Um Estudo Sobre a Efetividade do Método de Imputação Baseado no Algoritmo k-Vizinhos Mais Próximos

Gustavo Enrique de Almeida Prado Alves Batista
Maria Carolina Monard

Laboratório de Inteligência Computacional
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - USP
Av. do Trabalhador São-Carlense, 400 - Centro - Cx. Postal 668
São Carlos - São Paulo - Brasil - CEP 13560-970
{gbatista, mcmonard}@icmc.usp.br

Resumo. Qualidade de dados é uma preocupação central em Aprendizado de Máquina e outras áreas de pesquisa relacionadas à Descoberta de Conhecimento de Bancos de Dados. Um problema relevante em qualidade de dados é a presença de *valores desconhecidos*, também chamados de *valores ausentes*. Diversos métodos de tratamento de valores desconhecidos têm sido propostos na literatura, entre os mais promissores estão os métodos de *imputação*. Em um trabalho anterior dos autores foi mostrado que o algoritmo k-vizinhos mais próximos provê bons resultados comparado com outros métodos de tratamento de valores desconhecidos. Entretanto, existem algumas situações nas quais o método de imputação baseado no algoritmo k-vizinhos mais próximos não é capaz de superar outros métodos mais simples. O objetivo deste artigo é realizar um estudo sobre as características dos dados que podem levar a um desempenho ruim do método de imputação baseado no algoritmo k-vizinhos mais próximos.

1 Introdução

Qualidade de dados é uma preocupação central em Aprendizado de Máquina e outras áreas de pesquisa relacionadas à Descoberta de Conhecimento de Bancos de Dados. Uma vez que a maioria dos algoritmos de aprendizado induzem conhecimento estritamente a partir de dados, a qualidade do conhecimento extraído é amplamente determinada pela qualidade dos dados de entrada.

Um problema relevante em qualidade de dados é a presença de *valores desconhecidos*, também chamados de *valores ausentes*. Valores desconhecidos ou ausentes consistem na não medição dos valores de um atributo para alguns casos. Os valores desconhecidos podem ter diversas fontes como a morte de pacientes, defeitos em equipamentos, recusa por parte de entrevistados em responder determinadas perguntas, entre outras. Apesar da freqüente ocorrência de valores desconhecidos em conjuntos de dados, muitos analistas de dados tratam os valores desconhecidos de forma bastante simplista. Entretanto, o tratamento de valores desconhecidos deve ser cuidadosamente pensado, caso contrário, distorções podem ser introduzidas no conhecimento induzido.

Na maioria dos casos, os atributos de um conjunto de dados não são independentes entre si. Dessa forma, aproximações para os valores desconhecidos podem ser determinadas por meio da identificação de relações entre os atributos. *Imputação*[1] é um termo utilizado para denotar um procedimento que substitui os valores desconhecidos de um conjunto de dados por valores estimados. Essa abordagem permite que o tratamento de valores desconhecidos seja independente do algoritmo de aprendizado utilizado, o que permite ao analista de dados selecionar o método de tratamento de valores desconhecidos mais apropriado para cada conjunto de dados.

Neste trabalho, o desempenho do algoritmo k-vizinhos mais próximos [1] é analisado como um método de imputação. Em [2] é mostrado que o algoritmo k-vizinhos mais próximos pode fornecer bons resultados, em geral, superiores aos obtidos pelas abordagens internas utilizadas pelo algoritmos de aprendizado CN2 [4] e C4.5 [8] e pela imputação pela média ou moda. Entretanto, também é mostrado que em situações especiais, a imputação pelo algoritmo k-vizinhos mais próximos pode fornecer resultados inadequados. O principal objetivo deste trabalho é analisar mais detalhadamente quais características de um conjunto de dados podem fazer com que o método de imputação baseado no algoritmo k-vizinhos mais próximos não forneça bons resultados.

Este trabalho está organizado da seguinte forma: na Seção 2 é apresentada a metodologia utilizada nas análises experimentais apresentadas neste trabalho; na Seção 3 são discutidos os resultados obtidos em três conjuntos de dados; por fim, na Seção 4 são apresentadas as conclusões deste trabalho.

[1] *Imputation*.

Conjunto de dados	#Exemplos	#Duplicados ou conflitantes (%)	#Atributos (quanti., quali.)	Classes	% Classes	Erro Majoritário
Breast	699	8 (1,15%)	9 (9,0)	`benign`	65,52%	34,48%
				`malignant`	34,48%	para a classe `benign`
Sonar	208	0 (0,00%)	60 (60,0)	`M`	53,37%	46,63%
				`R`	46,63%	para a classe `M`
TA	151	45 (39,13%)	5 (1,4)	`1`	32,45%	65,56%
				`2`	33,11%	para a classe `3`
				`3`	34,44%	

Tabela 1: Descrição resumida dos conjuntos de dados.

2 Metodologia

Os experimentos foram realizados com os conjuntos de dados Breast, Sonar e TA do repositório UCI [3]. O conjunto de dados **TA** não possui valores desconhecidos. Os conjuntos de dados **Breast** e **Sonar** possuem poucos valores desconhecidos, no total 16 casos ou 2,28%, e 37 casos ou 5,36%, respectivamente, os quais foram removidos antes do início dos experimentos. A principal razão para não utilizar dados com valores desconhecidos é a preocupação em ter todo o controle sobre os valores desconhecidos nos conjuntos de dados. Por exemplo, é desejável que os conjuntos de teste não possuam valores desconhecidos. Caso algum conjunto de teste possua valores desconhecidos, então a habilidade do indutor em classificar exemplos com valores desconhecidos corretamente pode influenciar nos resultados. Essa influência não é desejável uma vez que o objetivo é analisar a viabilidade dos métodos de tratamento de valores desconhecidos.

Na Tabela 1 são apresentadas algumas das principais características dos conjuntos de dados utilizados neste estudo. Nela são apresentados, para cada conjunto de dados, o número de exemplos (#Exemplos), o número e percentual de exemplos duplicados (que aparecem mais de uma vez) e conflitantes (com os mesmos valores de atributos, mas com classe diferente), o número de atributos (#Atributos), o número de atributos quantitativos e qualitativos, a distribuição da classe e o erro majoritário. Essas informações foram obtidas utilizando o utilitário *info* da biblioteca $\mathcal{MLC}$++ [5].

Para estimar a taxa de erro dos métodos de tratamento utilizados foi aplicado o método de reamostragem *10-fold cross validation*. Valores desconhecidos foram introduzidos artificialmente nos conjuntos de treinamento. Os conjuntos de treinamento com valores desconhecidos foram fornecidos diretamente aos indutores C4.5 e CN2 e também aos métodos de imputação. A aplicação dos métodos de imputação gera novos conjuntos de treinamento sem valores desconhecidos, os quais foram fornecidos aos indutores.

Ambos indutores C4.5 e CN2 possuem estratégias internas que permitem integrar as informações presentes em exemplos com valores desconhecidos aos modelos gerados por esses sistemas. O método IMPUTAÇÃO PELA MÉDIA OU MODA consiste em substituir todos os valores desconhecidos de um atributo pela sua média ou moda, dependendo se o atributo é quantitativo ou qualitativo, respectivamente.

Ao final das 10 iterações do método *10-fold cross-validation*, a taxa de erro verdadeira de cada método de tratamento de valores desconhecidos pode ser estimada por meio do cálculo da média das taxas de erro em cada iteração. Por fim, o desempenho dos indutores C4.5 e CN2 aliados ao método de imputação baseado no algoritmo K-VIZINHOS MAIS PRÓXIMOS pode ser analisado e comparado com os desempenhos dos métodos utilizados internamente pelos sistemas C4.5 e CN2 para aprender na presença de valores desconhecidos, e com o desempenho dos sistemas C4.5 e CN2 aliados ao método IMPUTAÇÃO PELA MÉDIA OU MODA.

Para inserir os valores desconhecidos nos conjuntos de treinamento, alguns atributos devem ser escolhidos, e parte dos valores desses atributos devem ser selecionados para serem modificados para desconhecido. Neste experimento foi decidido inserir valores desconhecidos nos atributos mais representativos de cada conjunto de dados. Essa decisão foi tomada pois deseja-se medir a efetividade dos métodos de tratamento de valores desconhecidos. Tal efetividade não pode ser medida se os atributos tratados forem não representativos, os quais provavelmente não seriam incorporados ao classificador pelo sistema de aprendizado.

Uma vez que encontrar os atributos mais representativos de um conjunto de dados não é uma tarefa trivial, foram utilizados os resultados de [6] para selecionar os três atributos mais relevantes de um conjunto de dados segundo diversos métodos de seleção de atributos tais como *wrappers* e filtros. Foram identificados como atributos mais relevantes aqueles atributos mais freqüentemente selecionados pelos métodos de seleção. Quando possível, tentou-se dar uma ordem de importância aos atributos, da seguinte forma: entre os atributos identificados como mais relevantes, aquele mais freqüentemente selecionado pelos métodos de seleção foi escolhido como o mais relevante, o segundo mais freqüentemente selecionado foi escolhido o segundo mais relevante, e assim por diante. Na Tabela 2 são apresentados os atributos selecionados em cada conjunto de dados, ordenados por relevância.

Com relação à quantidade de valores desconhecidos a serem inseridos nos conjuntos de treinamento, deseja-se analisar o comportamento de cada um dos métodos de tratamento com diferentes quantidades de valores desconhecidos.

Dessa forma, os valores desconhecidos foram inseridos nas seguintes porcentagens: 10%, 20%, 30%, 40%, 50% e 60% do total de exemplos no conjunto de treinamento. Os valores desconhecidos foram inseridos em um único atributo, em dois atributos e, por fim, nos três atributos selecionados como mais representativos.

Embora os valores desconhecidos possam ser inseridos em diferentes distribuições, decidiu-se inserir os valores de forma completamente aleatória, MCAR [7]. Dessa forma, a distribuição dos valores desconhecidos não está sob o controle do experimento, impedindo assim que os valores desconhecidos sejam inseridos de forma que beneficiem um ou outro método.

Os valores desconhecidos foram substituídos por valores estimados utilizado 1, 3, 5, 10, 20, 30, 50 e 100 vizinhos mais próximos, além da substituição pela média ou moda do atributo. Por motivos de espaço, somente os resultados com 10 vizinhos mais próximos, identificados como 10-NNI[2], são apresentados neste trabalho.

Conjunto de Dado	Atributos Selecionados		
	Identificador	Posição	Tipo
Breast Cancer	*Uniformity of Cell Size*	1	inteiro
	Bare Nuclei	5	inteiro
	Clump Thickness	0	inteiro
Sonar	*A10*	10	inteiro
	A0	0	inteiro
	A26	26	inteiro
TA	*English Speaker*	0	nominal
	Course Instructor	1	nominal
	Course	2	nominal

Tabela 2: Atributos selecionados como os mais representativos de cada conjunto de dados.

3 Resultados Experimentais

Nesta seção são apresentados e discutidos os resultados experimentais obtidos para os conjuntos de dados **Breast** — Seção 3.1, **Sonar** — 3.2, e **TA** — Seção 3.3.

3.1 Conjunto de Dados Breast

Embora a imputação com K-VIZINHOS MAIS PRÓXIMOS pode prover bons resultados [2], existem ocasiões em que seu uso deve ser evitado. Uma dessas situações pode ser ilustrada pelo conjunto de dados **Breast**. Esse conjunto de dados possui fortes correlações entre seus atributos. As correlações causam uma situação interessante: por um lado, o algoritmo K-VIZINHOS MAIS PRÓXIMOS pode predizer os valores desconhecidos com uma precisão bem superior à IMPUTAÇÃO PELA MÉDIA OU MODA; por outro lado, o indutor pode decidir não utilizar o atributo tratado, substituindo esse atributo por outro com alta correlação. Os resultados obtidos com o conjunto de dados **Breast** são mostrados na Figura 1, na qual pode ser visto que o método 10-NNI não supera os demais métodos de tratamento de valores desconhecidos.

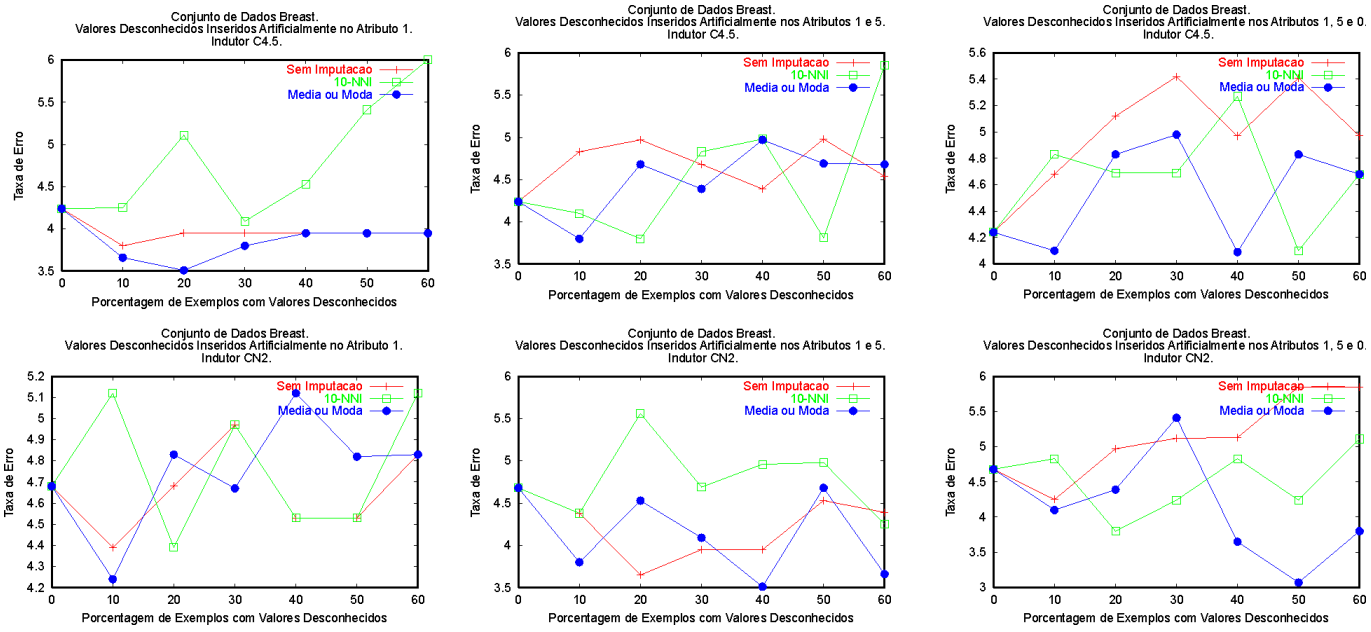

Figura 1: Comparação do método 10-NNI com as estratégias internas utilizada pelos indutores C4.5 e CN2 e com a IMPUTAÇÃO PELA MÉDIA OU MODA para o conjunto de dados **Breast**.

[2] *10-Nearest Neighbour Imputation.*

Como explicado anteriormente, o cenário causado pelo conjunto de dados **Breast** é interessante pois o método 10-NNI foi capaz de predizer os valores desconhecidos com uma precisão superior a IMPUTAÇÃO PELA MÉDIA OU MODA. Como os valores desconhecidos foram inseridos artificialmente nos dados, o erro médio quadrático (*mse*) entre os valores preditos e os valores reais pode ser medido. Esses erros são apresentados na Tabela 3 para os três atributos selecionados como mais relevantes no conjunto de dados.

Se o método 10-NNI é mais preciso em predizer os valores desconhecidos, então porque essa precisão não é traduzida em um classificador mais preciso? A resposta pode estar na alta correlação ente os atributos no conjunto de dados, e porque (ou conseqüentemente) o conjunto de dados **Breast** possui diversos atributos com poder de precisão similar.

Atributo	*mse* 10-NNI	*mse* MÉDIA OU MODA
0 (*Clump Thickness*)	$4,02 \pm 0,14$	$7,70 \pm 0,28$
1 (*Uniformity of Cell Size*)	$1,72 \pm 0,11$	$8,96 \pm 0,36$
5 (*Bare Nuclei*)	$4,23 \pm 0,30$	$13,29 \pm 0,46$

Tabela 3: Erro médio quadrático (*mse*) entre os valores preditos e os valores reais para os métodos 10-NNI e IMPUTAÇÃO PELA MÉDIA OU MODA — conjunto de dados **Breast**.

Para realizar uma análise mais profunda, é necessário verificar como cada atributo é utilizado pelo classificador induzido. Por exemplo, é interessante entender como o sistema C4.5 foi capaz de obter uma taxa de erro constante mesmo com uma grande quantidade de valores desconhecidos inseridos no atributo 1. Analisando as árvores de decisão geradas pelo indutor C4.5, é possível verificar que o C4.5 foi capaz de substituir o atributo 1 — *Uniformity of Cell Size* — pelo atributo 2 — *Uniformity of Cell Shape*. Essa substituição foi possível pois esses dois atributos possuem uma alta correlação (coeficiente de correlação linear $r = 0.9072$). De uma forma geral, para o conjunto de dados **Breast**, o indutor C4.5 foi capaz de trocar todos os atributos com valores desconhecidos por outros atributos, e ainda assim obter resultados similares ou melhores que os obtidos pelo método 10-NNI.

Utilizando o nível mais alto da árvore de decisão em que o atributo foi incorporado como uma medida heurística da importância do atributo no classificador, na Tabela 4 é mostrado que o indutor C4.5 foi capaz de descartar gradualmente os atributos com valores desconhecidos conforme a quantidade de valores desconhecidos aumentava. De forma similar, o indutor C4.5 mostra uma tendência de descartar os atributos com valores desconhecidos quando esses atributos são tratados pelo método IMPUTAÇÃO PELA MÉDIA OU MODA. Esse resultado é esperado uma vez que no método IMPUTAÇÃO PELA MÉDIA OU MODA todos os valores desconhecidos são substituídos por um mesmo valor, ou seja, a média ou moda do atributo. Conseqüentemente, o poder de discriminação do atributo, medido por diversos indutores por meio da *entropia* ou de outras medidas similares, tende a decrescer. O mesmo não ocorre quando os valores desconhecidos são tratados pelo método 10-NNI. Quando o método 10-NNI é utilizado, o indutor C4.5 mantém os atributos com valores desconhecidos como os atributos mais próximos da raiz da árvore de decisão. Essa situação poderia ter sido uma vantagem se o conjunto de dados **Breast** não possuísse outros atributos com poder de predição similar aos atributos selecionados.

% Desconhecidos	Sem Imputação			Média ou Moda			10-NNI		
	Atrib. 1	Atrib. 5	Atrib. 0	Atrib. 1	Atrib. 5	Atrib. 0	Atrib. 1	Atrib. 5	Atrib. 0
0%	1	2	3	1	2	3	1	2	3
10%	2	2	3	2	2	3	1	2	3
20%	-	2	3	-	3	3	1	2	3
30%	-	5	-	-	3	-	1	2	3
40%	5	4	-	3	-	-	1	2	3
50%	-	-	-	6	7	3	1	3	2
60%	-	5	-	-	3	-	1	2	3

Tabela 4: Nível da árvore de decisão no qual os atributos 1, 5 e 0 do conjunto de dados **Breast** foram incorporados pelo indutor C4.5. "-" significa que o atributo não foi incorporado à árvore de decisão. Nível 1 representa a raiz da árvore.

3.2 Conjunto de Dados Sonar

Para confirmar os resultados obtidos com o conjunto de dados **Breast**, foi incluído nos experimentos o conjunto de dados **Sonar**. O conjunto de dados **Sonar** possui características similares ao conjunto de dados **Breast**, uma vez que seus atributos possuem fortes relações entre si. Uma outra característica interessante do conjunto de dados **Sonar** é que esse conjunto de dados possui uma grande quantidade de atributos, 60 no total. Essa grande quantidade de atributos pode fornecer ao indutor diversas possibilidades durante a escolha dos atributos que irão compor o classificador. Na Tabela 5 são mostrados os índices de correlação linear entre os atributos selecionados e os atributos de maior correlação linear presentes no conjunto de dados.

Na Figura 2 são mostrados os resultados para o conjunto de dados **Sonar**. Da mesma forma que nos resultados obtidos para o conjunto de dados **Breast**, o sistema C4.5 foi capaz de substituir os atributos com valores desconhecidos por outros atributos com informações similares. Induzindo o mesmo classificador, o sistema C4.5 foi capaz

de apresentar a mesma taxa de erro, mesmo para grandes quantidades de valores desconhecidos. Diferentemente do conjunto de dados **Breast**, o método 10-NNI foi capaz de superar o sistema C4.5 em duas situações: quando os valores desconhecidos foram inseridos no atributo 10 e nos atributos 10, 0 e 26.

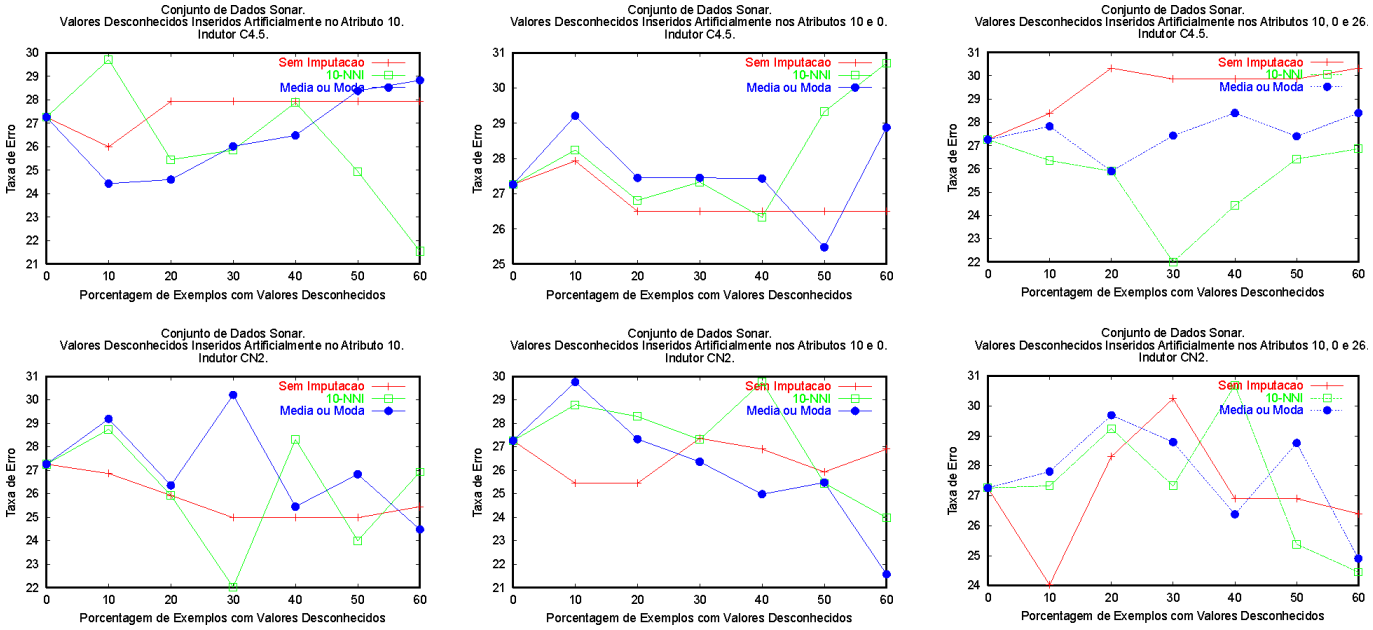

Figura 2: Comparação do método 10-NNI com as estratégias internas utilizada pelos indutores C4.5 e CN2 e com a IMPUTAÇÃO PELA MÉDIA OU MODA para o conjunto de dados **Sonar**.

Para o indutor CN2, não é possível dizer que um dos métodos foi superior aos demais. Os resultados apresentam alta variância, e freqüentemente um método apresenta a taxa de erro mais baixa para um determinado nível de valores desconhecidos, e logo em seguida apresenta a taxa de erro mais alta para o nível de valores desconhecidos imediatamente maior.

Atributo selecionado	Atributo de maior correlação	Índice de correlação linear r
A10	A9	$0,8531$
A0	A1	$0,7359$
A26	A25	$0,8572$

Tabela 5: Índice de correlação linear r entre os atributos selecionados como mais representativos e os atributos de maior correlação — conjunto de dados **Sonar**.

3.3 Conjunto de Dados TA

Uma outra situação na qual o uso de métodos de imputação baseados em modelos de predição não é recomendada é quando os atributos não possuem correlações entre si, ou quando o modelo de predição não é capaz de identificar e incorporar essas correlações. O conjunto de dados **TA** ilustra essa situação.

Um indicativo de que o algoritmo K-VIZINHOS MAIS PRÓXIMOS não é capaz de predizer adequadamente os valores desconhecidos é o fato que os valores imputados pelo algoritmo K-VIZINHOS MAIS PRÓXIMOS possuem uma distância média dos valores reais maior ou semelhante ao da IMPUTAÇÃO PELA MÉDIA OU MODA. Novamente, como os valores desconhecidos foram introduzidos artificialmente, é possível medir o erro médio quadrático (mse) entre os valores preditos e os valores reais. Esses valores são apresentados na Tabela 6.

Na Figura 3 são mostrados os resultados para o conjunto de dados **TA**. É possível observar que o método de imputação 10-NNI não foi capaz de superar os demais métodos de tratamento de valores desconhecidos. Sobretudo para o indutor CN2, o método 10-NNI provê os valores mais altos de taxa de erro, com exceção quando os valores desconhecidos são inseridos em 50% e 60% do conjunto de dados nos atributos 0, 1 e 2. Para o indutor C4.5, o método 10-NNI provê bons resultados somente quando os valores desconhecidos foram introduzidos somente no atributo 0. Nos demais resultados, o método 10-NNI provê

Atributo	mse 10-NNI	mse MÉDIA OU MODA
0 (*English Speaker*)	$7,48 \pm 0,84$	$6,71 \pm 0,81$
1 (*Course Instructor*)	$14,35 \pm 1,73$	$14,28 \pm 1,64$
2 (*Course*)	$12,87 \pm 1,69$	$12,88 \pm 1,56$

Tabela 6: Erro médio quadrático (mse) entre os valores preditos e os valores reais para os métodos 10-NNI e IMPUTAÇÃO PELA MÉDIA OU MODA — conjunto de dados **TA**.

resultados semelhantes ou piores de os fornecidos pelos demais métodos de tratamento de valores desconhecidos.

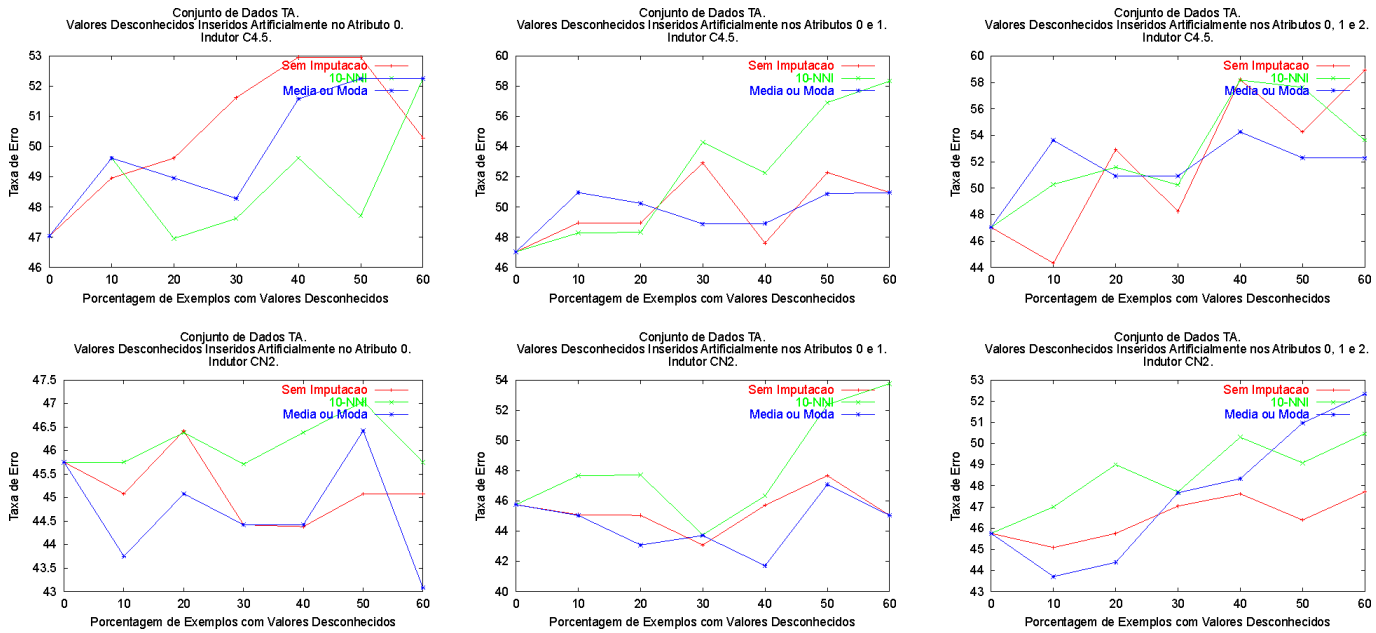

Figura 3: Comparação do método 10-NNI com as estratégias internas utilizada pelos indutores C4.5 e CN2 e com a IMPUTAÇÃO PELA MÉDIA OU MODA para o conjunto de dados **TA**.

4 Conclusão

O estudo realizado neste trabalho utilizando o método de imputação baseado no algoritmo K-VIZINHOS MAIS PRÓXIMOS mostra que mesmo um método avançado de imputação somente é capaz de aproximar os valores reais, não conhecidos, dos valores desconhecidos. Os valores preditos são geralmente mais bem comportados, uma vez que eles são preditos em conformidade com os valores dos outros atributos. Nos experimentos conduzidos, quanto mais atributos com valores desconhecidos são introduzidos, e quanto maior é a quantidade de valores desconhecidos, mas simples são os classificadores induzidos. Dessa forma, a imputação de valores desconhecidos deve ser cuidadosamente aplicada, sob o risco de simplificar demasiadamente o problema em estudo.

Referências

[1] D. W. Aha, D. Kibler, and M. Albert. Instance-based Learning Algorithms. *Machine Learning*, 6:37–66, 1991.

[2] G. E. A. P. A. Batista and M. C. Monard. An Analysis of Four Missing Data Treatment Methods for Supervised Learning. *Applied Artificial Intelligence*, 17(5):519–533, 2003. http://www.icmc.usp.br/~gbatista.

[3] C.L. Blake and C.J. Merz. UCI Repository of machine learning databases, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[4] P. Clark and T. Niblett. The CN2 Induction Algorithm. *Machine Learning*, 3(4):261–283, 1989.

[5] Ron Kohavi, Dan Sommerfield, and James Dougherty. Data Mining Using $\mathcal{MLC}$++: A Machine Learning Library in C++. *International Journal on Artificial Intelligence Tools*, 6(4):537–566, 1997.

[6] H. D. Lee, M. C. Monard, and J. A. Baranauskas. Empirical Comparison of Wrapper and Filter Approaches for Feature Subset Selection. Technical Report 94, ICMC-USP, 1999. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_94.ps.zip.

[7] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley and Sons, New York, 2 edition, 2002.

[8] J. R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, CA, 1988.