Um Estudo Comparativo do Uso de Abordagens de Comitês de Regressão para Imputação *hot-deck*

Thiago da Silva Pereira[1], **Eduardo Bezerra**[1], **Jorge A. Soares**[1]

[1]CEFET/RJ, Rio de Janeiro/RJ, Brazil

thiago.pereira@aluno.cefet-rj.br

{ebezerra,jorge.soares}@cefet-rj.br

Abstract. *An essential problem in data preprocessing is related to deal with missing data. A possible solution to this problem is hot-deck imputation, a technique comprised of two steps: first cluster similar records in the input dataset and then perform imputation in each separate cluster. However, selecting the best algorithm for the second step is a challenging task. This article presents a comparative study of hot-deck imputation considering two ensemble methods: Bagging and Adaboost. We evaluate these methods using datasets that show different correlations between their attributes, with varying missing value rates. Our results measuring the precision of imputed data by both techniques indicate that Adaboost results in better precision and reasonable processing time.*

Resumo. *Um problema essencial no pré-processamento de dados está relacionado a lidar com dados ausentes. Uma possível solução para esse problema é a imputação hot-deck, uma técnica composta de duas etapas: primeiro agrupar registros semelhantes no conjunto de dados de entrada e, em seguida, realizar a imputação em cada grupo separado. No entanto, selecionar o melhor algoritmo para a segunda etapa é uma tarefa desafiadora. Este artigo apresenta um estudo comparativo da imputação hot-deck considerando dois métodos de comitê: Bagging e Adaboost. Avaliamos esses métodos usando conjuntos de dados com diferentes correlações entre seus atributos, variando as taxas de valor ausente. Nossos resultados medindo a precisão dos dados imputados por ambas as técnicas indicam que o Adaboost resulta em melhor precisão e tempo de processamento razoável.*

1. Introdução

Dados ausentes podem prejudicar a busca de conhecimento em bases de dados, levando a conclusões enviesadas e incorretas. O estudo de estratégias para substituição desses valores ausentes é conhecido por imputação de dados. As técnicas mais simples envolvem substituir valores ausentes pela média ou pela moda. Outras induzem algum modelo de aprendizado de máquina, como a imputação *hot-deck*. Nela, os registros do conjunto de dados que possuem valores ausentes para algum atributo são agrupados por algum critério de similaridade [Ford, 1983; Christopher et al., 2019]. Em seguida, os valores ausentes são preenchidos em cada subgrupo.

A execução da imputação *hot-deck* por meio de métodos de aprendizagem de máquina suscita a questão de qual algoritmo utilizar. Comitês de regressores (*ensemble regressors*) representam uma categoria de algoritmos nos quais a tomada de decisão é

feita pela integração das respostas de vários modelos componentes [Zhang and Ma, 2012]. *Bagging* e *Adaboost* são exemplos de algoritmos que produzem comitês. Em *Bagging*, T regressores são treinados em paralelo, cada qual sobre um conjunto de treinamento gerado por meio da técnica de *bootstrap*, na qual os registros são selecionados de forma aleatória e com reposição. Durante a inferência, os T regressores são combinados utilizando a média simples sobre seus resultados. No *Adaboost*, os T componentes são treinados de forma sequencial. Pesos são atrelados aos registros e, a cada iteração, são recalculados de acordo com os resultados preditivos gerados na iteração anterior. No caso de acerto em um dado registro de treinamento, o peso é diminuído. Contudo, na ocorrência de erro de predição para um registro, o seu peso é aumentado, o que faz com que o próximo componente do comitê dê mais atenção à correta predição do exemplo.

Este trabalho compara experimentalmente das abordagens de imputação simples e *hot-deck*. Para a imputação *hot-deck*, utilizamos *comitês de regressão*. Apresentamos resultados obtidos com as abordagens *Bagging* e *Adaboost*. Realizamos as comparações sobre três conjuntos de dados com diferentes graus de correlação entre seus atributos. Também analisamos o efeito de diferentes porcentagens de valores ausentes (gerados sinteticamente sobre os atributos dos conjuntos de dados utilizados).

Este artigo está organizado como segue. Na Seção 2 apresentamos um visão geral da imputação de dados. Na Seção 3, apresentamos os experimentos comparativos realizados e fornecemos uma análise dos resultados obtidos. Na Seção 4, as conclusões.

2. Imputação de Dados

Um aspecto relevante na tarefa de imputação é o mecanismo de ausência dos dados. Existem três tipos de classificações para mecanismos de ausência: completamente aleatório (*MCAR – Missing Completely At Random*) - quando não se conhecem as causas que geraram a aleatoriedade; aleatório (*MAR – Missing At Random*) - a ausência de valores em uma coluna é causada pelas demais do conjunto de dados; e não aleatório (*NMAR – Not Missing At Random* ou *IM – Ignorable Missing*) - que ocorre quando a coluna com valores faltantes influencia única e diretamente a ausência.

Grande parte dos trabalhos sobre imputação realizam comparações entre algum algoritmo de aprendizado de máquina em relação à imputação por média (para atributos numéricos) ou moda (para atributos categóricos), comparam algoritmos comitês em relação a algoritmos não comitês ou comparam a imputação *hot-deck* com a imputação simples [Souza et al., 2018; Souza, 2019]. Essa situação ilustra a importância de se realizar estudos comparativos entre diferentes algoritmos comitês no contexto da imputação.

A imputação simples não considera as similaridades entre os registros do conjunto de dados, o que pode levar a resultados enviesados. Já a imputação *hot-deck* primeiramente aplica alguma estratégia de agrupamento aos registros para e, em seguida, aplica a imputação simples sobre cada grupo resultante. Com isso, a imputação *hot-deck* preserva a distribuição dos registros observados e garante que os valores imputados estejam dentro de um intervalo válido [Marker et al., 2002]. Diversas estratégias de agrupamento podem ser usadas na imputação *hot-deck*.

3. Avaliação Experimental

Nesta seção, apresentamos a avaliação experimental, considerando resultados relativos à precisão dos valores imputados e aos tempos de imputação.

3.1. Conjuntos de Dados

Levar em consideração a análise da correlação entre os atributos é importante para o processo de imputação. Altas correlações tendem a favorecer o resultado da imputação [Soares, 2007]. Neste sentido, os testes foram realizados em conjuntos de dados com diferentes graus de correlação entre seus atributos: (i) O *Breast Cancer Wisconsin (Original) Data Set*, possui registros referentes ao diagnóstico de câncer de mama e foram coletados no hospital de Winscosin. (ii) No *Pima Indians Diabetes*, estão registros referentes aos integrantes de uma tribo indígena conhecida por grande parte dos seus integrantes possuírem diabetes *mellitus*. (iii) Já em *AIDS Deaths - National Health and Family Planning Commission of China*, os dados foram coletados a partir do relatório mensal da Comissão Nacional de Saúde e Planejamento Familiar da China [Nan and Gao, 2018].

Para que os resultados fossem apresentados em uma mesma ordem de grandeza, foi aplicada a normalização *Min-Max* nos conjuntos de dados. Os registros que previamente possuíam valores nulos foram previamente retirados dos conjuntos de dados (*listwise deletion*), para utilizar o conjunto de dados completo, com vistas a potencializar a avaliação dos métodos empregados. A Tabela 1 apresenta o sumário dos conjuntos de dados.

Tabela 1. Sumário dos Conjuntos de Dados

Conjunto de Dados	Atributos	Registros	Registros após remoção dos valores ausentes
Pima Indians Diabetes	9	768	336
Breast Cancer	11	699	683
AIDS	33	78	78

3.2. Experimentos

A implementação da solução foi realizada com o *framework Spark*, versão 2.4.5 em um *cluster* composto por um nó *master* e 3 nós *workers*. Cada nó formado por um processador *Intel Xeon Platinum série 8000*, *quad-core* com 3,1 GHz e 16GB de memória RAM. O sistema operacional utilizado foi o *Linux Red Hat* 7.3.1-6 com *Kernel* versão 4.14.173. Para a execução dos comitês foi utilizada a biblioteca *Spark-Ensemble* [1].

Os percentuais de ausência foram gerados de forma artificial e aleatória. Sequencialmente, um atributo alvo é escolhido, as ausências são provocadas nele de acordo com um percentual, a etapa da imputação é iniciada, e após o término, o processo é repetido no atributo seguinte. Esses percentuais variaram de 10% até 30% com saltos de 10% seguindo o mecanismo de ausência MCAR, de acordo com o proposto por Soares [2007].

A imputação *hot-deck* implica na utilização de algum algoritmo para agrupar os valores do conjunto de dados. Utilizamos o algoritmo *k-Means* devido à sua simplicidade

[1] https://github.com/pierrenodet/spark-ensemble

e bom desempenho [Patil and Karthikeyan, 2020]. A medida de distância utilizada para o agrupamento com o *k-Means* foi a euclidiana: $d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$. Para identificar o valor ideal do parâmetro k no *k-Means*, utilizamos a técnica *elbow* [Syakur et al., 2018]. Este método testa a variância dos dados em relação ao número de grupos. Um valor ideal de k ocorre quando o aumento no número de grupos não produz um ganho significativo na qualidade dos grupos gerados.

Tanto para *Adaboost* quanto para *Bagging*, utilizamos *Árvores de Decisão* como componentes, com a profundidade máxima de cada árvore definida como 10. A quantidade de componentes por comitê foi igual a cinco.

Para identificar o grau de correlação entre os atributos dos conjuntos de dados, foi utilizada a técnica de *pearson* que é representada pela seguinte fórmula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

A análise das matrizes de correlação dos atributos dos conjuntos de dados indica que tanto *Breast Cancer* quanto *Aids* possuem alto percentual de correlação, pois seus atributos apresentam graus maiores do que 50%. Já a base *Pima Indians Diabetes* apresenta um percentual baixo de correlação. A Tabela 2 ilustra estes números.

Tabela 2. Consolidado dos coeficientes de correlações por conjunto de dados.

Conjunto de Dados	Percentual de Correlações Maiores que 50%	Nível de Correlação
Pima Indians Diabetes	9,38%	Baixa
Breast Cancer	78,00%	Alta
Aids	71,04%	Alta

A métrica utilizada em cada atributo foi a *Root Mean Squared Error (RMSE)* e para consolidar o erro observado por conjunto de dados, foi calculado a média dos erros (RMSE) obtidos em seus atributos. As Figuras 1, 2 e 3 apresentam o comparativo destes erros consolidados entre as imputações simples e *hot-deck*, e entre os algoritmos *Adaboost* e *Bagging*. A imputação *hot-deck* obteve os melhores resultados em todos os cenários nos três conjuntos de dados quando comparado à imputação simples. Na comparação dos resultados com *hot-deck*, *Adaboost* perde para *Bagging* somente no cenário da imputação em *Breast Cancer* com 10% de ausência. A maior diferença de resultado ficou em *Pima Indians Diabetes* com 30% de ausência. A Tabela 3 apresenta a média dos resultados obtidos na imputação hot-deck. Quando comparado o tempo de processamento, *Bagging* ganha de *Adaboost* em todos os cenários.

4. Conclusão

Neste artigo, apresentamos um estudo experimental comparativo sobre o processo de imputação de dados utilizando as técnicas imputação simples e *hot-deck* com comitês regressores *Adaboost* e *Bagging*. Do ponto de vista da precisão do dado imputado, os resultados revelam que *hot-deck* supera a imputação simples em todos os cenários testados. Com relação as comparações apenas entre *hot-deck*, *Adaboost* apontou resultados

Tabela 3. Média do erro (RMSE) da imputação hot-deck entre os algoritmos Adaboost e Bagging

	Breast Cancer				Pima Indians Diabetes				Aids			
	Imputação Simples		Imputação Hot-Deck		Imputação Simples		Imputação Hot-Deck		Imputação Simples		Imputação Hot-Deck	
Ausência	Adaboost	Bagging	Adaboost	Bagging	Adaboost	Bagging	Adaboost	Bagging	Adaboost	Bagging	Adaboost	Bagging
10%	0,23	0,21	0,20	0,19	0,35	0,27	0,23	0,25	0,14	0,17	**0,13**	0,14
20%	0,25	0,22	**0,18**	0,19	0,37	0,30	0,23	0,23	0,14	0,16	**0,13**	0,15
30%	0,24	0,21	**0,18**	0,20	0,35	0,32	**0,21**	**0,21**	0,15	0,17	0,14	0,15

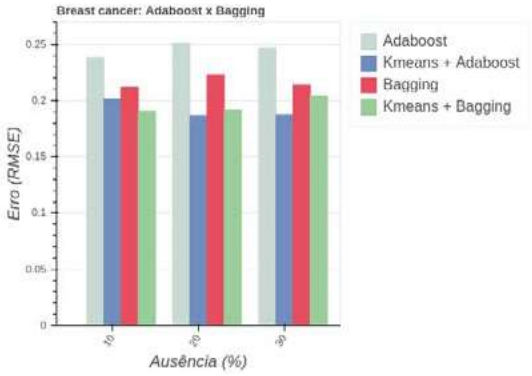

Figura 1. Média do erro (RMSE) em *Breast Cancer* com imputação Adaboost, Bagging e hot-deck

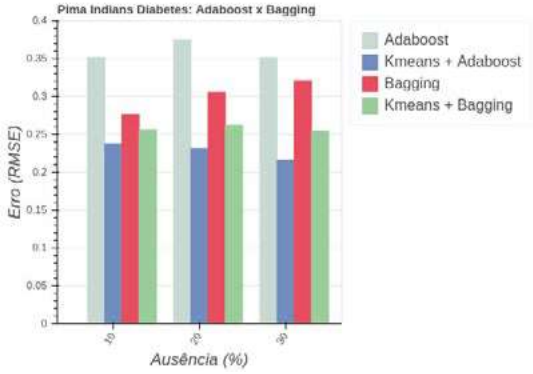

Figura 2. Média do erro (RMSE) em *Pima Indians Diabetes* com imputação Adaboost, Bagging e hot-deck

melhores que *Bagging*, salvo uma exceção, na imputação com 10% de ausência em *Breast Cancer*. Comparando os resultados entre todos os conjunto de dados, *Aids* mostrou os melhores resultados, seguido de *Breast Cancer* e *Pima Indians Diabetes*. No tempo de execução, *Bagging* superou *Adaboost*, na média, em todos os cenários testados, exceto nas imputações com 10% e 20% de ausência em *Aids*. Porém, *Adaboost* ainda se mostra uma alternativa razoável visto a maior média de tempo ser 333 segundos, observados na imputação com 20% de ausência no conjunto com a maior quantidade de dados: *Breast Cancer*.

A utilização de comitês de regressão na imputação *hot-deck* se mostrou uma alternativa promissora tanto no ponto de vista da precisão da reconstituição dos dados ausentes, quanto do ponto de vista do tempo de processamento. Trabalhos futuros consistem na comparação entre *Random Forest* e *Bagging*, utilização de outros comitês como o *Gradient boost* e o *Stacked Generalization*, assim como o uso de outros componentes.

Tabela 4. Média de tempo em segundos da imputação hot-deck entre os algoritmos Adaboost e Bagging

	Breast Cancer				Pima Indians Diabetes				Aids			
	Imputação Simples		Imputação Hot-Deck		Imputação Simples		Imputação Hot-Deck		Imputação Simples		Imputação Hot-Deck	
Ausência	Adaboost	Bagging	Adaboost	Bagging	Adaboost	Bagging	Adaboost	Bagging	Adaboost	Bagging	Adaboost	Bagging
10%	43	18	226	212	39	17	232	138	42	36	86	94
20%	35	17	333	250	29	15	245	160	45	37	89	105
30%	36	17	305	154	30	14	296	111	48	29	89	88

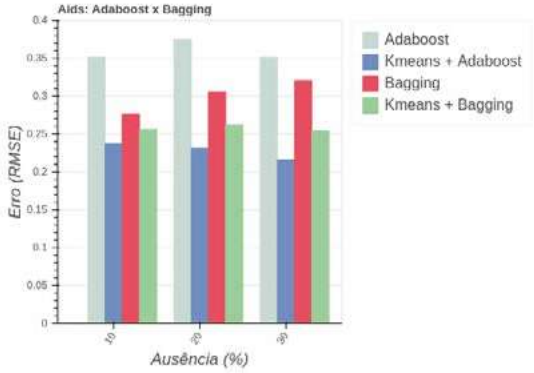

Figura 3. Média do erro (RMSE) em *Aids* com imputação Adaboost, Bagging e hot-deck

Referências

Samuel Zico Christopher, Titin Siswantining, Devvi Sarwinda, and Alhadi Bustaman. Missing value analysis of numerical data using fractional hot deck imputation. In 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), pages 1–6. IEEE, 2019.

B Ford. An overview of hot-deck procedures, incomplete data in sample surveys, 1 theory and bibliographies, vol. 2, w. Academic Press, 3, 1983.

David A Marker, David R Judkins, and Marianne Winglee. Large-scale imputation for complex surveys. Survey nonresponse, 329341, 2002.

Yongqing Nan and Yanyan Gao. A machine learning method to monitor china's aids epidemics with data from baidu trends. PloS one, 13(7):e0199697, 2018.

Pratik Patil and A. Karthikeyan. A survey on k-means clustering for analyzing variation in data. In G. Ranganathan, Joy Chen, and Álvaro Rocha, editors, Inventive Communication and Computational Technologies, pages 317–323, Singapore, 2020. Springer Singapore. ISBN 978-981-15-0146-3.

Jorge Soares. Pré-Processamento em mineração de dados: Um Estudo Comparativo em Complementação. PhD thesis, COPPE/UFRJ - Engenharia de Sistemas e Computação, 2007.

Rodrigo Tavares Souza. Appraisal-spark: Uma abordagem para imputação em larga escala. Master's thesis, CEFET/RJ - PPCIC, 2019.

Rodrigo Tavares Souza, Rafael Castaneda, Claudia Ferlin, Ronaldo Goldschmidt, Luis V. Carvalho Alfredo, and Jorge de Abreu Soares. Apoiando o processo de imputação com técnicas de aprendizado de máquina. In 33rd Brazilian Symposium on Databases (SBBD), pages 259–264, 2018.

MA Syakur, BK Khotimah, EMS Rochman, and BD Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP Conference Series: Materials Science and Engineering, volume 336, page 012017. IOP Publishing, 2018.

Cha Zhang and Yunqian Ma. Ensemble machine learning: methods and applications. Springer, 2012.