Apoiando o processo de imputação com técnicas de aprendizado de máquina

Rodrigo Tavares de Souza[1], Rafael Castaneda Ribeiro[1], Claudia Ferlin[2], Ronaldo Ribeiro Goldschmidt[3], Luis Alfredo V. Carvalho[4], Jorge de Abreu Soares[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)

[2]Pontifícia Universidade Católica do Rio de Janeiro (PUC/RJ)

[3]Instituto Militar de Engenharia (IME)

[4]Universidade Federal do Rio de Janeiro (UFRJ)

rodrigo.souza@eic.cefet-rj.br, rafael.ribeiro@cefet-rj.br, ferlin@inf.puc-rio.br, ronaldo.rgold@ime.eb.br, luisalfredo@medicina.ufrj.br, jorge@eic.cefet-rj.br

Abstract. *The task of imputation of missing data is an important challenge faced by data scientists. In this context, imputation techniques that improve the quality of the data entered are imperative. Exploring both machine learning techniques and variations of the classical imputation process can improve the quality of the imputed data. Hence, this article aims to evaluate the impact of the use of the k-neighbors algorithm faced to the use of the mean in the global imputation process as well as to explore the use of the hot-deck imputation technique with the clustering algorithm k-Means and imputation with k-NN. Results reveal an interesting reduction of absolute error obtained in the simulation in three databases with different characteristics.*

Resumo. *A tarefa de imputação de dados é um importante desafio enfrentado pelos cientistas de dados. Nesse contexto, torna-se imperativo dispor-se de técnicas de imputação que melhorem a qualidade do dado preenchido. Valer-se tanto de técnicas de aprendizado de máquina quanto de variações do processo clássico de imputação pode tornar possível a melhora da qualidade dos dados imputados. Assim, este artigo tem por propósito avaliar o impacto da utilização do algoritmo dos k-vizinhos mais próximos frente ao uso da média no processo de imputação global bem como explorar o uso da técnica de imputação hot-deck com o algoritmo de agrupamento k-Means e a imputação com k-NN. Os resultados revelam interessante redução da margem de erro obtida na simulação em três bases de dados com diferentes características.*

1. Introdução

O importante aumento da quantidade de dados gerenciados por sistemas informatizados é fator inegável na rotina das corporações. Todavia, esse crescimento maximiza um conhecido problema dos administradores de dados: as inconsistências de dados. Essas inconsistências são fruto primário dos diversos momentos onde bases de dados são integradas. Essas bases vêm de diversas fontes, que nem sempre recebem o devido cuidado. Podem também ocorrer por outros motivos, tais como falhas de equipamentos, na transmissão da mensagem, erros de preenchimento não tratados, falhas em rotinas de carga, entre outros [Han et al., 2011]. É um dos casos de inconsistência que demandam atenção é o da ausência de valores em bases de dados.

Essas ausências, dependendo de sua natureza e incidência, podem prejudicar sobremaneira a análise de dados por qualquer técnica produtora de informação, tais como a descoberta de conhecimento em bases de dados, os armazéns de dados (*Data Warehouse*), ou similares, e comprometer seus resultados [Little *et al*, 2002]. Para isso, o estudo de técnicas de complementação de dados ausentes (ou "Imputação") procura soluções para questões em aberto ligadas ao tema, fundamentalmente com base em métodos estatísticos e/ou técnicas de aprendizado de máquina [Farhangfar et al., 2007]. Assim, o objetivo desse trabalho é o de avaliar o impacto na qualidade do dado imputado (erro absoluto entre o que é calculado e o dado real) com técnicas de aprendizado de máquina, tanto na imputação global quanto local e, na ausência de uma técnica consagrada como estado da arte, frente à clássica imputação com o uso de média.

Vários trabalhos exploram a variação de técnicas de imputação, com vistas à melhoria da qualidade do dado imputado, tais como os de Luengo *et al* [2012] e Silva e Zárate [2014]. Porém, uma importante técnica com potencial de melhora da qualidade do dado imputado – ou seja, a diminuição do erro absoluto entre o valor real e o deduzido – é a imputação local ou *hot-deck* [Ford, 1983, Jerez et al., 2010]. Nesta técnica, divide-se o conjunto de dados em grupos, de forma que a composição do valor imputado se dê somente por elementos de tuplas que sejam similares aos elementos da tupla com valor ausente [Fuller et al., 2001]. Neste artigo, a imputação *hot-deck* é implementada por meio da combinação de dois algoritmos de aprendizado de máquina: o agrupamento utiliza o algoritmo *k-Means* [Han *et al*, 2011], e a imputação propriamente dita: o algoritmo dos *k*-vizinhos mais próximos (k-NN) [Han *et al*, 2011]. O desempenho desta abordagem é avaliado em três bases de dados com diferentes níveis de correlação entre seus atributos, variados percentuais de ausência em cada atributo de cada base, e múltiplos valores dos parâmetros dos algoritmos, como proposto por Soares [2007]. Espera-se com isso aproximar cada vez mais o dado imputado do dado real.

Além da introdução, o artigo está organizado da seguinte forma: a Seção 2 analisa a utilização de técnicas de aprendizado de máquina no processo de imputação global (que considera todos os valores presentes na coluna alvo da imputação) e local (que considera um subconjunto desses valores no processo de imputação). A Seção 3 explora os resultados com a utilização do algoritmo k-NN frente à média aritmética simples, e o uso da imputação local com o algoritmo de agrupamento *k-Means* e a imputação com o algoritmo dos *k* vizinhos mais próximos. Por fim, a Seção 4 tece as considerações finais do artigo.

2. Usando técnicas de aprendizado de máquina na imputação global e local

Existem diversos métodos para tratar a ausência de dados, ou seja, substituir valores ausentes por valores reais [Rubin,1988]. A tarefa de imputação tem como objetivo recuperar valores ausentes de maneira mais precisa, através de técnicas que variam desde a média simples, regressão linear, modelos preditivos específicos, até a utilização de algoritmos de aprendizado de máquina [Ford,1983, Jerez *et al*, 2010].

A maior parte dos trabalhos disponíveis na literatura realiza a tarefa de imputação, seja ela simples ou precedida de alguma outra técnica, com o cálculo da média aritmética simples dos valores presentes na coluna cujos valores são imputados (no caso de bases numéricas) ou com o uso da moda (em bases de dados categóricas). Apesar de simples, essa abordagem carrega consigo um considerável erro, por desconhecer qualquer similaridade das tuplas da base.

Nesse contexto, considera-se uma importante e difundida técnica de imputação que busca reduzir o desvio de similaridade entre os dados, classificando a priori a amostra, é denominada imputação local, ou *hot-deck* [Ford, 1983]. Nesta abordagem são utilizados somente grupos de objetos completos que possuam relação de similaridade com o dado ausente [Fuller *et al*, 2001]. O seu principal objetivo visa reduzir desvios

através da classificação da amostra [Ford, 1983], algo difícil de ser atingido [Soares, 2007].

Qualquer estudo envolvendo ausência de dados deve delimitar o mecanismo causador da ausência. Ausências de dados normalmente seguem um mecanismo de distribuição, mas também podem ocorrer de forma intermitente ou simplesmente ao acaso [Little, Rubin, 2002]. Trabalhos envolvendo a complementação de dados ausentes levam inevitavelmente em conta o mecanismo que causou a ausência dos dados, classificadas em três tipos [Little *et al*, 2002]: completamente aleatória – *Missing Completely at Random* (MCAR), aleatória – *Missing at Random* (MAR) e não aleatória – *Not Missing at Random* (NMAR).

3. Avaliação Experimental

Foram utilizadas três bases de dados numéricas com atributos classificadores no experimento, como proposto inicialmente por Soares [2007] e Castaneda *et al* [2008], disponíveis no repositório da Universidade da Califórnia, Irvine: *Iris Data Set*, *Breast Cancer Wisconsin (Original) Data Set*[1] e *Pima Indians Diabetes*[2] [Dua, Karra Taniskidou, 2017]. O conjunto de dados *Iris Data Set* relaciona as medidas de comprimento e largura das pétalas e caules de três espécies de plantas Iris. Já o dataset *Pima Indians* apresenta dados referentes a integrantes de uma tribo indígena, onde parte deles possui diabetes mellitus. Por fim, *Breast Cancer* possui dados do hospital de Wiscosin sobre o diagnóstico de câncer de mama, com dados relativos a pacientes que possuem esta doença. Os dados utilizados foram os originais, sem nenhum processo de normalização dos mesmos.

No que tange à simulação da ausência, adotou-se neste trabalho uma abordagem em largura: gerou-se ausência em todos os atributos de todas as bases, à exceção dos atributos identificadores e classificadores. Para cada base e atributo, provocou-se ausências de dados nas proporções de 10%, 20%, 30%, 40% e 50%, seguindo o mecanismo MCAR, conforme proposto por Soares [2007]. Limitou-se a ausência máxima em 50%, pois valores acima dessa taxa degeneram a base de modo que seu uso passa a ser questionável.

Adotou-se a mesma proposta ampla na variação do parâmetro *k* dos algoritmos de aprendizado de máquina aqui utilizados. O agrupamento precedendo a imputação é técnica conhecida e apresenta bons resultados [Little, Rubin, 2002]. Entretanto, existem poucos estudos que avaliam o número ideal de *k* vizinhos a serem considerados no algoritmo *k-NN* na utilização destes algoritmos no processo de imputação, bem como o número *k* de grupos do algoritmo *k-Means* [Soares, 2007]. Uma abordagem de variação dos valores do parâmetro *k* figura em Castaneda *et al* [2008], que utilizou os valores 1, 3, 5 e 10 para o algoritmo *k-NN* na composição de um *workflow* de imputação iterativa. De forma a explorar os possíveis valores de *k* para o *k-NN* e *k-Means*, foram utilizados todos os valores possíveis para esse parâmetro, nas três bases de dados, para os cinco percentuais de ausência [Soares, 2007]. A Tabela 1 detalha a configuração de *k* para os algoritmos, apresentando os valores mínimo e máximo do parâmetro para cada base, algoritmo e percentual de ausência.

[1] Disponíveis em http://archive.ics.uci.edu/ml/index.php.

[2] Disponível em https://www.kaggle.com/uciml/pima-indians-diabetes-database.

Tabela 1. Valores mínimo e máximo do parâmetro *k* para os algoritmos *k-Means* e *k-NN*

Base de dados	Algoritmo	Percentual de ausência				
		10	20	30	40	50
Iris Plants	k-NN	1-135	1-120	1-105	1-90	1-75
	k-Means	2-135	2-120	2-105	2-90	2-75
Pima Indians	k-NN	1-353	1-314	1-275	1-236	1-196
	k-Means	2-353	2-314	2-275	2-236	2-196
Breast Cancer	k-NN	1-614	1-546	1-478	1-410	1-341
	k-Means	2-614	2-546	2-478	2-410	2-341

As bases *Iris Plants* e *Breast Cancer* apresentam bons índices de correlação entre seus atributos e tendem a favorecer o processo de imputação [Soares, 2007]. Já a base *Pima Indians* apresenta baixo nível de correlação, sendo um dos desafios do experimento. Para aferir a atenuação geral (média) dos métodos em cada base, foram somadas as diferenças entre os erros dos mesmos, em cada percentual de ausência, sendo esse somatório divido por cinco (número de cenários de ausência), conforme especificado na Tabela 2.

Tabela 2. Total de correlações entre os atributos de cada base

Base de dados	Nº de correlações maiores que 50% / Nº de atributos	Correlação
Iris Plants	3 / 4	Alta
Pima Indians	3 / 8	Baixa
Breast Cancer	8 / 9	Alta

As Figuras 1, 2 e 3 mostram um comparativo médio de erro nos atributos de cada uma das bases com imputações feitas utilizando a média aritmética simples frente à complementação de dados ausentes realizada com o algoritmo *k-NN*. Os resultados na base *Iris Plants* são significativamente melhores, como por exemplo com 40% de ausência (erro médio de 75.75% e 9.32% para, respectivamente, os métodos de imputação média e k-NN).

Comportamentos interessantes são observados na base *Breast Cancer*. O uso do algoritmo *k-NN* frente à média revela ganhos consideravelmente animadores. No melhor caso, com 10% de ausência. Nesta situação, obteve-se um erro médio de 114.82% versus 37.46%. Além disso, tem-se que o erro oscilou na casa dos 37% para as diversas taxas de ausência, com desvio-padrão dos experimentos com a aplicação do método *k-NN* frente à média igual a 0.17%, frente a 1.46% na base *Iris Plants* e 2.27% na base *Pima Indians*. A Tabela 3 apresenta os desvios-padrão médios dos erros de imputação por base de dados e método de imputação. Esse é um resultado bem interessante, que demanda aprofundamento em bases com características similares. Isto é, um forte indicativo provavelmente reside no fato de que a alta correlação dos seus atributos ajuda sobremaneira o ganho.

Tabela 3. Desvios-padrão médios dos erros de imputação por base de dados e método de imputação

	k-NN			k-Means + k-NN		
	Iris Plants	Pima Indians	Breast Cancer	Iris Plants	Pima Indians	Breast Cancer
10%	5,85	33,04	37,46	5,59	31,86	36,59
20%	8,84	36,89	37,85	8,22	35,94	36,54
30%	7,59	39,09	37,61	7,29	34,88	37,13
40%	9,32	34,90	37,82	9,01	33,42	38,41
50%	9,20	36,59	37,79	9,12	36,19	38,40
Desvio-padrão	1,46	2,27	0,17	1,46	1,82	0,93

Já a base *Pima Indians*, cuja correlação entre os atributos é baixa, não se destaca como no caso anterior. Todavia, ainda assim os resultados são interessantes. Com 10% de ausência, temos erros médios de 43.15% para a média e 33.04% para o algoritmo dos *k* vizinhos mais próximos.

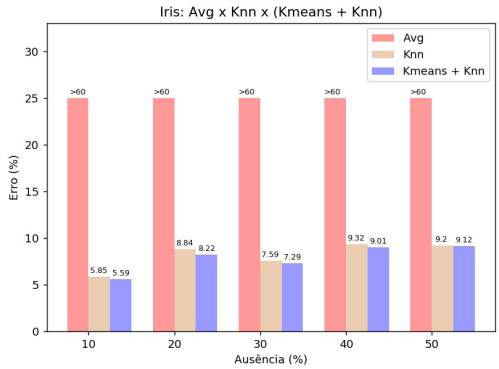

Figura 1. Erro médio na base *Iris Plants* com imputação por média, k-NN e hot-deck

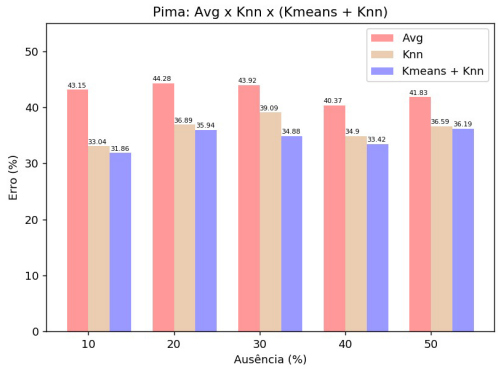

Figura 2. Erro médio na base *Pima Indians* com imputação por média, k-NN e hot-deck

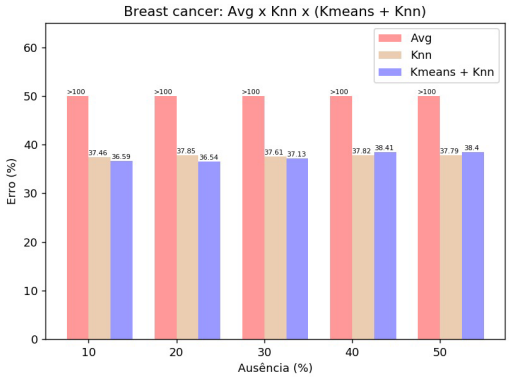

Figura 3. Erro médio na base *Breast Cancer* com imputação por média, k-NN e hot-deck

No que tange ao uso da imputação *hot-deck*, em praticamente todos os experimentos o erro médio de imputação diminui com o agrupamento precedendo a imputação. Apenas na base *Breast Cancer* essa situação se inverte com 40% e 50% de ausência. Todavia, os ganhos não são tão significativos quanto o são a utilização do algoritmo *k-NN* para imputação quando comparada ao uso da média para o mesmo objetivo, como pode ser observado na Tabela 4. Entretanto, os resultados também não desencorajam o investimento no tema, por ter de alguma forma contribuir com o ganho de uma técnica consolidada. As condições do estudo realizado neste artigo podem se revelar mais animadores em outro contexto.

Tabela 4. Diferença entre os erros médios entre a imputação com k-NN e *hot-deck* por percentual de ausência.

	Iris Plants			Pima Indians			Breast Cancer		
	k-NN	k-M + k-NN	Diff	k-NN	k-M + k-NN	Diff	k-NN	k-M + k-NN	Diff
10%	5,85	5,59	0,26	33,04	31,86	1,18	37,46	36,59	0,87
20%	8,84	8,22	0,62	36,89	35,94	0,95	37,85	36,54	1,31
30%	7,59	7,29	0,30	39,09	34,88	4,21	37,61	37,13	0,48
40%	9,32	9,01	0,31	34,90	33,42	1,48	37,82	38,41	-0,59
50%	9,20	9,12	0,08	36,59	36,19	0,40	37,79	38,40	-0,61

4. Conclusão

Neste artigo, experimentou-se o processo de imputação simples e local (*hot-deck*) utilizando duas consagradas técnicas de aprendizado de máquina: o algoritmo dos *k* vizinhos mais próximos para a imputação propriamente dita, e o agrupamento de dados com o algoritmo *k-Means* para a imputação local, além do uso clássico de imputação considerando apenas a média aritmética simples. Os resultados revelaram um promissor ganho de qualidade da imputação com o algoritmo *k-NN* frente ao uso da média aritmética simples. Ao adotar a técnica de imputação local, os resultados mostraram um ganho frente à imputação normal, mas com menos impacto. Esses resultados incentivam o investimento no tema, com a exploração de outras bases com diferentes características e com a utilização de outros algoritmos. Outro importante resultado revelou-se na imputação *hot-deck*, que teve o melhor resultado justamente onde apresenta os níveis de correlação mais desafiadores para o processo de imputação, a base *Pima Indians*. Utilizar a imputação *hot-deck* pode ser uma alternativa interessante para bases de dados que apresentem baixa correlação. Os próximos passos consistirão na aplicação de outras técnicas de aprendizado de máquina que extrapolem as tarefas de agrupamento, além da utilização de processamento de alto desempenho para lidar com bases de dados grandes.

Referências

Castaneda, R., Ferlin, C., Goldschmidt, R., Soares, J., Carvalho, L., Choren, R. (2008). Aprimorando Processos de Imputação Multivariada de Dados com Workflows. XXIII Simpósio Brasileiro de Banco de Dados (SBBD), pages 238–252.

Dua, D., Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Farhangfar, A., Kurgan, L., Pedrycz, W. (2007). A novel framework for imputation of missing values in databases. IEEE Transactions on Systems, Man, and Cybernetics.

Ford, B. L. (1983). An Overview of Hot-Deck Procedures. Incomplete Data in Sample Surveys, 1 ed., vol. 2, Academic Press.

Fuller, W. A., Kim, J. K. (2001). Hot Deck Imputation for the Response Model. Survey Methodology, v. 31, n. 2, pp. 139-149.

Han, J., Kamber, M., Pei, J. (2011). Data Mining: Concepts and Techniques, 3ed. Morgan Kaufmann, Waltham, Mass.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine.

Little, R. J. and Rubin, D. B. (2002). Statistical Analysis with Missing Data. John Wiley & Sons,New York, 2ed.

Luengo, J., García, S., Herrera, F., (2012), On the choice of the best imputation methods for missing values considering three groups of classification methods, Knowledge and Information Systems, v. 32, n. 1 (Jul.), p. 77–108.

Rubin, D. B. (1988). An overview of multiple imputation. In Proceedings of the Survey Research Section, American Statistical Association, pp. 79–84.

Silva, L. O., Zárate, L. E. (2014). A brief review of the main approaches for treatment of missing data. Intelligent Data Analysis, vol. 18, no. 6, pp. 1177-1198.

Soares, J. (2007). Pré-processamento em Mineração de Dados: um Estudo Comparativo em Complementação. Tese de Doutorado, COPPE/UFRJ.