Proceedings of the V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais
pp. 409–414, April 2–5, 2001 - Rio de Janeiro - RJ - Brazil

Reconhecimento de Padrões em Estatística: Uma Abordagem Comparativa

Carlos A. Ferreira[1], José F. Soares[1], Frederico R. B. Cruz[1]
[1]Departamento de Estatística, Universidade Federal de Minas Gerais
Caixa Postal 702, 30123-970 - Belo Horizonte - MG, Brasil
E-mails: {carlosaf,jfsoares,fcruz}@est.ufmg.br

Abstract

The aim of this paper is to present a comparative experimental study concerning the pattern recognition problem applied to statistics. We compare two well established methodologies, that is, the logistic regression and the classification and regression trees, with a compelling one which is based on neural networks. We present comparative results for two databases, one of them composed by binary variables and the other, by categorical variables. Preliminary results seem to indicate a superior performance of the neural network based method over the logistic regression and the classification and regression trees.

1. Introdução

A história do problema de reconhecimento de padrões na engenharia, especialmente em aplicações militares, é antiga, mas, por razões de custos elevados do *hardware* envolvido na aquisição e tratamento dos dados, este problema vinha sendo um assunto para especialistas, não sendo tão amplamente difundido. Entretanto, isso tem mudado e sua aplicabilidade ampliada. Assim, têm surgido sistemas automáticos para substituir especialistas humanos, muitas vezes com desempenho até superior ao deles.

Neste artigo, apresentaremos uma comparação entre as redes neurais artificiais, uma metodologia emergente na área estatística, com duas outras, bem conhecidas e utilizadas, quais sejam a regressão logística e as árvores de classificação e regressão. Esta comparação é feita com dados de controle de processos hospitalares.

Um trabalho de comparação semelhante ao aqui proposto foi realizado recentemente [1], no qual estas metodologias foram aplicadas ao problema de previsão de inadimplência no pagamento de empréstimos contraídos em instituições financeiras na Alemanha. Entretanto, não conseguimos encontrar na literatura um estudo que levasse em consideração tanto respostas binárias, quanto categóricas, como fazemos aqui.

Organizamos o texto da seguinte forma. As metodologias "tradicionais", *i.e.*, a regressão logística e as árvores de classificação e regressão são apresentadas na seção 2. As redes neurais artificiais, na seção 3. Na seção 4, são apresentadas e discutidas as comparações entre as três metodologias. As conclusões e observações finais da seção 5 encerram o artigo.

2. Metodologias "tradicionais"

2.1. Regressão logística

Através da regressão logística é possível estabelecer a relação entre uma variável resposta dicotômica, normalmente representada pelos termos *sucesso* e *fracasso*, e variáveis explicativas categóricas ou contínuas. A variável resposta pode também receber o nome de variável dependente e as variáveis explicativas podem ser chamadas de variáveis independentes, explanatórias ou covariáveis. Matematicamente, o modelo logístico é apresentado a partir da seguinte expressão:

$$\log\left\{\frac{P[Y(\mathbf{x})=1]}{P[Y(\mathbf{x})=0]}\right\} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p, \quad (1)$$

onde $P[Y(\mathbf{x})=1]$ se refere à probabilidade de sucesso para a variável resposta, $P[Y(\mathbf{x})=0]$, à probabilidade de fracasso, x_i, à i-ésima componente do vetor de variáveis explicativas $\mathbf{x}$ e $\beta_1,\beta_2,\ldots,\beta_p$, aos seus respectivos coeficientes. O termo β_0 é denominado intercepto. O índice $\mathbf{x}$, associado à variável resposta Y, explicita a sua dependência em relação ao vetor de variáveis explicativas $\mathbf{x}$.

A idéia básica do modelo logístico consiste em estabelecer uma relação linear entre as variáveis explicativas (ou alguma transformação delas) e a variável resposta. O ajuste do modelo de regressão logística, dada uma amostra de observações independentes, constituídas pelos pares $(y_j,\mathbf{x}_j)$, consiste em estimar os valores dos parâmetros $\beta_0,\beta_1,\ldots,\beta_p$, a partir do **método da máxima verossimilhança** [2]. Em síntese, esse método retorna, para um dado conjunto de observações, valores para os parâmetros desconhecidos, de forma a maximizar a probabilidade de que os dados tenham sido originados da população correspondente.

Para covariáveis dicotômicas, a interpretação dos coeficientes se dá através da **razão das chances** (do inglês, *odds ratio*) [2]. A razão das chances é uma medida que representa o quão mais provável é a variável resposta assumir o valor positivo y=1, para a

covariável com valor x_i=1, do que para x_i=0. Se a variável *y* denotar, por exemplo, a presença (*y*=1) ou ausência (*y*=0) de câncer de pulmão e a covariável x_i caracterizar se uma pessoa é fumante (x_i=1) ou não (x_i=0) fumante, então uma razão de chances igual a 2 indica que o câncer de pulmão é duas vezes mais freqüente entre os fumantes. Maiores detalhes sobre a regressão logística podem ser facilmente encontrados na literatura [2].

2.2. Árvores de classificação e regressão

As árvores de classificação e regressão, CART (do inglês, ***C**lassification **A**nd **R**egression **T**ree*), podem ser consideradas como um modelo de regressão não-paramétrico, que têm por objetivo estabelecer uma relação entre um vetor de variáveis preditoras **x** e uma única variável resposta *y*. É ajustado mediante sucessivas divisões binárias no conjunto de dados, de modo a tornar os subconjuntos resultantes cada vez mais homogêneos, em relação à variável resposta. Essas divisões são convenientemente representadas por uma estrutura de árvore binária, na qual cada nó corresponde a uma divisão em uma covariável particular [3].

Em uma CART, tanto as variáveis explicativas quanto a resposta podem assumir valores contínuos ou não (categóricos). Se a variável resposta for numérica, o modelo recebe o nome de **árvore de regressão**; caso contrário, é tratada como uma **árvore de classificação**. É importante ressaltar que a CART pode ser utilizada como uma alternativa à regressão linear, quando a variável dependente for numérica, e também à regressão logística, se ela for categórica.

As componentes básicas de uma CART são os **nós** (do inglês, *nodes*) e as **regras de divisão** (do inglês, *splitting rules*). Os **nós** estão associados aos subconjuntos resultantes da aplicação de uma **regra de divisão** a determinado conjunto de dados. Normalmente, são representados através de elipses ou retângulos. As elipses são utilizadas para os nós intermediários, enquanto que os retângulos são utilizados para os nós terminais, também chamados de folhas. O primeiro nó de uma árvore é chamado de nó raiz e equivale ao conjunto de dados completo. Os nós gerados pela divisão de um nó já existente recebem o nome de descendentes e o nó que os originou é chamado de ascendente ou pai.

As regras de divisão são representadas por expressões do tipo:

$$\text{idade} < 14.5,$$

caso a covariável considerada para a divisão seja numérica. Para covariáveis categóricas, a representação se dá na forma:

$$x_i \in \{\text{A,B}\}.$$

Por exemplo, se a variável *cor* fosse considerada para uma subdivisão em determinada CART, as seguintes regras de divisão poderiam ser consideradas:

$$cor \in \{\text{branca, negra}\},$$
$$cor \in \{\text{negra, amarela}\},$$
$$cor \in \{\text{amarela}\}.$$

Para covariáveis categóricas, existem 2^{k-1}-1 possíveis divisões, onde *k* corresponde ao número de categorias possíveis para a variável. Geralmente, as implementações computacionais consideram que as regras de divisão são baseadas em apenas uma das covariáveis de cada vez. Isto significa que combinações lineares entre elas não são permitidas. Esta *heurística* se justifica devido ao fato de que, caso combinações fossem permitidas, haveria um número explosivo de possibilidades, tornando o algoritmo tão lento, a ponto de tornar-se sem utilidade prática.

Os modelos baseados em árvores possuem características que têm contribuído para o crescimento de sua popularidade entre a comunidade estatística [4,5], tais como facilidade de interpretação, tratamento de dados ausentes e captura automática de interações entre as covariáveis [6]. Maiores detalhes sobre o processo de construção de uma CART podem ser encontrados facilmente na literatura [5].

3. Redes neurais artificiais

As redes neurais artificiais surgiram na tentativa de implementar computacionalmente a maneira pela qual o cérebro humano trabalha. O cérebro humano possui cerca de 10^{11}–10^{12} neurônios [7], cada qual interligado a outras 10^4 células nervosas [8].

Cada neurônio realiza uma tarefa bastante simples e o tempo necessário para executá-la gira em torno de 10 milisegundos, ao passo que os processadores dos computadores atualmente utilizados necessitam de nanosegundos para realizar operações correspondentes. Isto equivale a dizer que o processamento nas unidades internas de um computador é cerca de 1 milhão de vezes mais rápido que o processamento realizado no cérebro humano!

No entanto, a despeito da baixa velocidade de processamento de seus elementos constituintes, o cérebro possui uma série de características desejáveis para os sistemas artificiais [9]:

- é robusto; células nervosas morrem diariamente sem afetar, significativamente, o desempenho de nosso cérebro;
- é bastante flexível, podendo se ajustar facilmente a um novo ambiente a partir da aprendizagem;
- pode lidar com informações difusas, probabilísticas ou inconsistentes;
- implementa o paralelismo na execução de suas instruções;

- é pequeno, compacto e dissipa muito pouca energia.

As redes neurais artificiais têm sido aplicadas na solução de problemas os mais variados, principalmente no que se refere ao reconhecimento de padrões [7,9]:

- reconhecimento de caracteres escritos à mão;
- identificação de células cancerosas;
- reconhecimento de voz;
- piloto automático de carros;
- reconhecimento de galáxias;
- detecção de fraudes de cartões de crédito;
- reconhecimento de impressões digitais;
- reconhecimento de maus pagadores;
- otimização de problemas de roteamento;
- fabricação de eletrodomésticos.

O ponto de partida para o desenvolvimento dos estudos acerca das redes neurais teve início a partir da publicação do trabalho de McCulloch e Pitts [10]. De forte fundamentação biológica, esse trabalho tinha como objetivo simular, computacionalmente, a maneira pela qual os neurônios do cérebro humano processavam as informações a eles apresentadas.

Matematicamente, cada neurônio de McCulloch e Pitts, Figura 1, recebe entradas binárias (±1) de outros neurônios ($1,x_1,x_2,\ldots,x_p$) ponderadas por um vetor de pesos ($w_0,w_1,w_2,\ldots,w_p$) e "dispara" se essa combinação linear resultar em um valor positivo. O parâmetro w_0 corresponde ao **intercepto** (bias) da combinação linear entre pesos e variáveis independentes.

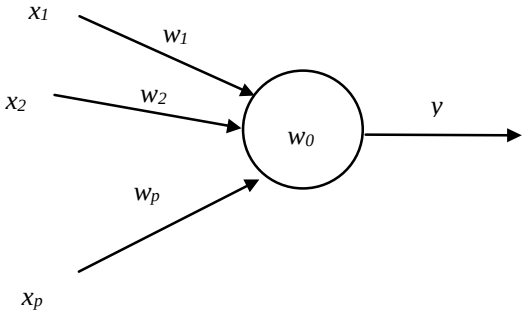

Figura 1: Neurônio de McCulloch e Pitts

Embora simples, o neurônio de McCulloch e Pitts pode ser considerado, sob o ponto de vista computacional, um poderoso instrumento. McCulloch e Pitts provaram que uma combinação de tais neurônios é capaz de realizar qualquer tipo de computação que um computador digital usual pode executar, embora não tão rapidamente [9].

Considerando um *perceptron* de múltiplas camadas, *i.e.*, uma rede neural, organizada em forma de camadas, cujas conexões ocorrem em apenas uma direção, como uma função não linear entre as variáveis de entrada e os pesos entre as conexões, temos que:

$$y_i = f(\mathbf{x}_i, \mathbf{w}). \quad (2)$$

Lembramos, novamente, que, sob o ponto de vista estatístico, os pesos **w** são os parâmetros que devem ser estimados para melhorar a aproximação da rede neural. Ao se apresentar à rede o *i*-ésimo padrão de entrada $\mathbf{x}_i = (x_{1_i}, x_{2_i}, \ldots, x_{p_i})$ e a saída correspondente, t_i (do inglês, *target*), a solução exata para a função (2) seria $y_i=t_i$ (o índice *i* aqui representado se refere a cada uma das saídas da rede). Como isso normalmente não ocorre, uma medida da qualidade da aproximação pode ser dada pela soma dos quadrados dos resíduos:

$$E_i[\mathbf{w}] = (t_i - y_i)^2 / 2. \quad (3)$$

Essa função é minimizada quando $y_i=t_i$ e aumenta à medida em que o valor predito se afasta do valor dado para cada padrão de entrada. Desta maneira, se a rede acertar em todas suas previsões, a função erro retornará 0 e o ajuste dos pesos não se fará necessário.

Quanto menor o valor do erro, melhor a aproximação obtida pela rede. Vários algoritmos para minimização da função E são conhecidos. Entre eles estão o de **Newton-Raphson** e os **algoritmos genéticos** [9]. Entretanto, o mais utilizado é, sem dúvida, o método chamado de *backpropagation of errors* ou, simplesmente, *backpropagation*. Maiores detalhes podem ser encontrados na literatura [11,12].

4. Resultados experimentais

Avaliaremos o erro de generalização pelas seguintes medidas de desempenho:

1) número de elementos incorretamente classificados;
2) somatório dos desvios absolutos entre a variável resposta t_i e o valor retornado pelo modelo y_i:
$$|y_i - t_i|;$$
3) somatório das distâncias Kullbak-Leiber:

$$\text{K - L} = t_i \log\left(\frac{t_i}{y_i}\right) + (1-t_i)\log\left(\frac{1-t_i}{1-y_i}\right), \quad (4)$$

onde y_i representa o valor predito pelo modelo, para o *i*-ésimo padrão de entrada $\mathbf{x}_i$, e t_i, o valor esperado.

4.1. Variável resposta binária

Para comparação entre as metodologias, para variável resposta binária, utilizamos o banco de dados referente às infecções hospitalares associadas aos procedimentos cirúrgicos *revascularizações do miocárdio*, realizados no *Hospital Vera Cruz*, localizado em Belo Horizonte, MG.

O conjunto de dados considerado é composto por 543 pacientes que foram submetidos à cirurgia, no período de jan/95 a dez/97. A Tabela 1 apresenta um resumo das variáveis que o compõem, sendo que a variável resposta é *infecção*.

A variável *contaminação* não é considerada na análise, por todas as cirurgias poderem ser consideradas limpas. O conjunto de dados é dividido em dois subconjuntos, um para treinamento e validação, composto por 362 elementos aleatoriamente escolhidos, outro para testes e verificação do erro de generalização, com 181 indivíduos.

Os detalhes do ajuste dos modelos, bem como as implementações, em *S-Plus*©, e podem ser encontrados no trabalho recente de Ferreira [13].

Tabela 1: Relação das variáveis do banco de dados analisado.

Variável	Tipo	Descrição
sexo	cat.	sexo dos pacientes
idade	num.	idade (anos) dos pacientes
permanência	num.	número de dias em que o paciente esteve internado no hospital antes da realização da cirurgia
cutpoint	cat.	recebe "s" se o cutpoint da cirurgia foi ultrapassado
cec	num.	tempo (horas) da circulação extra-corpórea
geral	cat.	recebe "s" caso o paciente tenha recebido anestesia geral
contaminação	cat.	L – cirurgia limpa PC – potencialmente contaminada C – contaminada I – infectada
trauma	cat.	indica se o paciente trauma pré-operatório
emergência	cat.	indica se o paciente foi internado em situação de emergência
prótese	cat.	indica se o paciente recebeu alguma prótese
profilático	cat.	indica se o paciente recebeu antibiótico profilático
asa	cat.	índice da sociedade americana de anestesiologia que se refere à gravidade do paciente; varia de 1 a 5; quanto maior o asa, maior a gravidade.
infeccao	cat.	indica se o paciente teve infecção hospitalar

Um resumo dos resultados pode ser visualizado na Tabela 2, na qual se nota que todos os modelos falharam na previsão das infecções ocorridas. Isto pode ser explicado pelo fato de as covariáveis utilizadas não conterem muita informação sobre a variável resposta e o evento ser muito raro, o que torna o problema de previsão de infecção hospitalar muito difícil.

Tabela 2: Resumo das medidas de desempenho dos modelos para o conjunto de testes com variável resposta binária

Modelo	Classificações Incorretas	Desvio Absoluto	Distância K-L
regressão logística	33	53,27	84,65
CART	33	54,60	85,97
rede neural	33	51,28	95,63

Com relação ao desvio absoluto, as redes neurais superaram ligeiramente os outros modelos, e o menor somatório das distâncias Kullback-Leiber foi observado para a regressão logística. A distância K-L atribui pesos maiores a erros maiores, razão pela qual as redes neurais saíram perdendo dos demais modelos. Apesar destas terem errado menos, do ponto de vista de vista de erro absoluto, alguns dos erros observados tiveram grandes desvios em relação ao valor conhecido, resultando na maior distância K.L.

4.2. Variável resposta politômica

Utilizamos o banco de dados referentes a 1003 pacientes que foram internados, entre nov/96 a mar/99, no *Hospital Pró-cardíaco*, Rio de Janeiro, RJ, com dor torácica e analisados com relação à possibilidade da ocorrência de infarto de miocárdio e angina instável. A ficha de admissão ao estudo coleta informações sobre as características da dor, história clínica do paciente e alguns exames, de baixo custo, realizado na admissão (p.e., eletrocardiograma, ausculta, pressão).

Após reparametrização, que é um procedimento necessário para a implementação das redes neurais [13], 95 variáveis são obtidas. A Tabela 3 apresenta algumas delas.

Tabela 3: Relação das variáveis reparametrizadas do banco de dados do Hospital Pró-cardíaco.

	Variável	Descrição	Valor	Tipo	Obs.
1	*pc.sexo*	sexo do paciente	m/f	int.	f=1,m=0
2	*pc.idade*	idade (÷ 100)	idade	num.	reparam.
3	*pc.interval*	duração da dor	interval	num.	reparam.
4	*pc.duracao*	se ausente à admissão, qual a duração (min)	duracao	num.	reparam.
5	*pc.loc.ret*	localização da dor	retroesternal	int.	sim=1, não=0
...	...	...	...	...	...
10	*pc.asp.ire*	aspecto da dor	início recente	int.	sim=1, não=0
...	...	...	...	...	...
14	*pc.irr.nao*	irradiação da dor	nenhuma	int.	sim=1, não=0
...	...	...	...	...	...
19	*pc.qua.nao*	qualidade da dor	nenhuma	int.	sim=1, não=0
...	...	...	...	...	...
25	*pc.pre.nao*	precipitação da dor	nenhuma	int.	sim=1, não=0
...	...	...	...	...	...
30	*pc.agr.nao*	agravante da dor	nenhum	int.	sim=1, não=0
...	...	...	...	...	...
37	*pc.ali.nao*	alívio para a dor	nenhum	int.	sim=1, não=0
...	...	...	...	...	...
43	*pc.sin.não*	sintomas da dor	nenhum	int.	sim=1, não=0
...	...	...	...	...	...
90	*pc.tipodor.f*	tipo da dor	A, B, C, D	fator	A = def. anginosa, B= prov. anginosa, C = prov. não anginosa, D = def. não anginosa
91	*pc.di1.iam*	diagnóstico final do investigador	iam	int.	sim=1, não=0
92	*pc.di1.ang*	diagnóstico final do investigador	angina instável	int	sim=1, não=0
93	*pc.di1.ind*	diagnóstico final do investigador	indeterminado	int.	sim=1, não=0
94	*pc.di1.nic*	diagnóstico final do investigador	não ICA	int.	sim=1, não=0
95	*pc.dia.fi*	diagnóstico final do investigador	IAM, angina instável, não ICA	fator	sim=1, não=0

Três análises são realizadas. A primeira tem como objetivo verificar se existe uma uniformidade na identificação do tipo de dor (que acomete o paciente), pelos médicos plantonistas, a partir do questionário preenchido na admissão. Esta resposta intermediária é muito importante, pois vários tratamentos dependem dela. As covariáveis reparametrizadas utilizadas estão identificadas pelos índices 1 ao 54, Tabela 3. Para as

redes neurais, as variáveis resposta são aquelas apresentadas na Tabela 4. Para a CART e a regressão logística, a variável é *pc.tipodor.f*. Estas variáveis correspondem à classificação da dor torácica, identificada na admissão.

Os modelos obtidos são utilizados para a previsão do tipo de dor de determinado paciente recém-admitido ao estudo. O esperado é que pacientes com mesmas características sejam classificados como possuidores do mesmo tipo de dor.

Tabela 4: Classificação da variável resposta para as redes neurais

	Variável	Tipo da Dor
86	*pc.tdd.A*	definitivamente anginosa
87	*pc.tdd.B*	provavelmente anginosa
88	*pc.tdd.C*	provavelmente não anginosa
89	*pc.tdd.D*	definitivamente não anginosa

O segundo problema consiste em verificar se há uniformidade nos diagnósticos do tipo da dor, a partir das covariáveis identificadas pelos índices 1 ao 85, Tabela 3, que se referem às características da dor e *aos resultados dos exames realizados na admissão*. As variáveis resposta são as mesmas utilizadas na primeira análise. O objetivo aqui é verificar se as características da dor (obtidas através dos questionários aplicados na admissão) e os resultados dos exames admissionais são os determinantes para a classificação do tipo de dor do paciente admitido ao estudo.

O terceiro problema é procurar identificar se existe uniformidade para os diagnósticos finais feitos pelos médicos investigadores, a partir das mesmas covariáveis utilizadas na segunda análise. O esperado é que pacientes com mesmas características da dor e com resultados nos exames admissionais semelhantes sejam similarmente diagnosticados. A variável resposta é *pc.dia.fi*, quando utilizamos regressão logística ou CART, e as apresentadas na Tabela 5, no caso de redes neurais.

Tabela 5: Classificação da variável resposta (diagnóstico final) para as redes neurais

	Variável	Tipo da Dor
91	*pc.di1.IAM*	IAM (infarto agudo do miocárdio)
92	*pc.di1.ang*	angina instável
93	*pc.di1.ind*	indeterminado
94	*pc.di1.nic*	não ICA

Para cada análise, o conjunto de dados foi dividido em dois subconjuntos. O primeiro, denominado **conjunto de treinamento e validação**, é composto por 2/3 dos elementos, aleatoriamente selecionados. O segundo, denominado **conjunto de testes**, é composto pelo 1/3 restante.

A análise a partir de redes neurais foi realizada através de uma função em *S-Plus*© [13], que cria perceptrons de apenas uma camada intermediária.

Para o primeiro modelo a ser construído, a variável resposta considerada se refere à dor torácica diagnosticada pelo médico plantonista, quando da admissão do paciente.

Para determinação do número de elementos que a camada intermediária deve possuir, utilizamos o procedimento de validação cruzada.

A Tabela 6 apresenta os resultados, para redes neurais com número de unidades na camada intermediária de 1 a 5.

Tabela 6: Resultados do procedimento de validação cruzada para a primeira análise

Unidades	**Classificações Incorretas**						
1	328	333	328	324	333	338	356
2	340	317	336	301	328	366	361
3	336	327	350	341	309	357	328
4	356	352	337	333	381	342	346
5	306	326	339	357	373	344	346

A análise da Tabela 6 nos sugere uma rede com duas unidades intermediárias, para a verificação do erro de generalização, pois são aquelas que apresentam, em média, os menores erros. Procedimentos similares (não apresentados) foram conduzidos para a segunda e terceira análises, chegando a redes com duas unidades e uma unidade na camada intermediária, respectivamente.

Tabela 7: Comparação entre os métodos

Análise	**Número de Elementos do Conjunto de Testes**	**Percentual de Acertos**		
		Regressão Logística	**CART**	**Redes Neurais**
1	333	37,54	43,84	49,25
2	293	45,39	48,46	46,42
3	293	43,00	42,66	43,69

Não detalharemos aqui as análises via regressão logística e CART, que podem ser encontradas no trabalho de Ferreira [13]. Passemos imediatamente ao resumo dos resultados. Pela Tabela 7, pode-se notar que as redes neurais mostraram-se superiores aos demais métodos, para as primeiras e terceira análises, sendo inferiores à CART, para a segunda análise.

5. Conclusões e observações finais

Para o conjunto de dados relativos às revascularizações de miocárdio, verificamos que todos os modelos falharam na previsão de infecção hospitalar, ou seja, todos previram a não-ocorrência de infecção para pacientes infectados (33 pacientes), Tabela 2. A conclusão a que chegamos é que as variáveis explicativas utilizadas na construção do modelo não contêm informações suficientes para explicar a variável resposta. Isto é equivalente a dizer que as variáveis disponíveis não são boas preditoras para este evento raro.

Os resultados apresentados na Tabela 2 apontam para uma superioridade das redes neurais em relação ao somatório dos desvios absolutos. Entretanto, esta metodologia teve o pior desempenho ao considerarmos o somatório das distâncias *Kullback-Leiber*. Conforme discutido anteriormente, esta distância penaliza erros de maior amplitude.

Para o banco de dados cuja variável resposta é politômica, as redes neurais se mostraram superiores aos demais modelos, para as primeiras e terceira análises, sendo inferiores à CART, para a segunda análise, conforme apresentado na Tabela 7.

A regressão logística é uma metodologia normalmente utilizada quando o objetivo é identificar os fatores de risco que possuem associação com a variável resposta [2]. Raramente é utilizada como ferramenta de previsão. Neste trabalho, contudo, para fins de comparação com a CART e as redes neurais, ela teve de ser avaliada neste aspecto. A vantagem desta metodologia em relação às demais aqui consideradas é o fato de haver uma teoria matemática bastante sólida (teoria da máxima verossimilhança) por trás de sua fundamentação, o que nos permite obter estimadores ótimos para os coeficientes das covariáveis analisadas. De posse destes estimadores, podemos obter intervalos de confiança e realizar testes de hipóteses para verificar a importância de cada uma das variáveis explicativas consideradas.

A CART é normalmente utilizada para identificar grupos homogêneos, sendo também utilizado na previsão de valores para a variável resposta considerada. Esta metodologia pode, a princípio, ser utilizada para a determinação de fatores de risco, uma vez que as variáveis explicativas presentes na árvores selecionada podem ser consideradas como significativas. Além disto, o tamanho do degrau relativo à divisão a partir de determinada covariável pode ser utilizada para medir a sua importância. Infelizmente, este método, considerado não paramétrico, não permite a formulação de testes de hipóteses e construção de intervalos de confiança.

Finalmente, as redes neurais podem ser consideradas como modelos não paramétricos e são utilizadas apenas nos casos nos quais o objetivo é prever o valor da variável resposta, não sendo indicadas para as situações nas quais o interesse está na identificação de fatores de risco. Em se tratando de previsão, elas têm demonstrado ser uma metodologia bastante eficiente, conforme corroborado pelos testes aqui apresentados, bem como em outros trabalhos [14].

Um observação importante é que, para avaliar a eficiência de um modelo ajustado por uma rede neural do tipo perceptron de múltiplas camadas, não há outra maneira que não verificar seu poder de generalização. Em termos práticos, isto significa medir seu desempenho por consideração de um conjunto de testes independente, supostamente advindo da mesma população que gerou o conjunto de treinamento.

Em trabalhos futuros, novos bancos de dados deveriam ser analisados, com diferentes características nas covariáveis preditoras, e comparações entre as metodologias deveriam novamente ser realizadas.

Agradecimentos

A pesquisa de Frederico Rodrigues Borges da Cruz é financiada parcialmente pelo CNPq, processo 301809/96-8, FAPEMIG, processo CEX-855/98, e pela PRPq-UFMG, aos quais ficam os seus agradecimentos.

Referências

[1] G. Arminger *et al*. Analyzing credit risk data: a comparision of logistic discrimination, classification tree analysis and feedfoward networks. *Computational Statistics*, 12(2):293-310, 1997.

[2] D. W. Hosmer e S. Lemeshow. *Applied Logistic Regression*, John Wiley & Sons, New York, 1989.

[3] J. F. Bell. Application of classification trees to the habitat preference of upland birds, *Journal of Applied Statistics*, 23(2-3):349-359, 1996.

[4] L. Breiman *et al*. *Classification and Regression Trees*. Pacific Grove, California: Wadsworth and Brooks, 1984.

[5] L. A. Clark. e D. Pregibon. Tree-based models. In *Statistical Models in S* (Eds. J. M. Chambers e T. J. Hastie), Pacific Grove, California: Wadsworth & Brooks/Coal Advanced Books e Software, pages 317-419, 1992.

[6] M. R. Segal. Tree-structured methods for longitudinal data. *Journal of the American Statistical Association - Theory and Methods*, 87(418):407-418, 1992.

[7] B. Cheng e D. M. Titterington. Neural networks: a review from a statistical perspective. *Statistical Science*. 9(1):2-54, 1994.

[8] B. D. Ripley. Neural Networks and Flexible Regression and Discrimination. In: *Statistics and Images II* (Ed. K.V. Mardia), Carfax, Abingdon, pages 39-57, 1994.

[9] J. Hertz *et al*. *Introduction to the Theory of Neural Computation*. Addison-Wesley Publishing Company, New York, 1995.

[10] W. S. McCulloch e W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulleting of Mathematical in Biophysics*, 5:115-133, 1943.

[11] K. Gurney. *Multilayer nets and backpropagation*. Department of Human Sciences. Brunel University, Uxbridge, Middx, UK, 1997.

[12] L. Smith. *An Introduction to Neural Networks*. Centre for Cognitive and Computational Neuroscience. Department of Computing and Mathematics. University of Stirling, UK, 1996.

[13] C. A. Ferreira. Comparação da Capacidade Preditiva da Regressão Logística, CART e Redes Neurais. *Dissertação de Mestrado*, Departamento de Estatística–ICEx–UFMG, Belo Horizonte, 1999.

[14] R. A. L. Faraco, P. P. Costa Jr., F. R. B. Cruz. Minimização do erro no algoritmo back-propagation aplicado ao problema de manutenção de motores. *Pesquisa Operacional*, 18(1): 22-36, 1998.