Recuperação de Dados Ausentes Através de Redes Neurais Artificiais - Estudo de Caso para uma Base de Dados Mercadológica

Luis E. Zárate, Bruno M. Nogueira e Tadeu R. A. Santos
Laboratório de Inteligência Computacional Aplicada
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais
Email: zarate@pucminas.br, bmnogueira@gmail.com, tadeu@javafree.com.br

***Resumo*— Dados ausentes em bancos de dados são hoje considerados um dos maiores problemas enfrentados na aplicação de *Data Mining*. No tratamento destes dados é necessário que as características do banco sejam preservadas, ou seja, que não haja informação perdida nem adicionada sem uma análise mais cuidadosa. O objetivo deste trabalho é mostrar como as Redes Neurais Artificiais junto com o conhecimento tácito do especialista no domínio, podem ajudar a recuperar informações dos atributos ausentes. Neste trabalho, esses dois elementos são combinados para recuperar dados ausentes numa base de dados mercadológicos.**

I. Introdução

Atualmente, KDD (*Knowledge Discovery in Data Base*) [1], onde o *Data Mining* está inserido, vem sendo aplicado aos mais diversos segmentos científicos e de mercado. Como exemplos, podem ser citadas as áreas industrial, financeira, de saúde, telecomunicações, de negócios entre outras, sempre com a mesma finalidade, a descoberta de conhecimento não óbvio e o auxílio para tomada de decisão.

Os dados sobre os quais é aplicado o processo KDD freqüentemente possuem dados ausentes ocasionados por circunstâncias não controladas. Entende-se por dados ausentes aqueles cujos valores não foram adicionados à base de dados, mas para os quais existe um valor real no meio do qual foram extraídos. A presença de valores ausentes em uma base de dados é um fato comum, podendo estar distribuído em diversos atributos, numa mesma instância (registro) ou de forma aleatória. Valores ausentes podem gerar sérios problemas na extração de conhecimento e na aplicação dos algoritmos de *Data Mining*.

Durante o processo da Descoberta de Conhecimento numa base de dados, um procedimento muito comum para lidar com dados ausentes consiste em eliminar o(s) atributo(s) ou a(s) instância(s) da base de dados que apresentam esses valores, impondo, desta forma, restrições ao conhecimento extraído. Outros procedimentos sugerem a substituição de valores ausentes por valores padrões ou valores médios em todas as ocorrências.

A eliminação de instâncias e/ou atributos pode acarretar também na perda de informações importantes relativos aos valores que estão presentes. Além disso, a substituição por valor padrão, mesmo o mais criterioso, pode introduzir na base informações distorcidas, que não estão contidas no evento e nas circunstâncias que a geraram [2] . A recuperação de dados ausentes torna-se, então, um ponto de extrema importância na descoberta de conhecimento em base de dados, requerendo predições cuidadosas dos valores, utilizando técnicas mais avançadas e elaboradas, além do conhecimento tácito de um especialista no domínio do problema [3]. Todas elas, em seu conjunto, visam a não distorção das informações.

Embora um grande número de técnicas usadas em *Data Mining* não lide com dados que contenham valores ausentes, existem outras que sobrelevam este problema em diferentes graus. Técnicas como classificador por vizinho mais próximo *nearest neighbor*, classificadores bayesianos e diversas técnicas estatísticas, não conseguem lidar com conjunto de dados com valores ausentes, tornando seu uso inviável para determinadas bases de dados [4]. Por outro lado, técnicas convencionais que lidam com bases de dados contendo pequeno número de valores ausentes, como árvores de decisão, podem ser utilizadas na tentativa de se retirar conhecimento dessas bases, porém experiências mostram que estas não apresentam resposta satisfatória quando o número de dados ausentes é muito grande.

Uma possível solução para substituição dos valores ausentes seria a utilização de métodos de aprendizado de máquinas. Estes são mais eficientes que os métodos estatísticos para tal fim, embora consumam maior tempo de processamento devido à sua maior complexidade. Dentre esses métodos encontram-se as Redes Neurais Artificiais (RNA) que conseguem aprender relações entre variáveis a partir das instâncias que lhe são mostradas. Associando fatores como capacidade de generalização das redes neurais e o conhecimento tácito do domínio de problema pelo especialista, essas redes podem ser utilizadas para predizer o valor dos atributos ausentes. Neste trabalho, esses fatores foram utilizados em conjunto a fim de recuperar dados ausentes numa base de dados mercadológicos.

O objetivo deste trabalho é apresentar uma aplicação de redes neurais para recuperação de informação em bases dados com massivos dados ausentes, oriundos do mercado varejista do ramo têxtil. Os dados são resultados de pesquisa de

mercado que procura identificar o perfil do setor. Por diversos motivos, vários campos da pesquisa considerada deixaram de ser respondidos, gerando uma base de dados com grande quantidade de dados ausentes. Dentre esses dados, verificou-se que aqueles relativos ao faturamento anual das lojas encontravam-se com grande índice de ausência (aproximadamente 33%). Dada a importância da informação deste campo para obtenção do perfil, tomou-se como objetivo deste trabalho estimar os intervalos de faturamento aos quais seus valores pertencem.

II. Base de Dados Considerada

A base de dados utilizada neste trabalho é oriunda de uma pesquisa de campo realizada junto a comerciantes varejistas do ramo têxtil, contendo dados ausentes em grandes proporções (Ver Figura 1, onde cada símbolo representa ausência de dados para um dos 71 atributos em uma amostra de 100 registros da base de dados). Em sua etapa inicial, cada registro considerado é uma loja pesquisada (inicialmente 634) e cada atributo é equivalente a um campo da pesquisa (totalizando 71). A descrição dos principais atributos da base de dados pode ser observada na Tabela I.

TABELA I

Alguns dos principais atributos da base de dados

Atributo	Descrição
<nom_com>	Nome comercial da loja
<endereço>	Endereço da loja
<sexo_prop>	Sexo do proprietário
<nome_prop>	Nome do proprietário
<atrativos_região>	Atrativos da região de localização da loja
<classe_cli>	Classe social dos clientes
<faturamento>	Faturamento anual bruto

Devido à grande quantidade de atributos, foram necessárias a filtragem e a seleção dos mais relevantes ao domínio do problema. Contudo, esta é uma tarefa que requer extrema cautela, uma vez que, se feita incorretamente, pode acarretar a perda de informações valiosas para a correta interpretação dos dados. Nesta seleção, com a ajuda de especialistas (conhecimento tácito), separou-se os atributos em grupos quanto à natureza de sua informação [5]: fatos e julgamentos.

Fatos são aqueles atributos cuja importância para a análise dos dados é considerada alta por fornecerem informações essenciais, enquanto que julgamentos são atributos de baixa importância, dos quais não se extrairiam informações relevantes para a análise dos dados, podendo, portanto, serem desconsiderados. Na Figura 2, é mostrada uma distribuição do grau de importância da informação contida num atributo, para análise do especialista no domínio do problema. Atributos considerados Fortemente-Fato (FF), Fato (F) e Julgamento-Fato (JF) serão considerados necessários, enquanto que atributos considerados como Julgamentos (J) e Fortemente-Julgamento (FJ) serão desconsiderados.

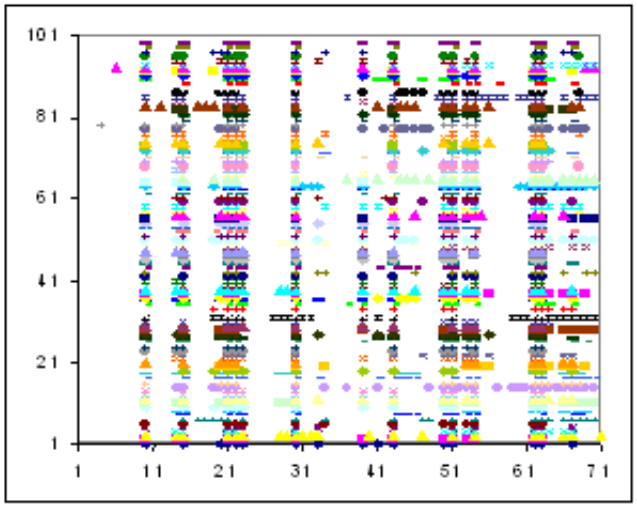

Fig. 1. Atributos x Registros da base de dados com massivos dados ausentes

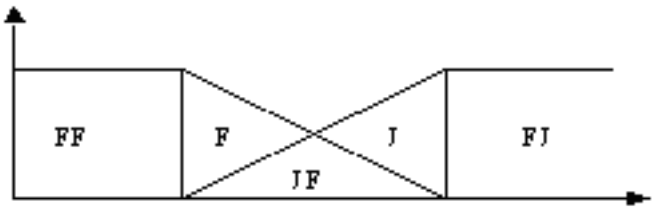

Fig. 2. Classificação de importância de um atributo

A classificação do grau de importância dos atributos faz parte da etapa de seleção do processo de descoberta de conhecimento em bases de dados (KDD).

III. Pré-processamento dos dados

A seguir serão apresentadas as diversas ações para preparação dos dados e seleção dos conjuntos de treinamento.

A. Recuperação de valores

Foi realizada a recuperação de valores de atributos que não estavam presentes na base de dados, mas que se encontravam disponíveis de forma indireta. Foram duas as ações aplicadas.

- Recuperação de valores por atributos relacionados: Por exemplo, a recuperação do atributo <sexo_prop> foi efetuada através do gênero do atributo <nome_prop>. O algoritmo utilizado foi uma adaptação do algoritmo *Shift-Or* para casamento aproximado de padrões [6].
- Recuperação de valores por substituição padrão: atributos não preenchidos que contemplavam valores default foram substituídos por estes.

B. Remoção de atributos

Os critérios adotados para a remoção de atributos são apontados a seguir:

- O primeiro critério foi a remoção de campos irrelevantes ao domínio do problema (por exemplo: <nom_com>,

<nom_prop>, etc). No total foram eliminados cinco atributos.

- O segundo critério adotado foi a eliminação de atributos com valores com pouca informação. Para isso foi aplicado o conceito de entropia [7].

$$H(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \, log_2 \, p_i \qquad (1)$$

onde: p_i é a probabilidade de ocorrência de um valor s_i para um atributo. Assim, para campos com valores praticamente constantes tem-se: $p_i \approx 1$, logo $H(s_i) \approx 0$. Foi eliminado um total de sete atributos.

- Atributos com grandes quantidades de valores ausentes não ajudam a caracterizar o domínio do problema. Portanto, foi determinado um limiar de 25% de dados faltantes para remover o atributo. Um único atributo <classe_cli> foi removido.
- Com auxílio do especialista foram classificados os atributos conforme a relevância com o domínio do problema. A remoção foi efetuada de acordo ao seguinte algoritmo:

```
PARA todos os atributos FAÇA
  SE inf_atrib == Julgamento OU
     inf_atrib == Fortemente-Julgamento
  ENTÃO remova atributo
```

Foi removido um total de 27 atributos.

C. Remoção de registros

Assim como atributos, registros com grande quantidade de dados ausentes foram eliminados. Estabeleceu-se um limiar de até 25% de valores ausentes para eliminar um registro. Um total de 46 registros foi retirado da base de dados.

D. Transformação dos atributos

Como a base de dados apresenta atributos dicotômicos, nominais, categóricos e ordinais, foi necessária a transformação destes para a forma numérica para que fossem apresentadas à RNA. As seguintes transformações foram aplicadas aos respectivos atributos:

- Atributos que já eram separados por faixas de valores no formulário de pesquisa ou que nele constituíssem apenas a marcação simples de uma das opções, tiveram números arbitrários associados a estas faixas/opções. Para o <faturamento>, foram estabelecidos 4 intervalos, Eq. 2.

$$\begin{aligned} \text{Faturamento} &\leq 61000 \in Fat_1 \\ 61000 < \text{Faturamento} &\leq 123000 \in Fat_2 \\ 123000 < \text{Faturamento} &\leq 377000 \in Fat_3 \\ \text{Faturamento} &> 377000 \in Fat_4 \end{aligned} \qquad (2)$$

- Campos dicotômicos foram substituídos pelo equivalente binário (0 ou 1).
- O dado composto <endereço> foi transformado em dado vetorial (Longitude, Latitude) através do GIS. A fim de possibilitar uma classificação numérica, representativa para os endereços, aplicou-se a técnica de clusterização para agrupar os registros (lojas). Foi aplicado o algoritmo K-Means, definindo a priori 10 clusters. A Figura 3 mostra a localização dos agrupamentos de lojas.

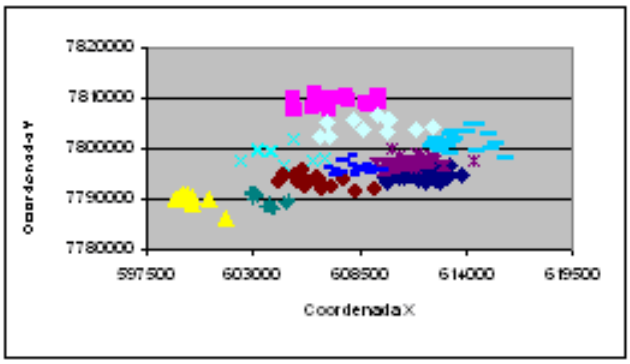

Fig. 3. Agrupamento das lojas

- Para atributos com marcações de múltiplas opções não associadas entre si, por exemplo, <atrativos_região>, foi considerado cada opção como um atributo independente, tornando-o dicotômico (marcação = 1, não marcação = 0). Assim, o número de atributos foi expandido para 81. Para outros casos de atributos com marcações de múltiplas opções associadas entre si, é recomendado o tratamento desses como dados circulares [8].

E. Identificação de dados inconsistentes

Devido à natureza mercadológica da base de dados considerada, é esperado que existam inconsistências no campo <faturamento> devido à falsa informação fornecida. De forma a detectar essas inconsistências nos registros, foi aplicada a técnica de clusterização (*K-Means*) separadamente aos grupos pré-classificados de faturamento.

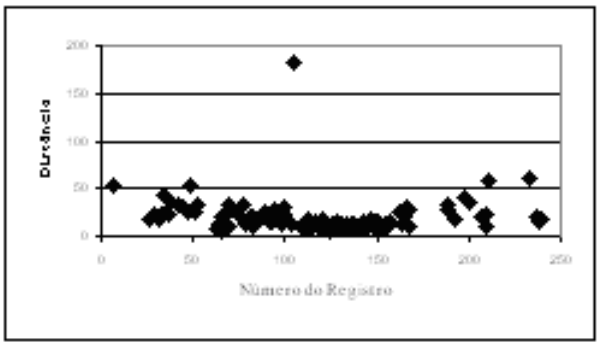

Fig. 4. *Cluster* 1 do Intervalo Fat_1

A Figura 4, mostra a distribuição dos registros do *Cluster* 1 para Fat_1. A Figura 5 mostra as distâncias dos registros aos centróides correspondentes, para o mesmo intervalo. É

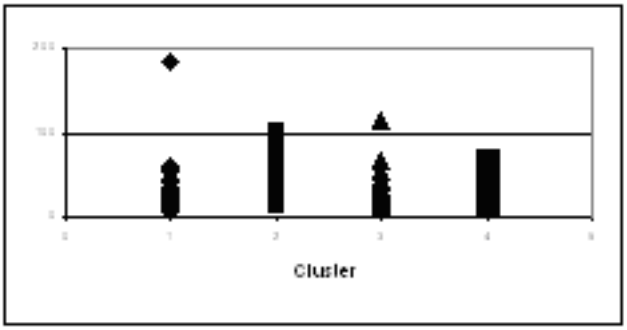

Fig. 5. Intervalo Fat_1 (4 *Clusters*)

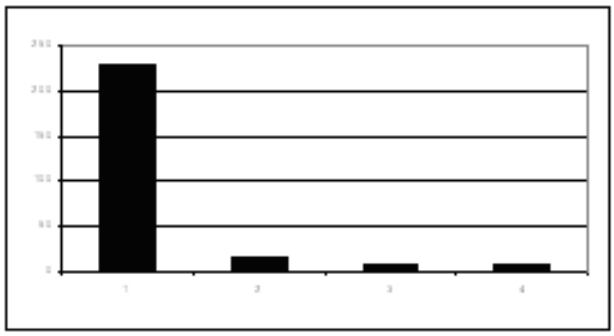

Fig. 6. Intervalo de Faturamento x Freqüência

possível observar em ambas figuras a existência de objetos discrepantes (outliers), os quais foram retirados do grupo de instâncias do intervalo Fat_1.

Os objetos retirados de cada intervalo Fat_1 obedecem às seguintes regras

$$RS_i = \{O_{j_k} \in F_i \mid li_i \leq dist(O_{j_k}, Cent_{k_i}) \leq ls_i\} \quad (3)$$

onde RS_i é o conjunto de registros selecionados para o i-ésimo intervalo de faturamento; F_i é o i-ésimo intervalo de faturamento; O_{j_k} é o objeto j vinculado ao *cluster* k; $Cent_{k_i}$ é o k-ésimo centróide do i-ésimo intervalo de faturamento; li_i é o limite inferior de distância para o i-ésimo intervalo de faturamento; ls_i é o limite superior de distância para o i-ésimo intervalo de faturamento. Para cada intervalo de Faturamento Fat_i, os valores li_i e ls_i s são mostrados na Tabela II.

TABELA II
LIMITES DAS DISTÂNCIAS AOS CENTRÓIDES DOS RESPECTIVOS *clusters*

Fat_i para i =	li_i	ls_i
1	0	60
2	10	20
3	20	50
4	20	30

IV. MONTAGEM DOS CONJUNTOS DE TREINAMENTO

Após o processo de limpeza, restaram 256 registros com 81 atributos. Sendo o objetivo deste trabalho a recuperação da informação acerca do faturamento anual, o atributo <faturamento> foi considerado como a saída da rede, obedecendo, a priori, as faixas de intervalos da expressão 2.

A Figura 6 mostra uma distribuição heterogênea do número de registros para as 4 faixas de faturamento. De forma a evitar uma polarização da rede neural, através de uma escolha aleatória dos conjuntos, foram selecionados 6 registros de cada faixa gerando um total de 24 conjuntos de treinamento. Os demais registros foram reservados para validação da rede.

Devido à natureza do problema, onde a falsidade da informação (faturamento anual) é muito latente, pela experiência do especialista no domínio do problema as faixas foram reduzidas para duas. A regra fundamentada no conhecimento tácito, aplicada neste trabalho é expressa como:

Sejam o conjunto Limites:

$$Limites = \{LimI_i \in \Re, LimS_i \mid LimI_i < LimS_i, \\ i = 1..N, \text{ com } LimI_{i+1} = LimS_i\}$$

e o conjunto de intervalos de faturamento Φ *dado por:*

$$\Phi = \{F_i; LimI_i \leq F_i < LimS_i, i = 1..N\}$$

O valor de faturamento informado é definido por:

$$VInf = \{x \in F_k; p(x \in F_k) = p_k\}$$

onde p_k *é a probabilidade de* $x \in F_k$.
Se $x \in F_k$ *é informação falsa, então*

$$p(LimI_{k+1} \leq x < LimS_n) > p(x \in F_k) > \\ p(LimI_1 \leq x < LimS_{k-1})$$

em outras palavras:

$$p(x < LimS_N) - p(x \leq LimS_{k-1}) > p_k > \\ p(x < LimS_{k-1}) - p(x \leq LimI_1)$$

Portanto é possível reduzir os intervalos de faturamento sem introduzir perda de informação significativa.

O novo conjunto Φ^* de intervalos de faturamento pode ser expresso como:

$$\Phi* = \{F_i; LimI_i \leq F_i < LimS_i, \, i = 1.k, N\} \quad (4)$$

de onde:

$$Faturamento \leq 61000 \in Fat_1 * \\ 61000 < Faturamento \in Fat_2* \quad (5)$$

V. Representação neural do problema

Neste trabalho foram treinadas duas Redes Neurais Multicamadas, constituída por neurônios do tipo *Perceptron* [9] dispostos em uma camada de entrada, uma camada oculta e uma camada de saída, de maneira que todos os neurônios de camadas subseqüentes são interligados. As redes utilizadas possuíam 80 entradas cada uma, com 4 e 2 saídas e 60 neurônios na camada escondida. Como função de ativação foi escolhida a função sigmóide. Para as RNAs, todas as entradas consideradas (representadas pelo conjunto E, Tabela I) foram mapeadas em 4 e 2 saídas binárias distintas e mutuamente exclusivas respectivamente (Eq. 6).

$$f(E) \underrightarrow{RNA}(Fat_1, Fat_2, Fat_3, Fat_4) \tag{6}$$

A. Preparação dos dados para o treinamento

Para treinamento das redes neurais, os dados de entrada E foram submetidos a um processo de normalização:

- A fim de melhorar a convergência do processo de treinamento, o intervalo de normalização foi escolhido como [0.2, 0.8];
- Os dados foram normalizados seguindo as seguintes expressões:

$$f^a(L_0) = L_n = (L_0 - L_{min})/(L_{max} - L_{min})$$
$$f^b(L_n) = L_0 = L_n * L_{max} + (1 - L_n) * L_{min}$$

L_{min} e L_{max} foram computados como segue:

$$L_{min} = L_{sup} - (N_s/(N_i - N_s)) * (L_{inf} - L_{sup})$$
$$L_{max} = ((L_{inf} - L_{sup})/(N_i - N_s)) + L_{min}$$

onde L_{sup} é o valor máximo de uma variável, L_{inf} é o valor mínimo e N_i e N_s são os limites para a normalização (neste caso, N_i = 0.2 e N_s = 0.8).

B. Treinamento e validação das redes neurais

O algoritmo adotado para treinamento da rede neural foi o algoritmo retro-propagação do erro. Para iniciar o processo de treinamento, valores aleatórios entre -1 e 1 foram atribuídos aos pesos das conexões.

Uma primeira rede neural com 4 saídas, correspondente aos quatro intervalos de faturamento, descritos na seção anterior, foi treinada. Como mencionado anteriormente, devido à natureza do problema, onde a falsidade da informação (faturamento anual) é muito latente, pela experiência do especialista no domínio do problema, sabe-se que as pessoas tendem a informar um faturamento menor ao real. Isso pode distorcer os dados utilizados para treinamento das redes neurais. De forma a verificar este fato foi treinada uma rede neural SEM eliminação de inconsistências da etapa do pré-processamento da seção III. Esta rede foi treinada com uma taxa de aprendizado média de 0.4 e após aproximadamente 2.000.000 de iterações foi atingido o erro global de 0,03. O processo de validação alcançou somente o valor de 32% de acertos. Este resultado mostra a natureza dos dados e a difícil recuperação da informação.

Uma segunda rede ainda com 4 saídas, COM eliminação de inconsistências da etapa do pré-processamento foi treinada. A taxa de aprendizado média foi de 0.4 e após 1.800.000 iterações foi atingido o erro global de 0,03, gerando um acerto para os 24 conjuntos de treinamento de 100%. Para validar a rede foram escolhidos 234 conjuntos obtendo um acerto de 59% dos casos.

Pelo conhecimento tácito do problema, expresso através da Equação 4, pesquisas revelam que informações de faturamento dentro dos intervalos 2, 3 e 4 são mutuamente exclusivos em relação ao intervalo 1.

Pelo motivo acima, uma rede neural com 2 saídas, correspondente aos dois intervalos de faturamento expressos através da Equação 5 foi treinada. A taxa de aprendizado foi de 0,4 e após 10.000 iterações foi atingido o erro global de 0,03 gerando um acerto para os 20 (10 de cada intervalo) conjuntos de treinamento de 100% (ver Tabela III). Para validar a rede foram escolhidos 237 conjuntos obtendo um acerto de 78,9% de 218 registros que indicaram Fat_1 e 42,1% de 19 registros que indicaram Fat_2 (Ver Tabela IV).

TABELA III
AMOSTRA DE RESULTADOS DA OPERAÇÃO DA REDE NEURAL PARA DADOS DO CONJUNTO DE TREINAMENTO

$Fat1_{real}$	$Fat2_{real}$	$Fat1_{rede}$	$Fat2_{rede}$
1	0	0,98122406	0,035555333
1	0	1,0229939	-0,002362222
1	0	1,0403508	-0,020859748
1	0	0,9790199	0,03197062
1	0	1,0032363	0,018320978
0	1	0,011625737	1,0110211
0	1	0,005974889	0,9993787
0	1	0,010991782	1,0116637
0	1	0,001197398	1,0012825
0	1	-0,013810873	0,98733854

VI. Conclusões

A presença de valores ausentes em uma base de dados é um fato comum podendo estar distribuído em diversos atributos, num mesmo registro ou de forma aleatória. Valores ausentes podem gerar sérios problemas na extração de conhecimento e na aplicação dos algoritmos de *Data Mining*.

A eliminação de instâncias e/ou atributos com dados ausentes pode acarretar a perda de informações. A substituição por valor padrão, pode introduzir na base informações distorcidas, que não estão contidas no evento e nas circunstâncias que a gerou.

Existem diversas técnicas que lidam com dados ausentes, mas todas elas falham quando existe uma massiva ausência de dados. Neste trabalho, têm sido combinados a capacidade de generalização das redes neurais e o conhecimento tácito de um

TABELA IV

AMOSTRA DE RESULTADOS DA OPERAÇÃO DA REDE NEURAL PARA DADOS DO CONJUNTO DE VALIDAÇÃO

$Fat1_{real}$	$Fat2_{real}$	$Fat1_{rede}$	$Fat2_{rede}$
1	0	0,8915733	0,06579086
1	0	1,0340279	-0,015005767
1	0	-0,21267533	1,2092355
1	0	0,7708286	0,22973764
1	0	1,0256788	-0,005170882
1	0	0,47972614	0,5181096
1	0	0,86554575	0,13493285
1	0	0,35712177	0,6749023
1	0	0,92061925	0,10486618
1	0	0,6344732	0,3975113
0	1	0,17450619	0,83285236
0	1	-0,2352078	1,2431346
0	1	0,2502669	0,7804699
0	1	0,9150722	0,11073053
0	1	0,92756677	0,093660355
0	1	0,6901374	0,27347565
0	1	-0,22690809	1,2352334
0	1	0,9503801	0,06381345
0	1	0,73711133	0,24804366
0	1	-0,041136175	1,0182058

especialista no domínio do problema. Estas, em seu conjunto, visam a não distorção das informações numa base de dados mercadológica.

O resultado geral de 75,9% de acerto indica que as redes neurais junto com o conhecimento tácito são elementos necessários para recuperar informação em bases de dados com massiva ausência de dados. Deve ser notado que o ponto crítico é a identificação de registros outliers, os quais devem ser analisados e verificados se estes correspondem a uma inconsistência ou a um novo padrão. Em futuros trabalhos pretende-se aplicar algoritmos de clusterização que lidam melhor com bancos de dados desbalanceados.

AGRADECIMENTOS

Os autores gostariam de agradecer o financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq - Brasil. Projeto CT - INFO/MCT/CNPq n 031/2004.

REFERÊNCIAS

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth e R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, EUA: MIT Press, 1996.

[2] D. Pyle, *Data Preparation for Data Mining*, EUA: Morgan Kaufmann, 1999.

[3] M. Hofmann e B. Tierney, *The involvement of human resources in large scale data mining projects*, Irlanda: Proceedings of the 1st international symposium on Information and communication technologies, Trinity College Dublin, 2003, pp. 103 - 109.

[4] Y. Fujikawa, *Efficient Algorithms for Dealing with Missing values in Knowledge Discovery*, Japão: School of Knowledge Science - Japan Advanced Institute of Science and Technology, 2001.

[5] D. J. Morgan, *The Thinker's Toolkit - 14 Powerful Techniques for Problem Solving*, EUA: Times Business, 1998.

[6] C. H. Charras, T. Lecroq, *Handbook of Exact String Matching Algorithms*, Reino Unido: Lightning Source, 2004.

[7] C. E. Shannon, *The Mathematical Theory of Communication*, EUA: Bell System Technical Journal, 1948.

[8] N. I. Fisher, *Statistical Analysis of Circular Data*, Austrália: Cambridge Univesity Press, 1995.

[9] S. Haykin, *Redes Neurais: Princípios e Práticas*, Brasil: Bookman, 2001.