UMA PROPOSTA PARA O TRATAMENTO DE VALORES DESCONHECIDOS UTILIZANDO O ALGORITMO K-VIZINHOS MAIS PRÓXIMOS

Gustavo E. A. P. A. Batista e Maria Carolina Monard

Universidade de São Paulo
Av. do Trabalhador Sãocarlense, 400.
Caixa Postal 668, CEP 13560-970, São Carlos - SP, Brasil.
Telefone: +55-16-261-3955, Fax: +55-16-273-9751.
{*gbatista,mcmonard*}*@icmc.sc.usp.br*

Resumo— O crescente interesse em métodos capazes de aprender a partir de dados, tais como Redes Neurais Artificiais e algoritmos de Aprendizado de Máquina, pode ser medido através das inúmeras aplicações dessas tecnologias em diferentes áreas como telecomunicações, produção e manufatura, bem como da criação de uma nova área de pesquisa conhecida como Descoberta de Conhecimento em Bancos de Dados (KDD). Diversos pesquisadores têm reportado que a qualidade dos dados extraídos diretamente de bancos de dados não é boa. De uma forma geral, os problemas encontrados em dados do "mundo real" são mais complexos do que aqueles encontrados em dados provenientes de repositórios. Dessa forma, algumas técnicas de pré-processamento de dados largamente difundidas entre os pesquisadores da comunidade devem ser revisadas. Um exemplo é o tratamento de valores desconhecidos. Muitas vezes esses tratamentos são muito simplistas, como por exemplo substituir os valores desconhecidos de um atributo pela média dos valores conhecidos do atributo (no caso de atributos contínuos) ou pelo valor mais freqüente do atributo (para atributos discretos). Métodos como esses são adequados somente para conjuntos de dados com poucos valores desconhecidos e com valores desconhecidos distribuídos aleatoriamente. Caso essas pré-condições não forem satisfeitas, algum conhecimento inválido pode ser introduzido nos dados. Dados com uma grande quantidade de valores desconhecidos, ou com valores desconhecidos distribuídos de forma não aleatória, devem ser tratados por métodos mais robustos, como por exemplo os métodos baseados em modelos. Métodos baseados em modelos consistem em criar um modelo para predizer os valores desconhecidos. Neste trabalho analisamos o desempenho do algoritmo k-vizinhos mais próximos como um método baseado em modelo para o tratamento de valores desconhecidos. O desempenho do algoritmo k-vizinhos mais próximos é comparado com o desempenho do algoritmo interno utilizado pelo sistema C5.0, um dos melhores sistemas de Aprendizado de Máquina, para tratar valores desconhecidos.

Abstract— The increasing interest in methods that are able to learn from data, such as Artificial Neural Networks and Machine Learning systems, can be measured by numerous applications of these technologies in different areas such as telecommunications, production and manufacturing, as well as the creation of a new research area known as Knowledge Discovery from Databases (KDD). Many researchers have reported that the quality of the data extracted from "real world" databases is not good. The data quality problems found in these databases are usually more complex than those found in data from repositories. In this way, some data pre-processing techniques broadly used by the research community should be revised. An example is the missing data treatment. Frequently, the unknown values of an attribute are substituted by the mean value (or the most frequent value) of the attribute. This sort of technique is only suitable for datasets with few and randomly distributed missing values. If these pre-conditions are not satisfied, invalid knowledge can be introduced into the data. Data with high level of missing values or with non-randomly distributed missing values should be treated by more robust methods, for instance model-based methods. Model-based methods consist in creating a predictive model to predict the values of the missing data. In this work we analyse the performance the k-nearest neighbour algorithm as a model-based method to treat missing values. The performance of the k-nearest neighbour is compared to the performance of the internal algorithm used by the system C5.0, one of the most successful Machine Learning system, to treat missing values.

Key Words— Missing data treatment, data mining, machine learning, k-nearest neighbour

1 Introdução

Valores desconhecidos constituem um problema sério em Aprendizado de Máquina, Redes Neurais, Estatística e outras áreas de pesquisa que utilizam dados como fonte de informação. Eles ocorrem quando parte dos valores de um atributo (ou variável) não foram registrados. Os motivos que levam a essa perda de informação são diversos e variam de aplicação para aplicação, os mais comuns são erros cometidos ao entrar informações manualmente, falhas em equipamentos tais como sensores, recusa de entrevistados em responder às perguntas, entre outros.

Técnicas de Aprendizado de Máquina e Redes Neurais têm sido aplicadas a uma grande variedade de domínios como por exemplo, detecção de fraudes (David and Goyal, 1993), telecomunicação (Manila et al., 1995), manufatura e produção (Mangano and Auriol, 1996). Muitos dos dados provenientes desses domínios possuem grandes quantidades de valores desconhecidos. Por exemplo, em (Lakshminarayan et al., 1999) é reportado que um banco de dados industrial com registros de manutenção e teste de instrumentação estava incompleto em mais de 50% dos seus valores.

Sob tais condições é muito custoso descartar os exemplos (ou registros) com valores desconhecidos, uma vez que grande parte dos dados seria

perdida. Uma outra solução é tratar os valores desconhecidos, substituindo-os por valores estimados. Existe uma grande variedade de métodos capazes de estimar os valores desconhecidos, desde os mais simples, tal como uma substituição pela média (para atributos contínuos) ou moda (para atributos discretos) dos valores conhecidos do atributo, até os mais sofisticados, como a construção de um modelo para predizer os valores desconhecidos.

Métodos como a substituição pela média ou moda são amplamente difundidos por serem de fácil implementação. Entretanto, o emprego de técnicas simples como essas pode levar à introdução de distorções nos dados, sobretudo quando existe uma grande quantidade de valores desconhecidos e/ou os valores desconhecidos não estão dispersos aleatoriamente nos dados.

Desta forma, as características dos valores desconhecidos devem ser analisadas para que um tratamento apropriado seja encontrado. Dados com uma grande quantidade de valores desconhecidos, ou com valores desconhecidos distribuídos de forma não aleatória, devem ser tratados por métodos mais robustos, como por exemplo os métodos baseados em modelos. Métodos baseados em modelos consistem em criar um modelo para predizer os valores desconhecidos. Para construir tal modelo, o atributo com valores desconhecidos é utilizado como classe (atributo de saída) e os demais atributos são utilizados como entrada para um sistema indutivo.

Neste trabalho é proposto um método baseado em modelos utilizando o algoritmo k-vizinhos mais próximos. São realizados experimentos para verificar a viabilidade desse método e os resultados são comparados com o algoritmo interno para tratar valores desconhecidos do sistema C5.0 (Quinlan, 1988), um dos melhores algoritmos de Aprendizado de Máquina. O algoritmo k-vizinhos mais próximos é uma boa escolha para ser utilizado como o algoritmo indutivo de um método baseado em modelo, uma vez que esse algoritmo pode ser utilizado tanto para classificação (para predizer valores de um atributo discreto) quanto regressão (para predizer valores de um atributo contínuo). Ainda, o k-vizinhos mais próximos pode ser facilmente adaptado para lidar com exemplos com múltiplos valores desconhecidos.

Este trabalho está organizado da seguinte forma: na Seção 2 são apresentadas duas características importantes dos valores desconhecidos que devem ser analisadas antes que um método seja escolhido para tratá-los: distribuição e quantidade de valores desconhecidos. Na Seção 3 são descritos os métodos baseados em modelos para tratar valores desconhecidos, e em especial o método que utiliza o algoritmo k-vizinhos mais próximos. Na Seção 4 são apresentados alguns resultados experimentais na tentativa de analisar a eficiência do método baseado em modelos proposto utilizando o k-vizinhos mais próximos para tratamento de valores desconhecidos. Esses resultados são comparados com a performance do algoritmo interno para tratar valores desconhecidos utilizado pelo C5.0, o qual é uma versão aprimorada do sistema C4.5 (Quinlan, 1988). Neste trabalho, estamos interessados sobretudo em analisar o comportamento desses dois métodos quando a quantidade de valores desconhecidos é bastante alta. Finalmente, a Seção 5 apresenta as conclusões deste trabalho.

2 A Importância da Distribuição e da Quantidade dos Valores Desconhecidos

De uma forma geral, a distribuição de valores desconhecidos é mais importante do que a quantidade desses valores (Tabachnick and Fidell, 1996). Valores desconhecidos dispostos aleatoriamente nos dados podem ser considerados um problema menos sério do que quando esses valores não estão aleatoriamente distribuídos. Por outro lado, valores desconhecidos distribuídos não aleatoriamente são um problema sério independente de quão poucos existam, uma vez que esses valores podem afetar a generabilidade dos resultados. Neste caso, torna-se necessário utilizar algum método para estimar e substituir os valores desconhecidos.

Quando poucos valores, comparado com o total de exemplos do conjunto de dados, são desconhecidos, e esses valores estão dispostos de forma aleatória, então, provavelmente, esses valores pouco influenciarão nos resultados das análises. Nesse caso, a maioria dos procedimento para manipular valores desconhecidos deve fornecer resultados semelhantes. Sendo assim, a simples remoção dos casos com valores desconhecidos é uma das formas mais simples e rápidas de solucionar o problema. Entretanto, se existir uma grande quantidade de valores desconhecidos, então esses casos podem fazer falta para o algoritmo indutivo, influenciando na qualidade final do modelo gerado. Infelizmente, ainda não existem diretrizes seguras que indiquem a quantidade de valores desconhecidos que pode ser tolerada por uma amostra de certo tamanho (Tabachnick and Fidell, 1996).

3 Métodos Baseados em Modelos

A criação de modelos é um método sofisticado para estimar valores desconhecidos. Neste caso, o atributo que possui valores desconhecidos é utilizado como classe, e os demais atributos do conjunto de dados são utilizados como entrada para o modelo. O maior argumento a favor dessa abordagem é que, freqüentemente, os atributos possuem relações entre si. Sendo assim, essas correlações podem ser utilizadas para criar um modelo de classificação ou regressão (dependo do tipo do

atributo com valores desconhecidos). E, muitas dessas relações existentes entre os atributos podem ser mantidas, se elas forem capturadas pelo modelo.

Uma desvantagem dessa abordagem é que os novos valores gerados pelo modelo geralmente são mais bem comportados do que deveriam, ou seja, uma vez que o valor desconhecido é predito a partir de outras variáveis, é provável que ele seja mais consistente com essas variáveis do que o valor real seria. Um segundo problema é que existe a necessidade da correlação entre atributos. Ou seja, se não houver alguma relação entre os atributos do conjunto de dados e o atributo com valores desconhecidos, então a estimativa do modelo não será boa. Por fim, se o atributo com valores desconhecidos é contínuo, então muitos modelos de regressão serão capazes apenas de predizer valores dentro da faixa de valores dos casos utilizados na criação do modelo.

Neste trabalho estamos propondo o uso do algoritmo k-vizinhos mais próximos para predizer os valores desconhecidos. Entre os principais benefícios da utilização dessa abordagem pode-se citar:

1. k-vizinhos mais próximos pode ser utilizado tanto para predizer atributos discretos (a moda do atributo entre os k vizinhos mais próximos) quanto contínuos (a média do atributo entre os k vizinhos mais próximos);

2. Não é necessário criar um modelo individual para cada atributo com valores desconhecidos. O k-vizinhos utiliza os próprios exemplos de treinamento como modelo. Como conseqüência, o algoritmo tem a facilidade de trabalhar utilizando qualquer atributo como atributo classe, para isso basta alterar quais atributos serão utilizados pela medida de distância;

3. Essa abordagem pode facilmente tratar exemplos com múltiplos valores desconhecidos.

Entre as principais limitações da utilização do algoritmo k-vizinhos mais próximos, pode-se citar:

1. Toda vez que o algoritmo k-vizinhos mais próximos faz a busca pelos exemplos mais semelhantes, é necessário percorrer todos os exemplos do conjunto de dados. Essa limitação pode ser muito grave, no caso de KDD, uma vez que essa área de pesquisa procura analisar grandes bases de dados. Existem na literatura diversos trabalhos na tentativa de superar essa limitação. Entre os métodos mais conhecidos está a criação de um conjunto de treinamento para o k-vizinhos mais próximos com somente os exemplos mais prototípicos (Wilson and Martinez, 2000), e a utilização de estruturas de dados como as *kd-trees* para a indexação dos exemplos.

4 Análise Experimental

O principal objetivo dos nossos experimentos é avaliar a eficiência do tratamento de valores desconhecidos utilizando o algoritmo k-vizinhos mais próximos e comparar com o desempenho do método interno utilizado pelo C5.0. Deve ser ressaltado que o C5.0 é um dos melhores algoritmos de Aprendizado de Máquina. Nos experimentos realizados, os valores desconhecidos são introduzidos artificialmente em diferentes quantidades no conjunto de exemplos. As performances de ambos os métodos de tratamento de valores desconhecidos são comparadas utilizando as taxas de erro estimadas para cada método. Nós estamos especialmente interessados em analisar o comportamento desses dois métodos quando a quantidade de valores desconhecidos é bastante alta, pois, como já mencionado, diversos pesquisadores da comunidade de KDD têm reportado encontrar bancos de dados com grandes quantidades de valores desconhecidos (Lakshminarayan et al., 1999).

Os experimentos foram realizados utilizando o conjunto de exemplos Breast Cancer (Merz and Murphy, 1998). A Tabela 1 mostra algumas das principais características desse conjunto de exemplos.

N° de atributos	9
N° de atributos contínuos	9
N° de exemplos	699
N° de exemplos com valores desconhecidos	16
N° de classes	2
N° de exemplos - classe “benign”	458 (65.52%)
N° de exemplos - classe “malignant”	241 (34.48%)

Tabela 1. Principais características do conjunto de exemplos Breast Cancer.

Resumidamente, os experimentos consistem em dividir o conjunto de exemplos em 10 pares de conjuntos de treinamento e teste segundo o método 10-fold cross validation. Essa divisão foi feita utilizando os recursos do ambiente AMPSAM (Batista and Monard, 1997). Para cada par de conjuntos de treinamento e teste são realizados os passos descritos na Figura 1. Inicialmente, valores desconhecidos são inseridos aleatoriamente no conjunto de treinamento. São utilizadas duas cópias desse conjunto de treinamento, uma é fornecida diretamente ao algoritmo C5.0 sem qualquer tipo de tratamento de valores desconhecidos. A outra cópia é tratada com a substituição dos valores desconhecidos por valores estimados utilizando o algoritmo k-vizinhos mais próximos. Uma vez que os valores desconhecidos foram tratados, o conjunto de treinamento é fornecido ao C5.0 que gera um classificador. Os dois classificadores gerados pelo C5.0, ou seja, o classificador gerado a partir dos dados com valores desconhecidos e o

classificador gerado a partir dos dados tratados, são utilizados para classificar o conjunto de teste. Ao final das 10 iterações, tem-se uma estimativa da taxa de erro verdadeira através da média das taxas de erro obtidas em cada iteração, juntamente com o desvio padrão. Por fim, é possível analisar e comparar o desempenho de classificação do algoritmo C5.0 aliado ao método de tratamento de valores desconhecidos, com o desempenho do algoritmo C5.0 utilizando o seu próprio método para tratar valores desconhecidos.

Os 16 exemplos com valores desconhecidos presentes originalmente no conjunto de exemplos Breast Cancer foram removidos antes do início dos experimentos. Desta forma, o conjunto passou a ter 683 casos sem valores desconhecidos. O principal motivo para essa remoção é que queremos ter controle absoluto sobre os valores desconhecidos do conjunto de exemplos. Por exemplo, durante o particionamento do conjunto de exemplos em conjuntos de treinamento e teste, é desejável que os conjuntos de teste não tenham algum exemplo com valores desconhecidos. Caso algum conjunto de teste tenha valores desconhecidos, então, a capacidade do algoritmo de aprendizado em classificar corretamente exemplos com valores desconhecidos pode influenciar nos resultados. Isso é indesejável uma vez que o que se deseja medir neste trabalho é a capacidade do algoritmo em aprender com conjuntos de treinamento que possuam valores desconhecidos, e a viabilidade dos métodos de tratamento de valores desconhecidos.

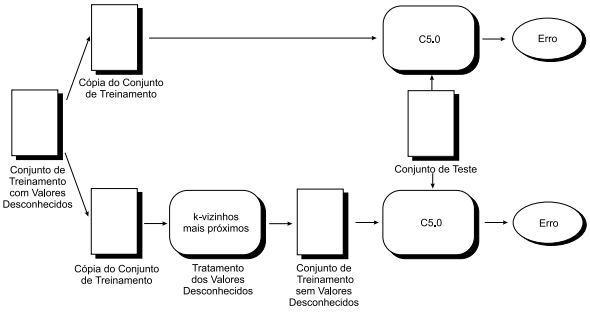

Figura 1. Metodologia utilizada nos experimentos.

O próximo passo é inserir valores desconhecidos somente nos exemplos dos conjuntos de treinamento. Para isso é necessário decidir quais atributos terão seus valores alterados para desconhecido, e qual a proporção de valores desconhecidos a ser inserida em cada atributo. Para decidir quais atributos devem ter os seus valores alterados para desconhecido, é razoável escolher os atributos mais representativos do conjunto de exemplos. Caso contrário, as análises de eficácia dos métodos utilizados podem ser comprometidas com o tratamento de atributos que não são representativos, e desta forma não são utilizados no modelo induzido pelo sistema de aprendizado. Uma vez que encontrar os atributos mais representativos do conjunto de exemplos não é uma tarefa trivial, decidimos encontrar alguns dos atributos mais importantes do conjunto de exemplos para o C5.0. Para isso, o C5.0 foi executado com todos os 683 exemplos e os três atributos mais próximos da raiz da árvore foram selecionados. Esses três atributos são: "Uniformity of Cell Size", "Uniformity of Cell Shape" e "Bare Nuclei". Vale notar que não se pode assegurar que esses atributos serão posteriormente utilizados nos modelos induzidos pelo sistema de aprendizado. A existência de algum outro atributo com a mesma informação (alto índice de correlação) com um dos atributos selecionados, pode fazer com que o C5.0 opte por não utilizar algum desses três atributos selecionados.

Quanto à quantidade de valores desconhecidos a ser inserida nos conjuntos de treinamento, é necessário analisar como os métodos se comportam com diferentes quantidades de valores desconhecidos. Dessa forma, foram inseridos aleatoriamente valores desconhecidos nas proporções de 10%, 20%, 30%, 40% e 50% para cada atributo. Ou seja, quando é dito que foi inserido 20% de valores desconhecidos nos atributos "Uniformity of Cell Size" e "Uniformity of Cell Shape", por exemplo, isso significa que 20% dos valores do atributo "Uniformity of Cell Size" foram selecionados aleatoriamente e tiveram seus valores reais substituídos por desconhecidos. Logo após, 20% dos valores do atributo "Uniformity of Cell Shape" também tiveram seus valores selecionados aleatoriamente (independentemente das seleções realizadas para o outro atributo) e transformados em desconhecidos.

Por fim, os experimentos foram realizados com valores desconhecidos inseridos nos três atributos selecionados "Uniformity of Cell Size", "Uniformity of Cell Shape" e "Bare Nuclei" (Figura 2-a); com valores desconhecidos inseridos nos atributos "Uniformity of Cell Size" e "Uniformity of Cell Shape" (Figura 2-b); e com valores desconhecidos inseridos somente no atributo "Uniformity of Cell Size" (Figura 2-c). No início dos experimentos utilizamos alguns valores diferentes para o parâmetro k (número de vizinhos mais próximos) do algoritmo k-vizinhos mais próximos. Como os resultados foram semelhantes para os diversos valores escolhidos, foi decidido prosseguir com as análises utilizando somente o parâmetro $k = 3$.

Considerando os resultados da Figura 2-a, pode ser observado que o desempenho dos dois métodos foi bastante semelhante, mesmo para uma grande quantidade de valores desconhecidos. De uma forma geral, o método de tratamento baseado em modelos foi ligeiramente superior ao método padrão do C5.0, com exceção da taxa de erro obtida com um conjunto de treinamento com 30% de valores desconhecidos. O resultado obtido com 30% de valores desconhecidos parece não ser

condizente com os demais pontos do gráfico.

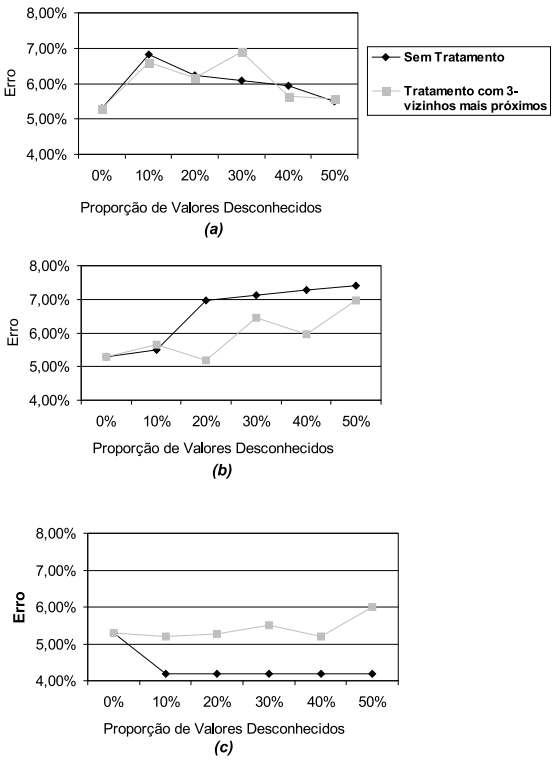

Figura 2. Resultados apresentados graficamente. Valores desconhecidos inseridos nos atributos "Uniformity of Cell Size", "Uniformity of Cell Shape" e "Bare Nuclei" (a); "Uniformity of Cell Size" e "Uniformity of Cell Shape" (b); e "Uniformity of Cell Size" (c).

No caso da Figura 2-b, o método baseado em modelos obteve resultados ligeiramente superiores aos resultados obtidos sem o tratamento de valores desconhecidos.

Por fim, a Figura 2-c mostra os resultados obtidos com valores desconhecidos inseridos somente no atributo "Uniformity of Cell Size". Curiosamente, o C5.0 foi capaz de obter melhores resultados sem o tratamento prévio dos valores desconhecidos do que com o tratamento prévio. Deve-se notar também que para todos os modelos criados com os dados com valores desconhecidos (10% a 50%) a taxa de erro acusada pelo C5.0 foi a mesma (4.18%). Com uma análise mais detalhada das árvores de decisão criadas pelo algoritmo, foi possível constatar que o C5.0 foi capaz de não utilizar o atributo "Uniformity of Cell Size" nos modelos induzidos, sempre que esse atributo possuía valores desconhecidos. Por outro lado, sempre que os valores desconhecidos desse atributo eram tratados através do método proposto, o atributo voltava a ser utilizado nos modelos. Ainda, é curioso verificar que o C5.0 conseguiu criar um modelo aparentemente mais preciso ignorando o atributo "Uniformity of Cell Size", o qual foi escolhido como raiz da árvore de decisão criada com o C5.0 executado com todos os dados do conjunto de exemplo. Como já foi dito anteriormente, o método que utilizamos para escolher os atributos que teriam seus valores alterados para desconhecido, e que supostamente deveriam ser os atributos mais importantes, não é a prova de falhas.

A Tabela 2-(a, b, e c) mostra os resultados numéricos dos gráficos apresentados na Figura 2-(a, b, e c), respectivamente, juntamente com o desvio padrão.

Analisando os resultados da Tabela 2-a é possível verificar que as taxas de erro obtidas por ambos os métodos de tratamento de valores desconhecidos não esta muito distante da taxa de erro obtida quando o modelo foi gerado sem valores desconhecidos (0%). Isso ocorre mesmo quando existe uma grande quantidade de valores desconhecidos (por exemplo, entre 40% e 50%). Para confirmar estatisticamente essa observação, é possível utilizar um teste de hipótese com o objetivo de verificar se existe uma diferença significativa (com 95% de confiabilidade) entre a taxa de erro obtida quando o modelo foi gerado com 0% de valores desconhecidos e quando o modelo foi gerado com valores desconhecidos. Neste trabalho foi utilizado o teste de hipóteses descrito em (Baranauskas and Monard, 2000). Caso esse teste de hipótese forneça um valor em módulo maior ou igual a 2, então existe uma diferença estatisticamente significativa entre as duas taxas erro com 95% de confiabilidade.

	%?	Sem Trat.	TH1	Com Trat.	TH2
(a)	0%	5.30 ± 0.60	-	5.30 ± 0.60	-
	10%	6.82 ± 0.51	-2.73	6.60 ± 0.79	-1.85
	20%	6.24 ± 0.63	-1.53	6.16 ± 0.94	-1.09
	30%	6.09 ± 0.58	-1.34	6.90 ± 0.73	-2.39
	40%	5.94 ± 0.65	-1.02	5.65 ± 0.84	-0.48
	50%	5.50 ± 0.69	-0.31	5.57 ± 0.59	-0.45
(b)	0%	5.30 ± 0.60	-	5.30 ± 0.60	-
	10%	5.50 ± 0.81	-0.28	5.65 ± 0.56	-0.60
	20%	6.98 ± 0.95	-2.11	5.20 ± 0.77	0.14
	30%	7.12 ± 0.98	-2.24	6.45 ± 0.81	-1.61
	40%	7.27 ± 0.92	-2.54	5.79 ± 0.59	-0.82
	50%	7.41 ± 0.92	-2.72	6.97 ± 0.69	-2.58
(c)	0%	5.30 ± 0.60	-	5.30 ± 0.60	-
	10%	4.18 ± 0.56	1.93	5.20 ± 1.01	0.12
	20%	4.18 ± 0.56	1.93	5.28 ± 0.68	0.03
	30%	4.18 ± 0.56	1.93	5.50 ± 0.78	-0.29
	40%	4.18 ± 0.56	1.93	5.20 ± 0.45	0.19
	50%	4.18 ± 0.56	1.93	6.01 ± 0.70	-1.09

Tabela 2. Resultados do tratamento de valores desconhecidos. Valores desconhecidos inseridos nos atributos "Uniformity of Cell Size", "Uniformity of Cell Shape" e "Bare Nuclei" (a); "Uniformity of Cell Size" e "Uniformity of Cell Shape" (b); e "Uniformity of Cell Size" (c).

Nos nossos experimentos, a taxa de erro obtida executando o algoritmo C5.0 sobre o conjunto de exemplos sem valores desconhecidos é comparada com as taxas de erro obtidas pelos dois métodos de tratamento de valores desconhecidos executados sobre um conjunto de treinamento com 10% a 50% de valores desconhecidos. Os resultados dos testes de hipótese são mostrados nas colunas *TH1* e *TH2*. Na coluna *TH1* são mostrados os resultados dos testes de hipótese comparando a taxa

de erro obtida com 0% de valores desconhecidos e as taxas de erro obtidas nos conjuntos de exemples com 10% a 50% de valores desconhecidos sem o tratamento prévio dos dados. Na coluna *TH2* são mostrados os resultados dos testes de hipótese comparando a taxa de erro com 0% de valores desconhecidos e as taxas de erro obtidas nos conjuntos de exemplos com 10% a 50% de valores desconhecidos previamente tratados com o k-vizinhos mais próximos.

Com os resultados da Tabela 2-a é possível constatar que somente existe uma diferença significativa (95% de confiabilidade) nos resultados em duas ocasiões. Na primeira, a taxa de erro obtida sem tratamento com 10% de valores desconhecidos é significativamente superior a taxa de erro do modelo gerado sem valores desconhecidos. Na segunda, o erro do modelo criado a partir dos dados tratados com 30% de valores desconhecidos é significativamente superior ao erro do modelo gerado sem valores desconhecidos. Com os resultados da Tabela 2-b é possível constatar que as taxas de erro obtidas sem tratamento prévio dos dados e com 20%, 30%, 40% e 50% de valores desconhecidos são significativamente superiores a taxa de erro obtida com 0% de valores desconhecidos. Por outro lado, somente a taxa de erro obtida após o tratamento de 50% de valores desconhecidos é significativamente superior a taxa de erro obtida com 0% de valores desconhecidos.

Finalmente, os resultados apresentados na Tabela 2-c não apresentam diferenças significativa entre os modelos gerados.

5 Conclusão

Neste trabalho foi analisado o comportamento de dois métodos para tratar valores desconhecidos: o método baseado em modelos utilizando o algoritmo k-vizinhos mais próximos, e o algoritmo interno para tratamento de valores desconhecidos do C5.0. Ambos métodos foram testados com diferentes quantidades de valores desconhecidos em diferentes atributos. Os resultados obtidos são bastante promissores. Os dois métodos obtiveram um desempenho muito bom, mesmo quando o conjunto de exemplos possui uma grande quantidade de valores desconhecidos, sendo que o método proposto, o qual utiliza o k-vizinhos mais próximos, obteve resultados superiores. Como trabalhos futuros, os métodos de tratamento de valores desconhecidos serão analisados em outros conjuntos de exemplos, e outros métodos tais como *Hot Deck* e *Cold Deck* (Little and Rubin, 1987) também serão analisados. Deve ser observado que neste trabalho está-se inserido os valores desconhecidos de forma aleatória. Em trabalhos futuros, será analisado o comportamento dos métodos quando os valores desconhecidos não estão aleatoriamente distribuídos. Nesse caso, existe a possibilidade de que padrões inválidos sejam criados no modelo. Para uma análise efetiva, será necessário analisar não somente a taxa de erro, mas também o conhecimento presente no modelo induzido por algoritmos indutivos simbólicos.

Agradecimentos. Os autores agradecem a CAPES e FINEP pelo auxílio parcial nesta pesquisa.

Referências

Baranauskas, J. A. and Monard, M. C. (2000). Reviewing some machine learning concepts and methods, *Technical Report 102*, ICMC-USP, São Carlos, SP.

Batista, G. E. A. P. A. and Monard, M. C. (1997). AMPSAM: Um ambiente computacional para medir a performance de sistemas de Aprendizado de máquina, *Anais do I Encontro Nacional de Inteligência Artificial - ENIA 97*, pp. 41–45.

David, A. and Goyal, S. (1993). Management of Cellular Fraud: Knowledge-Based Detection, Classification and Prevention, *Proceedings of 13th Int. Conference on AI, Expert Systems and Natural Language*, Vol. 2, pp. 155–164.

Lakshminarayan, K., Harp, S. A. and Samad, T. (1999). Imputation of Missing Data in Industrial Databases, *Applied Intelligence* **11**: 259–275.

Little, R. J. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, John Wiley and Sons, New York.

Mangano, M. and Auriol, M. (1996). Mining for OR, *ORMS Today, Special Issue in Data Mining* pp. 28–32.

Manila, H., Toivonen, H. and Verkamo, A. (1995). Discovering Frequent Episodes in Sequences, *Proceedings of KDD-95*, pp. 210–215.

Merz, C. J. and Murphy, P. M. (1998). UCI Repository of Machine Learning Datasets. **URL:** *http://www.ics.uci.edu/~mlearn/MLRepository.html*

Quinlan, J. R. (1988). *C4.5 Programs for Machine Learning*, Morgan Kaufmann, CA.

Tabachnick, B. G. and Fidell, L. S. (1996). *Using Multivatiate Statistics*, Haper Collins College Publishers.

Wilson, D. R. and Martinez, T. R. (2000). Reduction Techniques for Exemplar-Based Learning Algorithms, *Machine Learning* **38**(3): 257–286.