de uma rede de farmácia com unidades no Rio Grande do Sul. Na última seção são apresentadas as considerações finais do estudo, bem como as principais conclusões obtidas.

2 *Fundamentação teórica*

Saber se um cliente provavelmente honrará seus compromissos é uma informação imprescindível na hora de tomar uma decisão com vistas à concessão de crédito. Com isso, pode-se demonstrar que as instituições financeiras poderiam ter um acréscimo nos lucros se, na concessão de crédito, os critérios fossem mais rígidos. De posse da classificação fornecida por um modelo de previsão de risco de crédito, a empresa pode ter um diagnóstico preliminar do provável comportamento de novo cliente, aprovando ou não a concessão do crédito (VASCONCELLOS, 2004).

Steiner et al. (1999) salientam que os modelos quantitativos de previsão são muito utilizados para auxílio na análise de crédito, tendo como vantagens: o aumento do número de merecedores que terão o crédito aprovado, aumentando os lucros; o aumento do número de não merecedores que não receberão o crédito, diminuindo as perdas; solicitações de crédito analisadas com maior rapidez; critérios subjetivos substituídos pelas decisões objetivas; e, por fim, necessidade de menor número de pessoas para administrar o crédito.

Caouette et al. (1999) defendem que os sistemas de pontuação de risco de crédito são importantes por dispor ao credor o conhecimento que não estaria, de outra maneira, prontamente disponível. Esses autores acrescentam que há uma grande vantagem competitiva com a utilização dos modelos, pois um sistema de pontuação integrado permite operar em diversas regiões geográficas, envolvendo diversas pessoas e, mesmo assim, mantendo objetividade nas decisões.

Os modelos de previsão de risco vêm como uma ferramenta de auxílio ao crédito massificado, que é caracterizado pela avaliação de um grande número de solicitações de pequenos valores, já que a competitividade do mercado exige decisões rápidas. O analista informa os dados de seu potencial cliente no sistema de crédito e, imediatamente, o computador fornece a informação quanto à aprovação do crédito. Na verdade, o método estatístico utilizado para a construção do modelo leva em consideração o histórico da instituição com seus clientes, possibilitando a identificação das características capazes de diferenciar o bom do mau pagador (SILVA, 2006).

Com o rápido desenvolvimento da informática, a partir dos anos 70, os sistemas de pontuação de crédito baseados em modelos surgiram no negócio de financiamento a pessoas físicas e jurídicas como um dos métodos mais importantes de suporte à tomada de decisão para grandes volumes de solicitações de crédito (SANTOS, 2000).

Alguns autores como Silva (2003) e Caouette et al. (1999) têm citado a análise multivariada como uma ferramenta poderosa na avaliação do risco de inadimplência presente na concessão de crédito. Uma das vantagens de seu uso para elaboração de sistemas de pontuação é que os pesos a serem atribuídos aos índices são determinados por

métodos numéricos de ajuste de modelos, o que exclui a subjetividade no momento da análise ou mesmo o estado de espírito do analista do crédito.

Após a quantificação de cada característica ou variável de risco selecionada do tomador, obtém-se uma pontuação que determinará de maneira padronizada, consistente e objetiva, baseando-se nas probabilidades de reembolso calculadas, se o crédito pode ser concedido ou deve ser recusado (SANTOS, 2000).

Existem diversas técnicas multivariadas para a construção de modelos de previsão de risco de crédito, entre elas: regressão linear múltipla; programação linear; algoritmos genéticos; árvore de decisão; análise discriminante; regressão logística; redes neurais; e, mais recentemente, a análise de sobrevivência como exposto no trabalho de Andreeva et al. (2007). Na sequência, são descritas as três abordagens comparadas neste artigo.

2.1 Análise discriminante

A análise discriminante envolve a determinação de uma combinação linear de variáveis independentes que discriminarão melhor uma observação entre grupos definidos a *priori* (HAIR et al., 2005). Segundo Johnson e Wichern (2002), a análise discriminante é uma técnica multivariada com o objetivo de tratar dos problemas que envolvem separar conjuntos distintos e alocar novos objetos em conjuntos previamente definidos.

Conforme Johnson e Wichern (2002), a ideia de discriminar e classificar foi introduzida por Ronald A. Fisher no primeiro tratamento moderno dos problemas de separação de conjuntos em seu trabalho sobre espécies de plantas em 1935. O método proposto por Fisher consiste basicamente em separar um conjunto de objetos em duas classes pré-definidas.

Vários estudos deram sequência para a utilização desta técnica em diversas áreas. Em Finanças, um dos mais relevantes estudos na previsão do risco de crédito por meio de análise discriminante foi realizado por Edwards I. Altman, em 1968. Ele utilizou índices oriundos de demonstrações financeiras para prever a probabilidade de falência de empresas (CORRAR et al., 2007).

De acordo com Hair et al. (2005), a discriminação é alcançada estabelecendo os pesos discriminantes, que são calculados com o objetivo de maximizar a variância entre os grupos e, por consequência, minimizar a variância dentro dos grupos. A combinação linear para uma análise discriminante, também conhecida como função discriminante, tem a forma apresentada na Equação 1.

$$Z_k = a + W_1X_{1k} + W_2X_{2k} + ... + W_nX_{nk} \qquad (1)$$

em que, Z_k é o escore Z discriminante para o objeto k; a é o intercepto; W_i é o coeficiente discriminante para a variável independente i; e X_{ik} é a variável independente i para o objeto k.

Para testar a significância estatística da função discriminante, Hair et al. (2005) sugerem ainda a utilização de uma medida generalizada da distância entre os centroides dos grupos. Essa medida compara as distribuições dos escores discriminantes dos dois grupos. Quando a sobreposição das distribuições é pequena, considera-se que a função separa bem os grupos, mas se a sobreposição é grande, a função discriminante tem uma separação pobre dos grupos. A Figura 1 ilustra esse conceito, utilizando duas distribuições de escores (à esquerda, observa-se boa separação; à direita, a separação é mais pobre).

Se a significância do modelo foi verificada, costuma ser de interesse a criação de matrizes de classificação para verificar de forma mais precisa o poder de discriminação da função. Para isso é necessária a determinação de um escore de corte, ao qual cada elemento é comparado para determinar em qual grupo será classificado. Tal escore de corte pode ser obtido por uma média ponderada dos centroides dos grupos, como é apresentado na Equação 2.

$$Z_c = \frac{N_AZ_B + N_BZ_A}{N_A + N_B} \qquad (2)$$

em que, Z_c é o valor de escore de corte crítico; N_A é o número de elementos no grupo A; N_B é o número de elementos no grupo B; Z_A é o centroide para o grupo A; e Z_B é o centroide para o grupo B.

Corrar et al. (2007) ressaltam que as variáveis independentes geralmente são métricas com valores contínuos, mas também podem assumir valores que representam categorias (alto e baixo, por exemplo). Na área de Marketing, há estudos que se encaixam nesta última situação, porém os autores destacam que técnicas como regressão logística

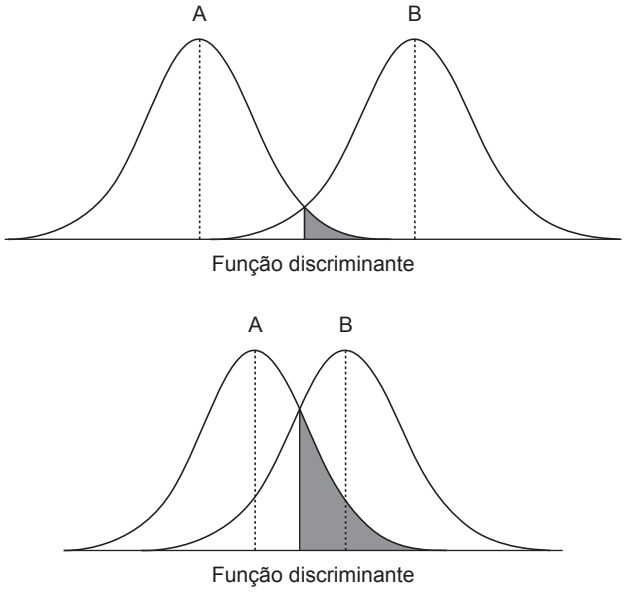

Figura 1. Representação univariada de escores Z discriminantes. Fonte: Hair et al. (2005).

e redes neurais são mais apropriadas a esse tipo de variável independente em razão das características de seus algoritmos.

Caouette et al. (1999) destacam que, embora não seja tão difundida quanto outras técnicas estatísticas, como por exemplo a regressão linear, a análise discriminante tem tido crescente utilização nas áreas de economia e finanças. Muitas das aplicações consistem na construção de modelos destinados à previsão de falência de empresas e à inadimplência de pessoas físicas.

2.2 Regressão logística

A utilização da técnica de regressão logística é adequada em muitas situações, porque permite que se analise o efeito de uma ou mais variáveis independentes (categóricas ou métricas) sobre uma variável dependente dicotômica, representando a presença (1) ou ausência (0) de uma característica (HOSMER; LEMESHOW, 1989).

A regressão logística é uma técnica que se caracteriza por descrever a relação entre várias variáveis independentes (X_j) e uma variável dependente binária (Y), codificada como 1 ou 0 (KLEINBAUM, 1996). Este modelo descreve o valor esperado de Y por meio da expressão apresentada na Equação 3.

$$E(Y) = \frac{1}{1 + exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j X_j \right) \right]} \quad (3)$$

O objetivo na análise de regressão logística é descrever o modelo matemático de Y em função dos valores de X_j e de β_j. Assim, utilizando o método de estimação da máxima verossimilhança, os parâmetros do modelo são ajustados (HOSMER; LEMESHOW, 1989). A expressão geral do modelo logístico é dada pelas Equações 4 e 5.

$$f(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

$$z = \beta_0 + \sum_{j=1}^{k} \beta_j X \quad (5)$$

em que z é conhecido com log odds, variando de $-\infty$ a $+\infty$ como se observa na Figura 2. Assim, a função logística $f(z)$ normaliza a saída do modelo para o intervalo [0,1], informando a probabilidade de ocorrência do evento de interesse.

De acordo com Hosmer e Lemeshow (1989), a regressão logística tornou-se, portanto, um método padrão de análise de regressão para variáveis medidas de forma dicotômica. Desta forma, a diferença principal da regressão logística quando comparada ao modelo linear clássico é que a distribuição da variável resposta segue uma distribuição binomial, e não uma distribuição normal.

A esse respeito, Hair et al. (2005) afirmam que a regressão logística se assemelha em muitos pontos à regressão linear, mas difere no sentido de prever a probabilidade de um evento ocorrer. Para obter um valor previsto delimitado entre zero e um, usa-se uma relação assumida entre as variáveis independentes e a variável dependente que lembra uma curva em forma de 'S', a distribuição sigmoide.

Os modelos lineares de regressão não podem acomodar tal relação entre as variáveis, já que ela é inerentemente não linear. Por isso a regressão logística foi desenvolvida para lidar especificamente com essas questões. A regressão logística deriva seu nome justamente dessa transformação logística utilizada com a variável dependente (HAIR et al., 2005). Ainda de acordo com esses autores, devido à natureza não linear da transformação logística, para a estimação dos coeficientes do modelo é recomendado o uso do método da máxima verossimilhança (ao invés do método tradicional de mínimos quadrados, utilizada na regressão linear).

O modelo de regressão logística é obtido pelo procedimento de comparação da probabilidade de um evento ocorrer com a probabilidade de não ocorrer. De acordo com Hair et al. (2005), esta razão pode ser expressa segundo a Equação 6.

$$\frac{Prob\ (evento\ \ ocorrer)}{Prob\ (evento\ \ não\ \ ocorrer)} = e^{B_0 + B_1 X_1 + \ldots + B_n X_n} \quad (6)$$

Os coeficientes estimados (B_0, B_1, ..., B_n) são medidas das variações na proporção das probabilidades, chamada de razão de desigualdade. São expressos em logaritmos, necessitando serem transformados para facilitar a interpretação. Um coeficiente positivo revela que um aumento naquela variável aumenta a probabilidade de ocorrência do evento, enquanto que um valor negativo significa o oposto.

Ao utilizar a técnica de regressão logística, o interesse pode estar na identificação do efeito de um fator de

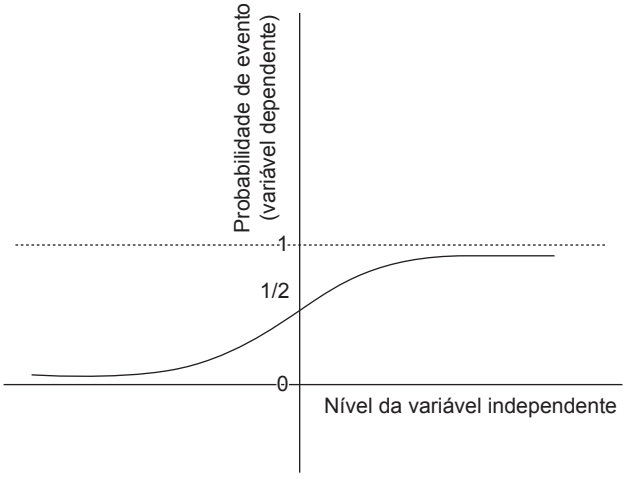

Figura 2. Forma da relação logística entre as variáveis.
Fonte: Hair et al. (2005).

risco específico ou em determinar quais são os vários fatores associados com a variável resposta. Segundo Hosmer e Lemeshow (1989), a função logística vem sendo bastante utilizada, não apenas pela simplicidade de suas propriedades teóricas, mas, principalmente, devido à sua simples interpretação como o logaritmo da razão de chances (*odds ratio*).

Para testar a significância dos coeficientes, Hair et al. (2005) sugerem o uso da estatística de Wald. Ela fornece a significância estatística para cada coeficiente estimado, de modo que o teste de hipóteses pode ocorrer como acontece na regressão múltipla. Outra semelhança com a regressão múltipla está no fato de que dados nominais e categóricos podem ser tomados como variáveis independentes do modelo por meio de codificação dicotômica. Além disso, o procedimento de seleção de modelos *stepwise* também está disponível.

Segundo Corrar et al. (2007), um dos motivos pelos quais a regressão logística tem sido muito utilizada é o pequeno número de suposições. Utilizando esta técnica, o pesquisador consegue contornar certas restrições encontradas em outros modelos multivariados.

Apesar de sua flexibilidade, existe o pressuposto de baixa correlação entre as variáveis independentes, já que o modelo de regressão logística é sensível à colinearidade entre as variáveis (HAIR et al., 2005). A utilização de variáveis altamente correlacionadas para a estimação do modelo pode ocasionar estimativas inflacionadas dos coeficientes de regressão (HOSMER; LEMESHOW, 1989).

Segundo Corrar et al. (2007), o método *stepwise* para escolha de variáveis para compor o modelo é considerado como uma das ações corretivas para os problemas de multicolinearidade. O procedimento de avaliação das variáveis independentes desconsidera variáveis que apresentem sinais de multicolinearidade, optando por manter no modelo apenas aquelas de maior significância estatística.

Portanto, o pesquisador que tem um problema que envolva uma variável dependente dicotômica não precisa apelar para métodos elaborados para suprir as limitações da regressão múltipla, nem precisa forçar-se a usar a análise discriminante, principalmente se suas suposições estatísticas não são satisfeitas. A regressão logística aborda satisfatoriamente esses problemas e oferece um método de análise desenvolvido especialmente para lidar com esse tipo de situação da forma mais eficiente possível (HAIR et al., 2005).

2.3 Redes neurais

As Redes neurais são uma das técnicas de tratamento de dados mais recentes e que tem despertado grande interesse tanto de pesquisadores da área de tecnologia quanto da área de negócios (CORRAR et al., 2007). Segundo Kovács (2002), dependendo do problema para o qual são aplicadas, as redes neurais têm apresentado desempenho superior a outros métodos de estatística multivariada. Por exemplo, em Subramanian et al. (1993), é apresentada a comparação da técnica com outros métodos estatísticos, em que os autores concluíram que as redes neurais apresentaram melhores resultados, em diversas circunstâncias, principalmente para análises de maior complexidade.

De acordo com Hair et al. (2005), redes neurais são uma abordagem diferente em relação as outras técnicas estatísticas multivariadas. A diferença não está somente na estrutura, mas também no processo, já que as redes neurais têm um elemento-chave: a aprendizagem. Essa é outra analogia com o cérebro humano, pela qual os erros de saída são retornados ao início da rede, sendo o modelo ajustado adequadamente.

A estrutura e a operação das redes neurais podem ser descritas por quatro conceitos: (1) o tipo de modelo de rede neural; (2) as unidades de processamentos (nós) que coletam informações, processam e criam um valor de saída; (3) o sistema de nós arranjados para transferir sinais dos nós de entrada para os nós de saída, por meio dos nós intermediários; e (4) o aprendizado pelo qual o sistema 'retorna' erros na previsão para ajustar o modelo (HAIR et al., 2005).

Haykin (2001) apresenta o elemento mais básico de uma rede neural (Figura 3). O nó é análogo ao neurônio do cérebro humano, recebendo informações de entrada e criando resultados de saída. O processamento dessa informação acontece pela criação de um valor somado no qual cada entrada é multiplicada por seu respectivo peso. Esse valor é então processado por uma função de ativação, gerando uma saída que é enviada para o nó seguinte. Em geral, a função de ativação é não linear, como a função sigmoide, da classe geral de curvas em forma de 'S' que incluiu a função logística. Outro elemento de entrada dos nós, chamado bias, utilizado também no processamento, funciona como uma constante da função aditiva.

Uma rede neural é um arranjo sequencial de três tipos de nós: de entrada, de saída e intermediários (ocultos ou escondidos). Os nós de entrada recebem os dados de cada caso e os transmitem para o restante da rede. Variáveis métricas necessitam apenas um nó para cada variável, já variáveis não métricas precisam ser codificadas, de forma que cada categoria é representada por uma variável binária (HAIR et al., 2005). Na Figura 4 é apresentado um modelo representativo do arranjo de uma rede neural.

Segundo Hair et al. (2005), um nó de saída recebe entradas e obtém um valor de saída, sendo este o resultado da previsão. É por meio das camadas ocultas e da função de ativação que a rede neural consegue representar as relações não lineares entre as variáveis.

Há dois tipos de redes neurais: não recorrentes e recorrentes. As redes não recorrentes não possuem realimentação de suas entradas com os dados de saída,

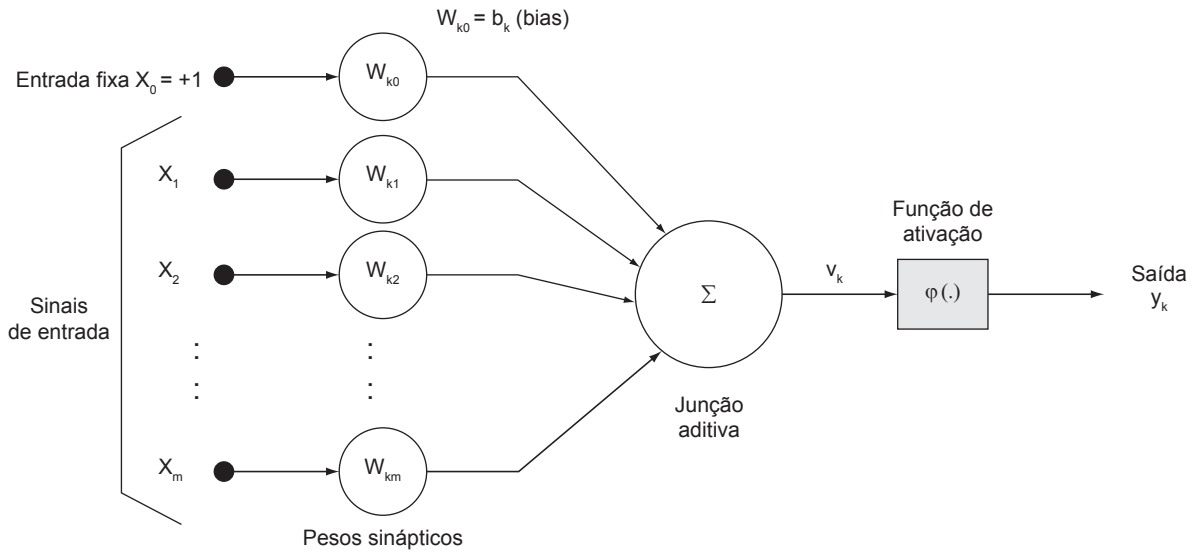

Figura 3. Modelo não linear de um nó de uma rede neural.
Fonte: Haykin (2001).

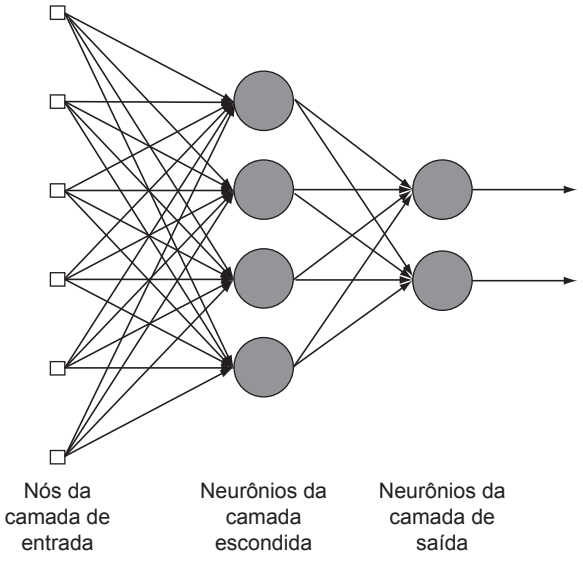

Figura 4. Modelo estrutural de uma rede neural.
Fonte: Haykin (2001).

por isso são ditas "sem memória". Nas redes recorrentes há a realimentação das entradas com os dados de saída, caracterizando um processo iterativo para diminuição do erro de predição (LAWRENCE, 1994; KOVÁCS, 2002).

A forma mais comum de treino da rede é a retropropagação. As variáveis de entrada são apresentadas aos nós e seu efeito se propaga através da rede, camada por camada, obtendo-se uma saída para a rede, sem alteração dos pesos sinápticos. O valor do erro é calculado, comparando-se a saída da rede com a saída esperada, e os pesos sinápticos são ajustados, tentando reduzir esse erro. Depois de treinada, a rede neural está apta a associar um conjunto de valores que são apresentados em suas entradas a um resultado de saída (HAYKIN, 2001).

Para Yu et al. (2002), o algoritmo de retropropagação do erro é essencial para muitos trabalhos atuais sobre aprendizado em redes neurais. Segundo Loesch e Sari (1996), o algoritmo pode ser dividido em 5 passos: (i) apresentação de um padrão de entrada e da saída desejada; (ii) cálculo dos valores de saída; (iii) ajuste dos pesos da camada de saída; (iv) ajuste de pesos das camadas escondidas; e (v) verificação da magnitude do erro

Ainda que o modelo de rede neural possa ser utilizado em situações que outras técnicas estatísticas, tais como regressão múltipla, análise discriminante e regressão logística seriam também indicadas, ele não informa sobre a importância relativa das variáveis independentes na predição devido à combinação não linear de pesos que ocorre na camada oculta. Nesse contexto, Hair et al. (2005) indicam a aplicação de redes neurais em problemas de previsão e classificação quando o interesse está na precisão de classificação e não na interpretação das variáveis independentes.

Saunders (2000) argumenta que a aplicação de técnicas não lineares, como redes neurais, à análise de risco de crédito promete uma melhora sobre os modelos mais antigos de pontuação de crédito com uso de técnicas lineares de estatística. Com a utilização de redes neurais consegue-se poder explicativo adicional, haja vista as complexas correlações e interações entre as variáveis independentes, que muitas vezes são realmente não lineares.

3 *Sistemática para desenvolvimento do modelo Previsão de Risco de Crédito – PRC*

Antes de apresentar a sistemática para desenvolvimento do Modelo de Previsão de Risco de Crédito, é apresentada uma visão geral das três técnicas que serão utilizadas neste

artigo (ver Quadro 1). O objetivo é expor as diferenças entre as técnicas que motivaram a sua escolha na realização deste estudo.

A sistemática utilizada para desenvolvimento dos modelos de previsão de risco de crédito proposta neste trabalho é constituída de seis etapas (ver Figura 5). As etapas para desenvolvimento do Modelo PRC são explicadas na sequência, contemplando desde os primeiros passos para iniciar a pesquisa até as atividades de implantação do modelo. Maiores detalhes podem ser vistos em Selau (2008).

3.1 Delimitação da população

A suposição básica para construir um modelo de previsão de crédito é que o padrão de comportamento dos clientes se mantém ao longo do tempo. Portanto, considerando que a construção do modelo é exclusivamente baseada na experiência do uso do crédito pela empresa, todos os dados utilizados no desenvolvimento são oriundos dos registros deste negócio. Os dados da amostra têm que constituir toda a informação conhecida dos clientes na hora da concessão do crédito e também seus *status* subsequentes como bons e maus pagadores.

Antes da definição dos parâmetros para a seleção da amostra, é necessário decidir para qual segmento da população o modelo construído vai ser utilizado. Em empresas de pequeno e médio porte, nas quais há somente um produto de crédito, pode ser a população (todos os clientes). Já para empresas grandes, nas quais são oferecidos diversos produtos de crédito, a população para o estudo deve ser limitada por tipo de produto.

Para a construção do modelo, é imprescindível que existam créditos concedidos pelo negócio e que os resultados da concessão tenham sido avaliados. Portanto, para que se possa desenvolver o modelo, inicialmente se deve definir

Quadro 1. Comparação de características das técnicas utilizadas no estudo.

Características / Técnicas	Processamento do relacionamento das variáveis	Aprendizado	Pressupostos para utilização
Análise discriminante	Linear	Não tem	Normalidade multivariada, homogeneidade de matrizes de variância; ausência de multicolinearidade
Regressão logística	Não linear, forma fixa	Não tem	Ausência de multicolinearidade
Redes neurais	Não linear, forma ajustável	Retropropagação	Não tem

Figura 5. Etapas para desenvolvimento do Modelo PRC.

os conceitos de desempenho aceitável e inaceitável. Quatro grupos devem ser separados no total de créditos concedidos: (i) os clientes que nunca utilizaram o crédito – sem uso; (ii) os clientes com pouco ou nenhum atraso – bons; (iii) os clientes em faixas de atrasos intermediárias – indeterminados; (iv) os clientes com atrasos consideráveis – maus. A definição de atrasos consideráveis deve ser feita pelo concessor, que definirá os atrasos que podem ser aceitos pelo negócio. Na construção do modelo, somente são utilizados os grupos de clientes bons e maus para facilitar a análise e para acentuar a separação de perfis.

3.2 Seleção da amostra

A análise das informações constantes no banco de dados da empresa pode ser feita pela observação detalhada da proposta de crédito. Por meio dela é possível identificar as possíveis variáveis que poderão fazer parte do modelo final e, portanto, devem ser listadas na amostra de estudo. Dentre as possíveis informações selecionadas, chamadas também de variáveis demográficas, pode-se citar: sexo, idade, escolaridade, estado civil, tipo de ocupação, tipo de residência, tempo no emprego atual, entre outras.

Para definição do período para extração da amostra é necessário observar um tempo entre a concessão do crédito e a verificação de seu desempenho de pagamento. É necessário que os clientes que farão parte da amostra tenham sido incluídos há no mínimo 12 a 18 meses, tempo necessário para se consolidar o comportamento dos clientes. Na utilização de técnicas multivariadas, o tamanho da amostra depende do número de variáveis independentes que farão parte do estudo para construção do modelo final. Desta forma, sugere-se a utilização de uma proporção de 20 observações para cada variável independente.

De posse do banco de dados, é efetuada uma análise exploratória, e todos os campos são analisados quanto ao seu conteúdo. Neste momento, devem ser verificadas questões quanto à qualidade de preenchimento, consistência dos campos e presença de observações faltantes (*missing*), eliminando dados inconsistentes ou atípicos.

Um último ponto a ser considerado quanto à amostra é a questão da divisão entre amostra de análise e teste, a fim de evitar qualquer tipo de viés. Portanto, para verificar se o poder preditivo do modelo é mantido para outras amostras provenientes da mesma população, são necessários testes para a sua validação. Não existem regras fixas quanto à partição da amostra. Devido à importância que a construção do modelo tem em relação ao seu teste, propõe-se a divisão da amostra total em 80% para análise e 20% para teste do modelo final.

3.3 Análise preliminar

O primeiro passo, antes de começar as análises das informações do banco de dados, trata-se da escolha das variáveis que entrarão na análise, podendo vir a integrar o modelo final. Com o uso de tabelas de contingência, calcula-se o risco relativo (RR) associado aos diferentes atributos (níveis) das variáveis independentes, dividindo-se o percentual de bons clientes pelo percentual de maus de cada atributo. Quanto mais os percentuais de bons e maus diferirem para os atributos de uma mesma variável, maior será a utilidade dessa variável para o prognóstico de desempenho futuro. Por exemplo, se a mesma fração de bons e maus clientes tem casa própria ou alugada, essa variável não provê nenhuma informação que ajude a estabelecer a probabilidade de um cliente vir a se tornar bom ou mau pagador.

Como regra geral, Lewis (1992) propõe que os atributos sejam agrupados, segundo o valor do risco relativo, em 7 classes, como apresentado na Figura 6. Atributos classificados como neutros não são utilizados na análise, já que não contribuem na diferenciação entre os grupos bom e mau.

Após a seleção dos atributos que farão parte da análise multivariada, passa-se para a criação de uma variável *dummy* para cada um deles (ex.: cada nível de escolaridade será uma variável *dummy*). Essa variável só assume dois valores: 1 ou 0 (ou um cliente possui formação superior ou não). Com esse artifício evitam-se problemas decorrentes da não linearidade dos atributos no cálculo da análise multivariada.

Neste estudo, o tratamento das variáveis seguiu os procedimentos propostos por Lewis (1992). Draper e Smith (1998) propõem uma forma alternativa de definição de variáveis *dummies*, que não é tratada neste artigo. Enquanto sugestão para trabalhos futuros, recomenda-se verificar se o procedimento utilizado na definição das variáveis *dummies* pode influenciar significativamente os resultados dos modelos.

3.4 Construção do modelo

Uma vez reduzidos os dados a agrupamentos de atributos, cuidadosamente escolhidos para todas as características, e criação das respectivas variáveis *dummies*, cada analista escolhe o método a ser utilizado para a modelagem. Neste estudo sugere-se a utilização de análise discriminante, de regressão logística e de redes neurais. Tais métodos encontram-se entre os mais utilizados para a construção de modelos de crédito e por isso foram os escolhidos.

A escolha do *software* a ser utilizado é um passo importante, devendo-se verificar suas características quanto aos recursos de análise e facilidade de uso. A construção dos modelos com uso das técnicas de análise discriminante e regressão logística pode ser feita utilizando-se os pacotes estatísticos comerciais. Para o treinamento e teste das redes neurais, é necessário um módulo adicional aos pacotes estatísticos convencionais ou um *software* específico para o desenvolvimento da técnica.

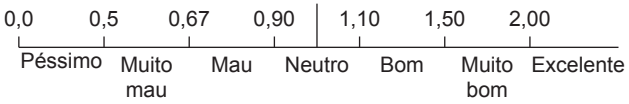

Figura 6. Classes de risco relativo para agrupamento.

A construção de um modelo adequado é uma tarefa complexa. É necessária, por exemplo, a avaliação de variáveis que devem entrar ou sair da análise para evitar problemas de multicolinearidade. Para a seleção das variáveis independentes que devem compor o modelo, pode-se utilizar o método *stepwise*, incorporado em muitos pacotes estatísticos, que automaticamente seleciona a melhor combinação de variáveis independentes para entrada no modelo.

Para seguir com a avaliação, é necessária a verificação dos pressupostos para cada modelo. A metodologia de redes neurais é mais flexível que as demais técnicas estatísticas, sendo que nenhum pressuposto precisa ser verificado. Os principais pressupostos da análise discriminante a serem verificados são: normalidade multivariada, homogeneidade de matrizes de variância e ausência de multicolinearidade. Na regressão logística, o pressuposto a ser verificado é o da ausência de multicolinearidade. Assim como na análise discriminante, tal suposição é atendida com a utilização do método *stepwise* para a seleção das variáveis independentes.

3.5 Escolha do modelo

Duas medidas de desempenho podem ser usadas para a escolha do melhor modelo: (i) percentual de classificações corretas; (ii) o valor do teste de Kolmogorov-Smirnov (KS) para duas amostras. O percentual de acerto nas classificações deve ser avaliado pelo cruzamento dos resultados observados e previstos pelo modelo. Desta forma, a taxa de acerto é medida pela divisão da quantidade de clientes corretamente classificados pelo total de clientes que fizeram parte da análise. Especialistas consideram satisfatórios os modelos com taxa de acerto superior a 65%. Com o cálculo do teste de KS, o que se busca é determinar a diferença máxima entre duas distribuições acumuladas. Obtendo-se uma diferença maior que 30 entre as distribuições de bons e maus pagadores, pode-se considerar que o modelo é eficiente na predição dos dois grupos.

3.6 Passos para implantação

Por meio dos critérios previamente definidos, o melhor modelo é escolhido. Com isso deve-se programar a implantação do modelo junto à empresa. O Departamento de Tecnologia da Informação da empresa deve adequar seus sistemas para receber o modelo final e programar sua utilização junto às demais áreas envolvidas.

Quanto à definição do ponto de corte, a escolha deve ser feita pela empresa, por se tratar de uma decisão mais estratégica do que científica, já que envolvem questões como níveis de aceitáveis de inadimplência e inclinação da alta administração no que concerne à exposição ao risco.

Uma das questões mais importantes na implantação do modelo é garantir que as propostas de crédito sejam avaliadas nas mesmas condições em que foram avaliados os clientes que constituíram a amostra utilizada no desenvolvimento do modelo. Deve-se garantir, portanto, que as informações obtidas no momento da análise de novas propostas apresentem condições semelhantes às coletadas no passado.

Sugere-se que, após um ano de utilização do modelo, uma revisão seja feita, seguindo os mesmos passos para a construção do modelo original. Além disso, a revisão é necessária quando houver mudança significativa na inadimplência, na lucratividade, nos prazos ou condições do negócio e, principalmente, no perfil da população. Tais alterações devem ser monitoradas por meio de relatórios de acompanhamento para o modelo.

4 *Resultados*

Nesta seção são apresentados os resultados obtidos em cada etapa da aplicação da sistemática.

4.1 Delimitação da população

No desenvolvimento desta pesquisa, foram utilizadas informações sobre os clientes de uma rede de farmácias com unidades em todo o Rio Grande do Sul. É oferecido aos clientes um cartão de crédito próprio como forma de facilitar o pagamento das compras, sendo este o único produto de crédito da empresa.

Para a definição de desempenho de pagamento, o cliente bom é definido como aquele que tem atrasos de até 30 dias e os clientes maus são aqueles com pelo menos um atraso superior a 60 dias. Como indefinidos são classificados os clientes com atrasos entre 31 e 60 dias. Além destes três grupos de clientes, foi separado um quarto grupo composto dos clientes que não tiveram nenhuma compra com o cartão.

4.2 Seleção da amostra

A identificação das informações disponíveis no sistema da empresa, que serviram como variáveis independentes para a análise, foi feita a partir da proposta que é preenchida pelos clientes no momento da solicitação do crédito. Nesta etapa, foram selecionadas 16 variáveis (sexo, idade, estado civil, escolaridade, renda, tipo de renda, profissão, tipo de ocupação, CEP residencial e comercial, tempo de serviço, crédito com outros estabelecimentos, tipo de residência, cidade de nascimento, ter filhos e pagar pensão).

O período de cadastro constante na amostra compreende informações de clientes aprovados de dezembro de 2005 a junho de 2006, de acordo com o período de 12 a 18 meses após a concessão. O total de clientes na amostra e a quantidade por tipo de cliente são apresentados na Tabela 2. Como somente os grupos bons e maus são considerados para o desenvolvimento do modelo, a amostra considerada se reduz a 11681 clientes.

Antes de passar para a separação das amostras de teste e análise, alguns dados tiveram de ser eliminados devido a problemas de preenchimento como inconsistências, *missing* e *outliers*, fazendo que com que a amostra de trabalho ficasse com um total de 11394 clientes.

De forma aleatória, foram separadas as amostras de análise e teste, na proporção de 80% e 20%, respectivamente. A amostra de teste, reservada para a posterior validação dos modelos construídos, selecionada aleatoriamente, ficou composta de 1631 clientes bons e 648 maus, num total de 2279 clientes. A amostra de análise ficou formada, portanto, por 6305 clientes bons e 2720 maus; um total de 9115 observações.

4.3 Análise preliminar

A análise para escolha das variáveis é realizada por meio do cálculo de risco relativo, dividindo-se o percentual de bons clientes pelo percentual de maus de cada atributo. Nesta fase inicial, quatro variáveis (tipo de renda, crédito em outros estabelecimentos, pagamento de pensão e renda) foram excluídas da análise por terem poder de discriminação muito baixo, em que o risco relativo dos seus diferentes atributos era próximo de 1.

Para incluir as variáveis profissão, cidade de nascimento, CEP residencial e comercial na análise, foi necessário agrupá-las, dado o grande número de atributos de cada uma delas. Para tal agrupamento foi utilizada a escala apresentada na Figura 3 e foi estipulada a ocorrência mínima de 30 observações em cada atributo para que pudesse ser considerado para classificação. Desta forma, o agrupamento das variáveis levou à criação de sete grupos, do cliente péssimo ao excelente. Cada grupo de risco é transformado em uma variável *dummy* (0 ou 1) que serão, portanto, as variáveis independentes para a construção do modelo. Para as demais informações do banco de dados também foram criadas variáveis *dummies*, obtendo-se um total de 69 possíveis variáveis independentes para os modelos. Na Tabela 3, é apresentado um exemplo para a variável idade. Detalhes referentes às demais variáveis podem ser vistos em Selau (2008).

Tabela 2. Total de clientes por tipo.

Tipo cliente	Quantidade	%
Mau	3655	21,5
Bom	8026	47,2
Indefinido	795	4,7
Sem Uso	4529	26,6
Total	17005	100

4.4 Construção do modelo Previsão de Risco de Crédito – PRC

Para a construção dos modelos com uso das técnicas de análise discriminante e regressão logística, foi utilizado o SPSS versão 13.0 (*Statistical Package for Social Science*). O *software* utilizado para treinamento e teste das redes neurais, empregada para a construção do terceiro modelo proposto, foi o *BrainMaker Professional* versão 3.7.

Nos testes iniciais para a construção do modelo discriminante e logístico, pelo método *stepwise*, utilizaram-se níveis de significância para a entrada e saída de variáveis do modelo de 5% e 10%, respectivamente. Para que algumas variáveis tivessem significância para entrar no modelo final, foi necessário fazer o agrupamento de *dummies* próximas, como por exemplo, os grupos de cidade de nascimento 1 e 2 (péssimo e muito mau) e os grupos de profissões 6 e 7 (muito bom e excelente), entre outros.

Dentre as 69 variáveis *dummies* relacionadas como possíveis variáveis independentes, apenas 26 tiveram poder discriminatório significativo para compor o modelo discriminante final. Na Equação 7 são apresentadas as variáveis significativas para a obtenção do escore do modelo discriminante. A especificação das variáveis é apresentada no Quadro 2.

$$\begin{aligned} Y = {} & 0{,}154 - 1{,}172\ DIDAD1 - 0{,}547\ DIDAD23 - \\ & - 0{,}315\ DIDAD4 + 0{,}272\ DIDAD6 + \\ & + 0{,}650\ DIDAD7 + 0{,}933\ DIDAD8 + \\ & + 0{,}366\ DSEXOF - 0{,}344\ DPRIM + \\ & + 0{,}257\ DSUP + 0{,}436\ DCASADO + \\ & + 0{,}492\ DTSERV67 + 0{,}641\ DTSERV89 - \\ & - 0{,}343\ DFILHO - 0{,}580\ DRES\_ALU - \\ & - 0{,}602\ DGCEPR12 - 0{,}329\ DGCEPRE3 - \\ & - 0{,}949\ DGCEPCO1 + 0{,}298\ DGCEPC56 + \\ & + 0{,}707\ DGCEPCO7 - 0{,}936\ DGPROF1 - \\ & - 0{,}399\ DGPROF2 + 0{,}212\ DGPROF5 + \\ & + 0{,}293\ DGPROF67 - 0{,}613\ DCIDNA12 - \\ & - 0{,}406\ DCIDNA3 + 0{,}457\ DCIDNA7 \end{aligned} \quad (7)$$

A interpretação da equação discriminante demonstra a ponderação atribuída a cada atributo para a separação dos clientes nos grupos. O sinal dos coeficientes de cada uma das variáveis indica o sentido para a classificação do tipo de cliente, sendo um indicativo de uma característica

Tabela 3. Criação de variáveis *dummies* para variável idade.

Variável	*Dummy*	Mau	Bom	Total	Risco relativo	Classe de risco
Até 20 anos	*DIDAD*1	457	406	863	0,38	Péssimo
21 a 25 anos	*DIDAD*2	421	656	1077	0,66	Muito mau
26 a 30 anos	*DIDAD*3	438	629	1067	0,61	Muito mau
31 a 35 anos	*DIDAD*4	344	604	948	0,75	Mau
36 a 40 anos	*DIDAD*5	283	656	939	0,99	Neutro
41 a 50 anos	*DIDAD*6	411	1261	1672	1,30	Bom
51 a 60 anos	*DIDAD*7	206	1035	1241	2,14	Excelente
Acima de 60 anos	*DIDAD*8	143	1108	1251	3,30	Excelente

para um cliente mau o sinal negativo, e de um cliente bom o sinal positivo.

Na medida em que as variáveis explanatórias estão padronizadas no intervalo 0 a 1, o efeito de cada variável pode ser avaliado diretamente pelo valor absoluto do respectivo coeficiente. Pode-se observar que, de acordo com o modelo discriminante, as variáveis que exercem maior influência sobre o risco do crédito são *DIDAD*1, *DIDAD*8, *DGCEPCO*1, *DGCEPCO*7 e *DGPROF*1. As variáveis *DIDAD*1 e *DIDAD*8 correspondem aos extremos da faixa de idades. O modelo indica que o risco de inadimplência é maior para o grupo de idade mais baixa (até 20 anos) e o risco é menor para o grupo de idade mais elevada (acima de 60 anos). As variáveis *DGCEPCO*1 e *DGCEPCO*7 correspondem aos extremos do CEP comercial. Assim, o bairro ou cidade onde as pessoas trabalham influencia o risco de crédito, possivelmente devido a questões econômicas ou culturais que moldam o comportamento das pessoas que trabalham nesses locais. Por fim, *DGPROF*1 corresponde a um dos extremos da variável profissão, reunindo as profissões que apresentam maior risco de inadimplência.

Para o emprego do modelo encontrado na previsão do risco de crédito, é necessário verificar os pressupostos da análise discriminante. Com a utilização do método *stepwise* garantiu-se que o pressuposto de fraca multicolinearidade fosse atendido, já que o método prioriza a inclusão de variáveis independentes com alto poder discriminatório e pouco correlacionadas entre si. A suposição de normalidade é verificada com a aplicação do teste de Kolmogorov-Smirnov para uma amostra. Este teste compara as frequências observadas com as esperadas segundo a distribuição normal. Quanto maior o valor do teste, maior é a distância entre a distribuição dos dados e a curva normal. O valor obtido no teste foi de 0,031 e significante (p = 0,000) o que leva a decisão de rejeitar a hipótese de ajuste à distribuição normal. Porém esse resultado pode ser explicado pela alta sensibilidade do teste, em função do grande tamanho da amostra. Ainda assim, o ajuste aproximadamente normal pode ser observado no formato do histograma apresentado na Figura 7.

O último pressuposto a ser verificado é o da homogeneidade de matrizes de variâncias e covariâncias dos grupos bom e mau. Tal premissa, necessária para a utilização da técnica, é verificada pelo teste denominado Box's M. Novamente, apesar de não existirem diferenças importantes entre as matrizes de variâncias e covariâncias dos grupos bom e mau (a razão média entre a maior e a menor variância era, em sua maioria, inferior a 1,6), o teste rejeitou a hipótese de homogeneidade (Hair et al., 2005, salientam que tal teste é extremamente sensível ao tamanho da amostra). Como as

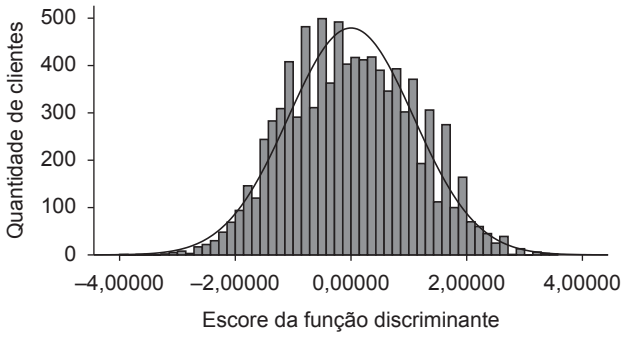

Figura 7. Distribuição dos escores da função discriminante.

Quadro 2. Especificação das variáveis utilizadas nos modelos.

Y = propensão de vir a ser um bom cliente	
*DIDAD*1 = idade até 20 anos	*DGCEPRE*3 = CEP residencial com mau desempenho
*DIDAD*23 = idade entre 21 e 30 anos	*DGCEPRE*5 = CEP residencial com bom desempenho
*DIDAD*4 = idade entre 31 e 35 anos	*DGCEPRE*6 = CEP residencial com muito bom desempenho
*DIDAD*6 = idade entre 41 e 50 anos	*DGCEPRE*7 = CEP residencial com excelente desempenho
*DIDAD*7 = idade entre 51 e 60 anos	*DGCEPCO*1 = CEP comercial com péssimo desempenho
*DIDAD*8 = idade superior a 60 anos	*DGCEPC*56 = CEP comercial com bom ou muito bom desempenho
DSEXOF = é do sexo feminino	*DGCEPCO*7 = CEP comercial com excelente desempenho
DPRIM = possui escolaridade primária (fundamental)	*DGPROF*1 = profissão com péssimo desempenho
DSUP = possui curso superior	*DGPROF*2 = profissão com muito mau desempenho
DCASADO = é casado	*DGPROF*5 = profissão com bom desempenho
*DTSERV*6 = tempo de serviço entre 37 e 60 meses	*DGPROF*67 = profissão com muito bom ou excelente desempenho
*DTSERV*7 = tempo de serviço entre 61 e 90 meses	*DCIDNA*12 = cidade de nascimento com péssimo ou muito mau desempenho
*DTSERV*67 = tempo de serviço entre 37 e 90 meses	*DCIDNA*3 = cidade de nascimento com mau desempenho
*DTSERV*89 = tempo de serviço superior a 90 meses	*DCIDNA*7 = cidade de nascimento com excelente desempenho
DFILHO = tem filhos	
DRES_ALU = tipo de residência alugada	
*DGCEPR*12 = CEP residencial com péssimo ou muito mau desempenho	

diferenças não eram relevantes (do ponto de vista prático), decidiu-se continuar a análise, acreditando-se que ela não ficaria prejudicada.

Para o modelo logístico, 29 variáveis *dummies* foram significativas para compor o modelo final, como é apresentado na Equação 8, que retorna a probabilidade de um proponente vir a ser um bom cliente.

$$
\begin{aligned}
Y = 1 / [+ \exp (&0{,}876 - 0{,}829\ DIDAD1 - \\
&- 0{,}409\ DIDAD23 - 0{,}252\ DIDAD4 + \\
&+ 0{,}232\ DIDAD6 + 0{,}644\ DIDAD7 + \\
&+ 1{,}047\ DIDAD8 + 0{,}327\ DSEXOF - \\
&- 0{,}287\ DPRIM + 0{,}270\ DSUP + \\
&+ 0{,}410\ DCASADO + 0{,}340\ DTSERV6 + \\
&+ 0{,}627\ DTSERV7 + 0{,}792\ DTSERV89 - \\
&- 0{,}293\ DFILHO - 0{,}547\ DRES\_ALU - \\
&- 0{,}392\ DGCEPR12 - 0{,}172\ DGCEPRE3 + \\
&+ 0{,}197\ DGCEPRE5 + 0{,}328\ DGCEPRE6 + \\
&+ 0{,}608\ DGCEPRE7 - 0{,}768\ DGCEPCO1 + \\
&+ 0{,}218\ DGCEPC56 + 0{,}472\ DGCEPCO7 - \\
&- 0{,}718\ DGPROF1 - 0{,}318\ DGPROF2 + \\
&+ 0{,}283\ DGPROF67 - 0{,}449\ DCIDNA12 - \\
&- 0{,}328\ DCIDNA3 + 0{,}592\ DCIDNA7)]
\end{aligned}
\qquad (8)
$$

As variáveis com sinais positivos revelam associações com ser bom pagador e as de sinais negativos com ser mau pagador. Ou seja, um proponente que tem idade acima de 41 anos (*DIDAD*6), tem curso superior (*DSUP*), é casado (DCASADO), entre outras, tem maior probabilidade de ser um bom pagador.

No que diz respeito à verificação de atendimento de suposições para utilização da técnica, somente a ausência de multicolinearidade deve ser confirmada. Com a utilização do método *stepwise* para escolha das variáveis independentes para compor o modelo, garantiu-se o atendimento deste pressuposto.

Analisando a Equação 8, pode-se observar que as mesmas variáveis que dominam o modelo discriminante também dominam o modelo de regressão logística. A única exceção é *DTSERV*89, que tinha aparecido numa posição intermediária no modelo discriminante, mas, no modelo de regressão logística, aparece como uma das mais influentes. Essa variável indica os clientes que possuem tempo de serviço maior do que 90 meses. Uma vez que o sinal é positivo, o modelo considera que as pessoas que estão em seus empregos há mais tempo possuem menor risco de inadimplência.

O *software* utilizado para a construção das redes neurais não tem nenhum tipo de método como o *stepwise*, utilizado na obtenção dos modelos anteriores, que seleciona automaticamente as variáveis com maior poder de explicação. Portanto, utilizou-se para o estudo das redes as variáveis que foram indicadas previamente pelos modelos discriminante e logístico, e também por um teste exploratório feito com a regressão linear.

Para a obtenção das redes neurais utilizou-se a função de ativação sigmoide e o algoritmo de aprendizado supervisionado de retropropagação de erro, com somente uma camada escondida. Várias redes foram criadas com diferentes quantidades de neurônios na camada escondida para verificar o desempenho quanto à predição dos bons e maus clientes. Para avaliar o desempenho das redes compararam-se os valores do erro quadrático médio e do teste KS para as amostras de análise e de teste. Os resultados das melhores redes construídas são apresentados na Tabela 4.

O modelo neural escolhido, com melhor desempenho para previsão do risco de crédito foi o modelo RN6. O modelo RN4 teve melhores resultados para a amostra de análise, porém o desempenho na amostra de teste é menor, evidenciando o excesso de encaixe da rede.

Uma avaliação conjunta dos três modelos (discriminante, logístico e redes neurais) é apresentada na Figura 8 com a distribuição dos bons e maus pagadores e a taxa de sinistro, que corresponde ao percentual de maus pagadores sobre o total de clientes.

Analisando o comportamento das curvas de distribuição dos bons e maus pagadores, verifica-se que os modelos conseguem separar os dois grupos de clientes, já que é possível observar a tendência de que os maus se concentram à esquerda da escala e os bons à direita. A queda na taxa de sinistro, à medida que aumenta o valor dos escores, também é um reflexo da separação obtida pelo modelo.

Para o emprego do modelo encontrado na previsão do risco de crédito com clientes da empresa é necessário verificar os pressupostos da análise discriminante. Com a utilização do método *stepwise* garantiu-se que o pressuposto da inexistência de multicolinearidade fosse atendido, já que o método prioriza a inclusão de variáveis independentes com alto poder discriminatório e também que sejam menos correlacionadas entre si.

As suposições de normalidade multivariada e de homogeneidade de variâncias da análise discriminante não são atendidas. Porém, Hair et al. (2005) salientam que os testes utilizados para verificar tais premissas são muito sensíveis ao tamanho da amostra. Por essa razão, a análise do modelo obtido para fins de previsão não fica prejudicada, já que o objetivo aqui é comparar o poder de predição dos modelos encontrados.

Tabela 4. Comparação dos melhores modelos neurais construídos.

Modelo	Nº neurônios camada oculta	KS	
		Análise	Teste
RN1	35	38,3	34,0
RN2	30	39,2	33,9
RN3	30	39,3	35,1
RN4	35	41,0	32,7
RN5	27	40,3	34,6
RN6	35	40,3	35,4

4.5 Escolha do modelo

Dois métodos adequados para verificar o poder de previsão dos modelos construídos são o percentual de classificações corretas (% bons classificados como bons + % maus classificados como maus) e o valor do teste KS (% de separação entre as distribuições acumuladas dos bons e maus pagadores). As duas medidas são avaliadas tanto na amostra de análise, utilizada para o desenvolvimento do modelo, como na amostra de teste, necessária para garantir que o modelo seja adequadamente utilizado para previsão. Os resultados das medidas para os três modelos são apresentados na Tabela 5.

Tanto na amostra de análise como na amostra de teste, os percentuais de acerto total encontrados para os três modelos são superiores a 65% e os valores para o teste KS são maiores que 30, valores mínimos para considerar um modelo com bom poder de separação.

Ao analisar os valores das duas medidas, observa-se que o resultado obtido pela regressão logística foi praticamente idêntico ao da análise discriminante. Já a separação obtida com o modelo neural foi ligeiramente superior aos demais modelos, o que pode ser explicada por sua abordagem diferenciada no relacionamento das variáveis.

4.6 Passos para implantação

Para programar o modelo de previsão de risco de crédito obtido no sistema da empresa, é necessário obedecer algumas diretrizes. É importante que as variáveis que compõem o modelo final tenham preenchimento obrigatório, garantindo sua devida ponderação para análise do risco do cliente.

A determinação do ponto de corte é um passo mais gerencial que metodológico, já que envolve questões importantes para a empresa como taxa de aprovação e proporção de clientes bons negados. Para nortear essa decisão, pode-se fazer uma análise de sensibilidade de ponto de corte. Desta forma é possível avaliar e escolher o melhor ponto de corte, levando em consideração o percentual de clientes bons que estariam sendo negados e de maus que

Tabela 5. Medidas de desempenho dos modelos construídos.

Modelo	Percentual de acerto		Valor de KS	
	Amostra de análise	Amostra de teste	Amostra de análise	Amostra de teste
Discriminante	73,2	72,1	36,7	30,7
Logístico	73,3	72,2	36,9	31,7
Neural	74,8	72,7	40,3	35,4

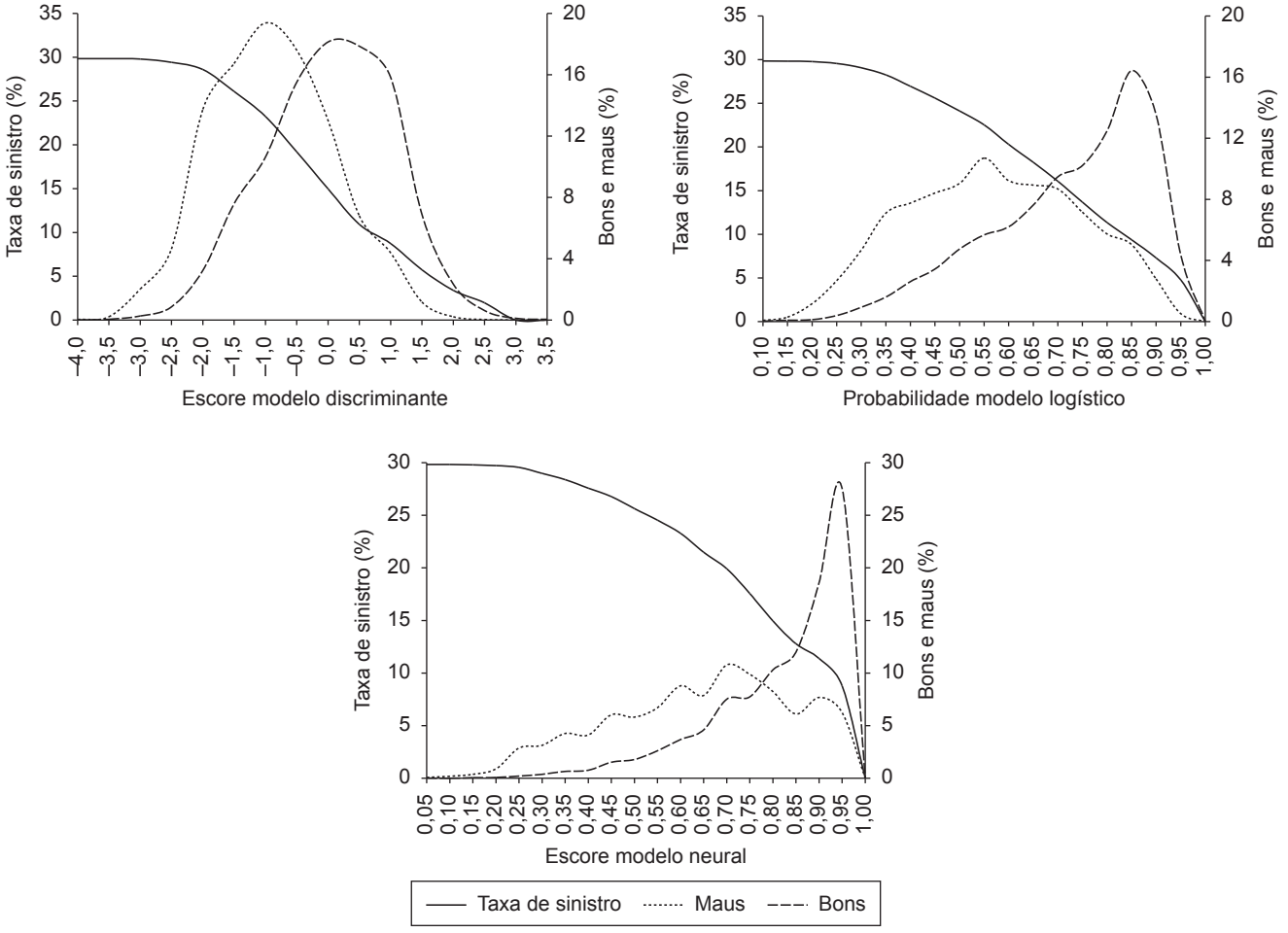

Figura 8. Taxa de sinistro e separação dos modelos construídos.

estariam sendo aprovados, bem como o percentual total de acerto do modelo.

Outra forma de utilizar os escores obtidos com o modelo de previsão de risco de crédito é o gerenciamento da concessão do crédito. Como as pontuações dão um ordenamento dos clientes de maior risco até os de menor risco, é possível utilizá-las para, por exemplo, decidir pela aprovação ou liberação de valores menores de limite, escolher clientes para participar de alguma promoção, flexibilizar a exigência por alguns documentos, entre outros.

Quanto aos cuidados na utilização do modelo, é importante que um proponente a crédito seja submetido a apenas uma análise pelo modelo, de maneira a garantir o sigilo e a eficiência da fórmula, ou seja, devem-se evitar simulações ou testes com o modelo. Neste sentido, recomenda-se também o acesso restrito às informações que compõem a fórmula e seus pesos, bem como o ponto de corte, ao menor número de pessoas possível, evitando-se fraudes no processo de crédito.

Para que o modelo seja adequadamente empregado como ferramenta de decisão, são necessários relatórios periódicos para acompanhamento de sua adequada utilização. O perfil dos clientes que estão solicitando crédito deve ser confrontado com o perfil da amostra de desenvolvimento do modelo, não devendo haver alterações significativas. Alterações de perfis devem ser analisadas cuidadosamente, pois podem evidenciar o mau uso do modelo ou a necessidade de reavaliação.

5 *Conclusão*

Este trabalho apresentou um modelo de previsão de risco de crédito, contemplando a comparação das técnicas de análise discriminante, regressão logística e redes neurais. As características especiais de cada uma das técnicas foram analisadas com o objetivo de encontrar o modelo com melhor desempenho na predição de bons e maus clientes.

Todos os passos para a obtenção de um modelo de previsão de risco de crédito foram apresentados, desde a definição da população até os passos para implementação no sistema da empresa. O nível de detalhe apresentado em cada etapa do método é inédito em trabalhos deste tipo. Assim, o Modelo PRC, desenvolvido neste estudo, pode servir como instrumento de apoio para pesquisadores de empresas construírem seus modelos adaptados.

Os três modelos construídos tiveram desempenhos satisfatórios na previsão dos clientes como bons e maus pagadores, obtendo-se 73,2%, 73,3% e 74,8% de acerto na classificação com as técnicas de análise discriminante, regressão logística e redes neurais, respectivamente. Desta forma, observou-se uma superioridade das redes neurais em relação às outras técnicas, explicada por sua abordagem não linear na combinação das variáveis.

O modelo neural teve melhor separação na previsão que os outros dois modelos, verificado também pelo maior valor para o teste KS (40,3). Porém sua programação no sistema da empresa pode ser considerada complexa, o que poderia levar a empresa, numa decisão estratégica, a escolher o modelo discriminante ou o modelo logístico.

Os resultados obtidos com a análise discriminante e com a regressão logística foram muito semelhantes em termos de desempenho, também evidenciados pelos valores do teste KS, 36,7 e 36,9, respectivamente. Desta forma, uma escolha entre os dois modelos seria dada pela análise dos pressupostos para utilização das técnicas e, neste caso, o modelo logístico apresenta algumas vantagens, pois possui um menor número de pressupostos a serem atendidos.

A utilização dos modelos de previsão de risco de crédito elimina a subjetividade da análise, criando um procedimento padronizado de decisão, que pode ser complementado com informações extras que não estejam contempladas no modelo matemático. Desta forma, é possível aumentar a velocidade da análise de crédito, o que pode permitir o aumento do número de clientes.

Como conclusão final, este estudo confirma que há diferentes técnicas que podem ser utilizadas para o tratamento de dados e predição do pagamento de um crédito concedido. Cada técnica tem suas características e pressupostos que devem ser avaliados para que o modelo construído possa efetivamente ser utilizado pela empresa para fazer previsões do risco de crédito.

Methodology for the construction and choice of credit risk prediction models

Abstract

Due to the growing consumer credit market and, therefore, insolvency indices, companies are seeking to improve their credit analysis by incorporating objective judgments. Multivariate techniques have been used to construct credit models. These models, based on consumer registration information, allow the identification of behavior standards concerning insolvency. The objective of this work is to propose a methodology for the construction of credit risk models and to evaluate prediction performance using three specific models: discriminant analysis, logistic regression, and neural networks. The proposed method (entitled PRC Model) embraces six steps: (i) population definition, (ii) sampling, (iii) preliminary analysis, (iv) model development, (v) model selection, and (vi) implementation steps. The PRC Model was applied to a sample of 17,005 customers of an organization which manages its own credit system and controls a pool of drugstores. The results for this specific database show slight superiority of neural networks over the other two techniques, which can be attributed to its non-linear approach when dealing with the combined effect of explanatory variables

Keywords: *Credit analysis. Discriminat analysis. Logistic regression. Neural networks.*

Referências bibliográficas

ANDREEVA, G.; ANSELLA, J.; CROOK, J. Modelling profitability using survival combination scores. **European Journal of Operational Research**, v. 183, n. 3, p. 1537-1549, dec. 2007.

BUENO, V. F. F. **Avaliação de risco na concessão de crédito bancário para micros e pequenas empresas**. Florianópolis, 2003. Dissertação (Mestrado em Engenharia da Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina.

CAOUETTE, J. B.; ALTMAN, E. I.; NARAYANAN, P. **Gestão do Risco de Crédito**. O próximo grande desafio financeiro. São Paulo: Qualitymark, 1999.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise Multivariada**: para cursos de Administração, Ciências Contábeis e Economia. São Paulo: Atlas, 2007.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 3 ed. [S.l.]: Wiley-Interscience, 1998.

GUIMARÃES, I. A.; CHAVES NETO, A. Reconhecimento de padrões: metodologias estatísticas em crédito ao consumidor. **RAE Eletrônica EAESP/FGV**, v. 1, n. 1, jul.-dez. 2002.

HAIR, J. F. et al. **Análise multivariada de dados**. 5 ed. Porto Alegre: Bookman, 2005.

HAYKIN, S. **Redes neurais**: princípios e prática. Tradução de Paulo Martins Engel. 2 ed. Porto Alegre: Bookman, 2001.

HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression**. New York: John Wiley & Sons, 1989.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 5 ed. Upper Saddle River: Prentice Hall, 2002.

KLEINBAUM, D. G. **Logistic regression: a self-learning text.** New York: Springer, 1996.

KOVÁCS, Z. L. **Redes Neurais Artificiais:** Fundamentos e Aplicações. 3 ed. São Paulo: Livraria da Física, 2002. 174 p.

LAWRENCE, J. **Introduction to Neural Networks – Design, Theory and Applications.** 6 ed. Nevada City, CA, EUA: California Scientific Software, 1994.

LEWIS, E. M. **An Introduction to Credit Scoring**. San Rafael: Fair, Isaac and Co., Inc. 1992.

LOESCH, C.; SARI, S. T. **Redes Neurais Artificiais:** Fundamentos e Modelos. Blumenau: FURB, 1996.

MENDES FILHO, E. F.; CARVALHO, A. C. P. L. F.; MATIAS, A. B. Utilização de redes neurais artificiais na análise de risco de crédito a pessoas físicas. In: SIMPÓSIO BRASILEIRO DE REDES NEURAIS, 3, 1996, Recife. **Anais...**

PEREIRA, S. L. G. Na mira do crédito. **GV Executivo**, v. 5, n. 1, p. 31-36, fev.-abr. 2006.

SANTOS, J. **Análise de Crédito**: Empresas e pessoas físicas. São Paulo: Atlas, 2000.

SAUNDERS, A. **Medindo o risco de crédito**: Novas abordagens para value at risk e outros paradigmas. Rio de Janeiro: Qualitymark, 2000.

SCHRICKEL, W. K. **Análise de Crédito**: concessão e gerência de empréstimos. 3 ed. São Paulo: Atlas, 1997.

SELAU, L. P. R. **Construção de Modelos de Previsão de Risco de Crédito**. Porto Alegre, 2008. Dissertação (Mestrado em Engenharia da Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Sul.

SILVA, J. P. **Gestão e análise de risco de crédito**. 4 ed. São Paulo: Atlas, 2003.

SILVA, J. P. Os dois lados do crédito. **GV Executivo**, v. 5, n. 3, p. 68-72, jul.-ago. 2006.

STEINER, M. T. A. et al. Sistemas especialistas probabilísticos e redes neurais na análise do crédito bancário. **Revista de Administração**, São Paulo, v. 34, n. 3, jul.-set. 1999.

SUBRAMANIAN, V.; HUNG, M. S.; HU, M. Y. An Experimental Evaluation of Neural Networks for Classification. **Computers & Operations Research**, v. 20, n. 7, p. 769-782, 1993.

VASCONCELLOS, R. S. **Modelos de Escoragem de Crédito Aplicados a Empréstimo Pessoal com Cheque**. Rio de Janeiro, 2004. Dissertação (Mestrado em Finanças e Economia Empresarial) – Escola de Pós-Graduação em Economia, Fundação Getúlio Vargas.

YU, X.; EFE, M. O.; KAYNAK, O. A general backpropagation algorithm for feedforward neural networks learning. **IEEE Trans. on Neural Networks,** v. 13, n. 1, p. 251–254, 2002.

Sobre os autores

Lisiane Priscila Roldão Selau

Departamento de Matemática e Estatística
Universidade Federal de Pelotas – UFPEL
Rua Gomes Carneiro, 1, Centro, CEP 96010-610, Pelotas – RS
e-mail: lisiane.selau@ufpel.edu.br

José Luis Duarte Ribeiro

Programa de Pós-Graduação em Engenharia de Produção
Universidade Federal do Rio Grande do Sul – UFRGS
Av. Osvaldo Aranha, 99, 5º andar, CEP 90040-020, Porto Alegre, RS
e-mail: ribeiro@producao.ufrgs.br

Agradecimentos. Os autores agradecem à CAPES e ao CNPq a concessão de bolsas que permitiram a realização desta pesquisa.

Recebido em 20/3/2008
Aceito em 18/8/2009