Uma Comparação Experimental de Métodos de Imputação de Valores Desconhecidos

Diego Furtado Silva, Gustavo E. A. P. A. Batista

[1] Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Laboratório de Inteligência Computacional – LABIC
Caixa Postal 668, 13560-970 – São Carlos, SP, Brasil

Abstract. *Data quality is a major concern in areas related to Knowledge Discovery from Databases. Any problems found in input data can seriously affect the knowledge induced by methods of such areas. The occurrence of missing data is among the potential problems. Missing data is constituted by values not measured in data collection. This paper presents a discussion and an empirical study on treatment of missing data. Experiments were performed on 17 data sets with missing values inserted artificially and treated by 9 different methods. We have evaluated the influence of the treatment methods on classification accuracy and the difference between the actual and the estimated values. Among the evaluated methods, imputation methods based on prediction models obtained better results, as they provide a better classification accuracy, and insert a lower imputation error.*

Resumo. *Qualidade de dados é uma preocupação central em áreas relacionadas à Descoberta de Conhecimento em Bases de Dados. Quaisquer problemas encontrados nos dados de entrada podem afetar seriamente o conhecimento induzido por métodos dessas áreas. A ocorrência de valores desconhecidos está entre os problemas em potencial. Valores desconhecidos são contituídos de valores não medidos durante a coleta de dados. Este trabalho apresenta uma discussão e um estudo empírico sobre o tratamento de valores desconhecidos. Experimentos foram realizados sobre 17 conjuntos de dados com valores desconhecidos inseridos artificialmente e tratados com 9 métodos distintos. Foram testadas a influência do tratamento desses valores sobre a taxa de acerto de classificação e a diferença entre valores reais e estimados. Entre os métodos avaliados, os métodos de imputação baseados em modelos de predição obtiveram melhores resultados, por resultarem em uma melhor classificação, além de inserirem um menor erro de imputação.*

1. Introdução

Qualidade de dados é uma preocupação central em Aprendizado de Máquina e outras áreas relacionadas à Descoberta de Conhecimento em Bases de Dados. Visto que grande parte dos algoritmos de aprendizado induz conhecimento a partir de dados, a qualidade do conhecimento obtido depende, em grande parte, dos dados de entrada. Assim, quaisquer problemas encontrados nos dados de entrada podem afetar seriamente o resultado desses processos.

Em Mineração de Dados, os dados utilizados são simples reflexos do mundo real, sendo que eles são, muitas vezes, representados de maneira simples, com um grau de

complexidade e detalhe muito inferior à realidade. Considerando a complexidade de detalhes das informações do mundo real, não importa o quão cuidadoso é o processo de obtenção de dados, sempre há a possibilidade de haverem falhas no decorrer do mesmo.

Dentre os problemas possíveis, está a ocorrência de valores desconhecidos, também chamados valores faltantes ou ausentes. Valores desconhecidos constituem na não medição de valores. Eles podem ter diversas fontes, como a recusa de entrevistados a responder certas perguntas, defeitos em equipamentos de medição, perda de formulários e muitas outras. A ocorrência desses valores pode impossibilitar o uso de alguns algoritmos que requerem que os dados estejam completos para sua execução, como Máquinas de Suporte Vetorial [Joachims 2002] e Redes Neurais [Haykin 1999].

Apesar dos valores desconhecidos serem bastante frequentes em conjuntos de dados reais e causarem problemas como os descritos, é muito comum que o tratamento desses valores seja realizado de maneira simplista. As técnicas mais utilizadas para o tratamento são ignorar casos que contenham valores desconhecidos e estimar valores para substituí-los, comumente pela média ou moda do atributo. Porém, ignorar casos acarreta na perda de informação, enquanto estimar valores para substituir os faltantes pode causar distorção nos dados, por exemplo, a substituição pela média ou moda tipicamente reduz artificialmente a variância dos atributos. Por isso, o tratamento de valores desconhecidos deve ser pensado cautelosamente.

O objetivo deste trabalho é estudar os valores desconhecidos e as implicações de seu tratamento. Para isso, é feito um estudo sobre as propriedades dos valores desconhecidos (Seção 2) e das principais técnicas de tratamento (Seção 3 e Seção 4). Foram também realizados experimentos com 9 técnicas de tratamento de valores desconhecidos em 17 conjuntos de dados dos repositórios UCI [Asuncion and Newman 2007] e *Weka*[1] (Seção 5).

2. Aleatoriedade dos Valores Desconhecidos

Em um conjunto de dados, a probabilidade de um exemplo possuir um valor desconhecido em um certo atributo pode não depender de nenhum outro valor no conjunto de dados ou pode ter uma relação direta com o valor real do atributo ou com valores de outros atributos. Aleatoriedade dos valores desconhecidos é o termo utilizado para descrever a distribuição dos valores desconhecidos entre os dados e pode ser dividida em três classes distintas [Little and Rubin 2002]:

Ausentes de forma totalmente aleatória

Valores desconhecidos distribuídos de forma completamente aleatória (MCAR[2]) ocorrem quando os valores desconhecidos de um atributo não estão relacionados nem com valores do próprio atributo nem com valores de qualquer outro atributo do conjunto de dados, ou seja, a probabilidade de se encontrar um valor desconhecido em um certo atributo é a mesma em qualquer exemplo do conjunto de dados. Esse é o caso em que os valores desconhecidos possuem maior grau de aleatoriedade. Um exemplo prático de valores desconhecidos de forma completamente aleatória é quando ocorre a perda aleatória de partes de formulários contendo respostas de entrevistados.

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] *Missing Completely At Random*

Ausentes de forma aleatória

Valores desconhecidos dispostos de forma aleatória (MAR[3]) são aqueles que acontecem em casos em que a probabilidade de ocorrer valores desconhecidos em um atributo não está relacionada com valores do próprio atributo, mas com valores de um ou mais outros atributos do conjunto de dados. Por exemplo, se o fato de os entrevistados recusarem a responder sobre suas rendas depender não de suas rendas, mas da empresa em que trabalham ou do cargo que ocupam, esses valores estarão distribuídos de forma aleatória.

Ausentes de forma não aleatória

Valores desconhecidos distribuídos de forma não aleatória (NMAR[4]) ocorrem quando os valores desconhecidos de um atributo dependem dos valores do próprio atributo e, possivelmente, de valores de um ou mais outros atributos do conjunto de dados. Por exemplo, se o fato de os entrevistados se recusarem a responder sobre suas rendas depender diretamente de suas rendas, e talvez de algum outro atributo, serão gerados valores desconhecidos de forma não aleatória.

3. Tratamento de Valores Desconhecidos

Existem diversos métodos para realizar o tratamento de valores desconhecidos. Alguns deles removem casos, outros utilizam estatísticas dos atributos, observando as correlações dos atributos ou não. Muitos dos casos foram desenvolvidos para pesquisas de opinião e possuem limitações no contexto de Descoberta de Conhecimento em Bases de Dados.

Basicamente, existem três categorias de métodos de tratamento de valores desconhecidos [Little and Rubin 2002]:

Ignorar ou descartar dados

Uma maneira simples de lidar com valores desconhecidos é ignorar ou descartar dados que possuam tais valores. Uma abordagem mais utilizada para esse método é a remoção de qualquer caso que possua algum valor desconhecido. Essa abordagem é conhecida como análise de casos completos e está disponível, muitas vezes como método padrão, na maioria dos programas estatísticos. Outra abordagem existente consiste na remoção de exemplos e/ou atributos com grandes quantidades de valores desconhecidos. Ambas abordagens devem ser utilizadas apenas quando os valores desconhecidos são distribuídos de forma completamente aleatória, visto que, caso contrário, a remoção de elementos que possam possuir relação com os demais pode introduzir grandes distorções nos dados.

Estimativa de parâmetros

Uma outra forma de tratar valores desconhecidos é utilizando procedimentos de máxima verossimilhança (ML[5]), como forma de estimar parâmetros para valores observados nos dados. Procedimentos ML que utilizam variações do algoritmo EM[6] [Dempster et al. 1977] podem estimar parâmetros de um modelo na ocorrência de valores desconhecidos.

Imputação

[3] *Missing At Random*
[4] *Not Missing At Random*
[5] *Maximum likelihood*
[6] *Expectation-Maximization*

Imputação é o nome dado à técnica de substituir valores desconhecidos por valores estimados para os mesmos. Existem diversas maneiras de realizar tal substituição. Formas mais simples de realizar esta operação são utilizar estatísticas obtidas diretamente dos dados, como a média ou moda do atributo. Essa técnica depende exclusivamente dos valores do atributo em questão, ignorando qualquer relação dele com os demais atributos do conjunto de dados. Porém, existem diversas técnicas diferentes de imputação e muitas delas são mais sofisticadas, mantendo a relação entre os atributos. Métodos de imputação são melhor descritos na próxima seção.

4. Métodos de Imputação

Métodos de imputação, como dito, são aqueles que substituem valores desconhecidos por valores estimados para os mesmos. Existem várias maneiras de substituir valores por meio de imputação. Essas maneiras podem ser divididas em cinco principais categorias, bastante difundidas na literatura [Batista 2003]:

Substituição de casos

O método de substituição de casos foi formulado para atender necessidades das pesquisas de opinião, não sendo de aplicação prática em áreas de Descoberta de Conhecimento em Bases de Dados. Essa técnica consiste em, no caso de haver valores desconhecidos, descartar o caso, obter um novo e inserí-lo nos dados. Por exemplo, caso alguém se recuse a responder alguma pergunta durante a entrevista, suas respostas serão descartadas e outra pessoa será entrevistada em seu lugar.

Conhecimento do domínio

Para realizar a imputação de dados, é possível contar com a experiência de um especialista do domínio de aplicação. O método consiste em um especialista se familiarizar com os dados, estimando valores para substituir os desconhecidos. O grande mérito desse método é que a estimativa do especialista pode ser bastante precisa, além de não haver perda de informação, como no caso de descartar casos. Entretanto, as estimativas do especialista ficam restritas aos dados existentes, podendo, assim, haver um direcionamento do conhecimento a ser aprendido.

De uma forma geral, esse método é seguro quando o especialista está bastante familiarizado com o domínio da aplicação, o conjunto de dados é grande e o número de valores desconhecidos é pequeno. Ainda assim, o processo é manual e lento, além de nem sempre se contar com um especialista para realizar o trabalho.

Média ou moda

Método bastante utilizado, a imputação por média ou moda consiste em substituir um valor desconhecido pela média do atributo, no caso em que o valor desconhecido se encontra em um atributo quantitativo, ou pela moda, caso o atributo em que o valor se encontra seja qualitativo.

Em casos em que não se possui muita informação sobre os dados, ou quando o atributo com valor desconhecido não tem muita relação com os demais, a média é uma boa estimativa para o valor desejado. Ainda, há a vantagem desse método ser conservador, pois a média do atributo não é alterada. Porém, a variância desse atributo diminui, ou seja, a dispersão dos valores desse atributo é reduzida. Além disso, as correlações entre atributos podem ser alteradas com o uso dessa técnica.

Hot Deck e Cold Deck

Os métodos *Hot Deck* e *Cold Deck* são, basicamente, divididos em dois estágios. O primeiro estágio consiste em particionar os dados em *clusters*, ou seja, agrupar os dados com a utilização de um algoritmo de aprendizado de máquina não supervisionado. Em seguida, a estimativa para o valor desconhecido é feita associando-se cada exemplo com valor desconhecido a um *cluster* e calculando a média ou moda dos atributos com valor desconhecido utilizando somente os exemplos do *cluster* ao qual o exemplo foi associado. A diferença entre os métodos *Hot Deck* e *Cold Deck* é que eles geram os *clusters* de maneira diferente. O primeiro utiliza todos os dados para gerar o agrupamento, enquanto o segundo utiliza apenas os exemplos que não possuem valores desconhecidos.

Esses métodos podem dar estimativas fracas para os valores, porém usam agrupamento para tentar associar exemplos a fim de melhorar essas estimativas, se comparadas às obtidas em todo o conjunto.

Modelos de predição

Modelos de predição são métodos sofisticados para tratar valores desconhecidos. Eles consistem em estimar valores com o uso de um modelo preditivo, ou seja, o atributo com valores desconhecidos é utilizado como classe e os demais atributos são utilizados como dados de entrada para o modelo de predição. A principal vantagem desses métodos é que eles procuram manter as correlações entre os atributos, pois essas informações são utilizadas para construir o modelo preditivo. Porém, esse método também possui limitações, como a necessidade dos atributos possuírem correlações entre si. Caso contrário, o modelo preditivo pode não ser preciso na estimativa dos valores. Além disso, os valores estimados por esse procedimento são, muitas vezes, mais comportados do que seriam os valores reais, ou seja, os valores preditos são mais consistentes com o conjunto de atributos.

Uma desvantagem a se considerar sobre essa categoria de métodos de tratamento de valores desconhecidos é que muitos modelos de predição trabalham exclusivamente com classes discretas, como no caso dos algoritmos *NaiveBayes* e *J4.8*, re-implementação em Java do sistema de aprendizado *C4.5* [Quinlan 1993]. Visto que conjuntos de dados com atributos contínuos, passíveis de possuir valores desconhecidos, são muito comuns, é necessário que haja a possibilidade de tratá-los. Para que isso seja possível, uma solução é discretizar os atributos contínuos que possuam valores desconhecidos. Outra solução é utilizar outro método de tratamento no caso de atributos contínuos. Neste trabalho, alguns modelos de predição que lidam apenas com classes discretas são utilizados de forma híbrida com ambas soluções citadas. Ainda, o algoritmo *K-Vizinhos Mais Próximos* é um modelo de predição que permite trabalhar tanto com classes discretas quanto com classes contínuas, não havendo a necessidade de discretização ou alternância de método para cada tipo de atributo com valores desconhecidos.

5. Experimentos

Para estudar empiricamente os valores desconhecidos, foram realizados experimentos sobre conjuntos de dados com valores na forma *MCAR*. Para isso, foram escolhidos 17 conjuntos de dados sem valores desconhecidos e selecionados atributos para conter tais valores. Então, foram inseridos valores desconhecidos artificialmente de forma aleatória

nesses conjuntos. Essa prática garante que o estudo seja feito sobre a mesma distribuição, no caso, da forma *MCAR*.

Foi escolhido inserir valores desconhecidos em 10%, 20%, 30%, 40% e 50% dos exemplos de cada um dos conjuntos selecionados, sendo distribuídos em 1, 2 e 3 atributos por vez. Assim, foram gerados 255 diferentes conjuntos de dados com valores desconhecidos.

Uma tarefa necessária para o desenvolvimento dessa avaliação foi a escolha dos atributos em que os valores desconhecidos seriam inseridos. Para essa tarefa foram utilizados dois algoritmos de seleção de atributos existentes no *Weka* [Witten and Frank 2005], *CfsSubsetEval* e *ReliefFAttributeEval*. O primeiro algoritmo avalia o valor preditivo de cada atributo individualmente, gerando um $ranking$ em que aqueles atributos que possírem maior relação com a classe e menor correlação com os demais atributos recebem maior pontuação. O segundo algoritmo, por sua vez, atribui pontuações de relevância com base em amostras de exemplos. Os atributos escolhidos para a avaliação foram aqueles que tiveram maior diferença entre as pontuações.

Os conjuntos de dados e atributos selecionados são descritos na Tabela 1. Os atributos se encontram na ordem em que foram selecionados, ou seja, ordenados pela diferença entre as pontuações citadas, e com uma indicação se são contínuos (cont.) ou discretos (disc.). No caso de conjuntos que tiveram valores desconhecidos inseridos em mais de um atributo, um exemplo pode receber valores desconhecidos em 1, 2 ou 3 desses atributos, de forma totalmente aleatória.

Tabela 1. Conjuntos de Dados e Atributos Selecionados Para Serem Inseridos Valores Desconhecidos.

Identificador	Atributos	Exemplos	Atributos selecionados
Abalone	9	4176	3 (cont.), 2 (cont.), 8 (cont.)
Autorship	71	841	60 (cont.), 10 (cont.), 70 (cont.)
Bupa	7	344	6 (cont.), 4 (cont.), 3 (cont.)
Contraceptive Method	10	1473	2 (disc.), 8 (disc.), 4 (cont.)
PIMA - Diabetes	10	1473	2 (cont.), 4 (cont.), 3 (cont.)
Letter Recognition	17	20000	13 (cont.), 11 (cont.), 2 (cont.)
Magic04	11	19019	10 (cont.), 4 (cont.), 5 (cont.)
Nursery	9	12960	8 (disc.), 2 (disc.), 3 (disc.)
Oil Spill	49	936	45 (cont.), 24 (cont.), 18 (cont.)
OptDigits	64	5619	32 (cont.), 25 (cont.), 29 (cont.)
PenDigits	17	10991	15 (cont.), 11 (cont.), 14 (cont.)
PG_1	237	3043	1 (cont.), 184 (disc.), 166 (disc.)
Satimage	37	6434	17 (cont.), 21 (cont.), 13 (cont.)
Segment	20	2310	11 (cont.), 19 (cont.), 17 (cont.)
Sonar	61	208	37 (cont.), 43 (cont.), 42 (cont.)
Splice	62	3190	30 (disc.), 33 (disc.), 31 (disc.)
Waveform 5000	41	5000	40 (cont.), 39 (cont.), 38 (cont.)

Após a construção dos conjuntos de dados, foram realizados testes em relação à influência do tratamento dos valores desconhecidos na classificação. Para isso, foram executados testes $10 \times 10-fold\,cross-validation$ com o classificador K-Vizinhos Mais Próximos, sendo que, a cada iteração, os valores desconhecidos são tratados no conjunto destinado a treinamento, ou seja, os nove $folds$ destinados a treinamento são tratados antes de se construir o classificador.

Para realizar o tratamento dos valores desconhecidos nos conjuntos de dados, foram utilizados os métodos $Hot\ Deck$ (HD), $Cold\ Deck$ (CD), Média ou Moda (MMI) e os algorimos baseados em modelos de predição K-Vizinhos Mais Próximos (KNNI), $J4.8$ e $NaiveBayes$, além do algoritmo Remoção de Casos (RM), que descarta qualquer exemplo com valores desconhecidos. Nos casos dos métodos de imputação com $J4.8$ e $NaiveBayes$, os valores contínuos foram tratados com discretização ($J48I - DISC$ e $NBI - DISC$) e com o uso do K-Vizinhos Mais Próximos ($J48 - KNN$ e $NBI - KNN$). Os testes com classificação também foram realizados com o conjunto de dados sem tratamento (NONE), para ser utilizado como parâmetro de comparação.

A Figura 1 mostra os resultados da taxa média de acerto por método de tratamento. Os resultados obtidos revelam que os métodos de imputação sem discretização apresentaram melhor desempenho, sendo que o algoritmo de imputação com K-Vizinhos Mais Próximos teve, mesmo que pouco, melhores resultados. A classificação com os conjuntos de dados tratados pelo algoritmo de Remoção de Casos também se mostrou eficiente, superando os testes com conjuntos tratados por métodos que utilizam discretização e com conjuntos de dados sem tratamento, se aproximando, mas não superando, dos resultado obtidos com algoritmos de imputação sem discretização. Os algoritmos de imputação que utilizam discretização obtiveram resultados inferiores aos demais testes, inclusive os realizados com conjuntos sem tratamento.

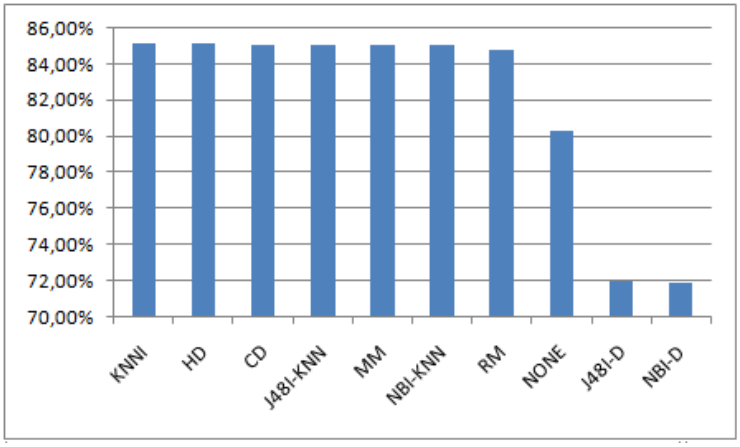

Figura 1. Influência do Tratamento na Classificação

Ainda, é possível notar na Figura 2, que mostra os resultados das classificações de acordo com a porcentagem de exemplos com valores desconhecidos, que as curvas relativas ao tratamento com métodos de imputação sem discretização e remoção de dados possuem uma queda, sendo que a da remoção de casos é levemente mais acentuada.

Outro experimento realizado diz respeito à qualidade dos dados imputados, se aplicando apenas aos métodos de imputação sem discretização. Esse teste se baseia em comparar os dados de um conjunto após o tratamento com o conjunto original, medindo, assim, possíveis distorções causadas nos dados. A diferença entre os conjuntos de dados é calculada valor a valor utilizando a distância Euclidiana entre dois exemplos x_i e x_j, Equação 1, onde $a_r(x)$ é o valor do a-ésimo atributo no exemplo x. A seguir, para melhor visualização, o erro total é dividido pelo número de exemplos do conjunto, a fim de se obter um erro médio distribuído por exemplos. Esse teste não se aplica aos métodos de

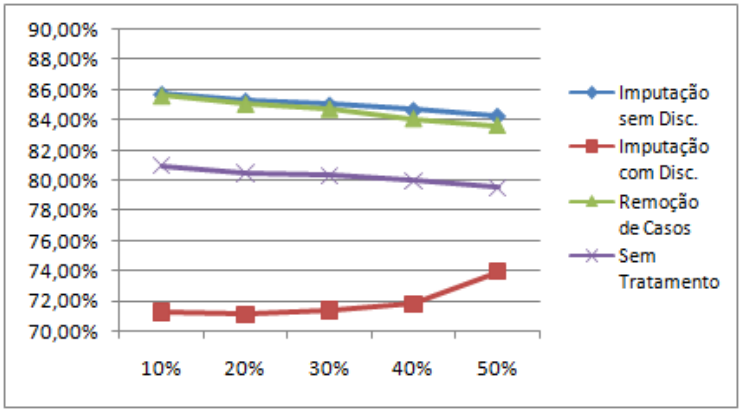

Figura 2. Influência na Classificação por Porcentagem de Exemplos com Valores Desconhecidos

Remoção de Casos nem aos de imputação com discretização porque os conjuntos finais não são compatíveis com os originais, ou seja, possuem menos exemplos ou atributos de tipos diferentes.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \tag{1}$$

A média dos erros obtidos no tratamento por cada técnica é mostrada na primeira parte da Figura 3. Com a realização do experimento, foi possível perceber uma proximidade muito grande dos resultados obtidos entre *Cold Deck* e *Hot Deck* e entre os algoritmos baseados em modelos de predição. Por esse motivo eles foram agrupados, utilizando-se a média entre eles, para a construção da segunda parte da Figura 3, que representa a curva do erro dos conjuntos em relação à porcentagem de exemplos com valores desconhecidos. É possível perceber que os métodos baseados em modelos de predição tiveram menor erro se comparados ao conjunto original, ao contrário do algoritmo de tratamento por média ou moda, que obteve pior resultado dentre os métodos testados.

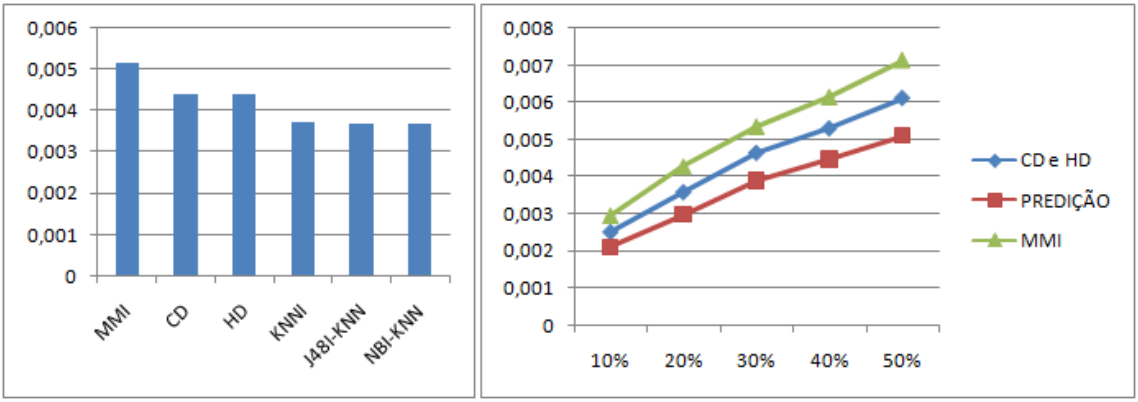

Figura 3. Erro médio nos conjuntos de dados tratados por métodos de imputação e em relação à porcentagem de exemplos com valores desconhecidos

6. Discussões e Conclusões

Os experimentos executados em relação à influência dos tratamentos de valores desconhecidos na classificação mostraram que a discretização, utilizada em alguns dos métodos de tratamento, pode ser bastante prejudicial. Isso ocorre devido à perda de informação que esses métodos provocam. A remoção de casos é um outro método que causa esse problema, entretanto se mostrou eficiente, provavelmente porque os conjuntos de dados possuem valores desconhecidos distribuídos de forma completamente aleatória.

Dentre os algoritmos de imputação, os testes referentes à influência dos tratamentos na classificação não mostrou diferenças muito significativas, embora o tratamento com o método K-Vizinhos Mais Próximos tenha apresentado, levemente, um melhor desempenho na avaliação. Um dos motivos possíveis é o fato de nem sempre os exemplos que passaram por tratamento serem utilizados como vizinhos mais próximos, ou até mesmo os atributos que passaram por tratamento não contribuírem muito no cálculo das distâncias, como em casos de conjuntos de dados com muitos atributos. Outros classificadores poderiam ser utilizados para testes, mas o mesmo tipo de situação poderia ocorrer. Por exemplo, se fosse utilizado um classificador baseado em árvores de decisão, os atributos tratados poderiam não ser utilizados em nós da árvore. Apesar disso, os testes comparativos mostraram que os algoritmos de imputação que inserem valores com menor erro são aqueles que utilizam modelos de predição para realizar o tratamento.

Assim, concluímos que algoritmos de tratamento de valores desconhecidos são uma importante ferramenta em Aprendizado de Máquina e áreas relacionadas, desde que haja um cuidado na escolha do método. Dentre os diversos métodos de imputação, os mais recomendados são os baseados em modelos de predição, como o método baseado no algoritmo K-Vizinhos Mais Próximos, por contribuírem no desempenho da classificação e não causar grandes distorções nos dados. Ainda, o método de remoção de dados pode ser utilizado, desde que se tenha conhecimento que a distribuição dos valores desconhecidos seja totalmente aleatória ou tais valores se encontrem em pequena proporção entre os dados.

7. Trabalhos Futuros

Neste trabalho, foram apresentados estudos sobre a influência do tratamento de valores desconhecidos distribuídos de forma completamente aleatória dentro de conjuntos de dados. Futuramente, pretendemos utilizar uma metodologia semelhante para analisar essa influência em conjuntos de dados com valores desconhecidos nas formas aleatória e não aleatória, divulgando os resultados em futuros congressos científicos.

Referências

Asuncion, A. and Newman, D. (2007). UCI machine learning repository.

Batista, G. E. A. P. A. (2003). Pré-processamento de Dados em Aprendizado de Máquina Supervisionado. Tese de Doutorado, ICMC-USP, `http://www.icmc.usp.br/~gbatista`.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via em algorithm (with discussion).

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines : Methods, Theory and Algorithms*. Springer.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley and Sons, New York, 2nd edition.

Quinlan, R. J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.