International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:8, 2013

A Testbed for the Experiments Performed in Missing Value Treatments

Dias de J. C. Lilian, Lobato M. F. Fábio, and de Santana L. Ádamo

***Abstract*—**The occurrence of missing values in database is a serious problem for Data Mining tasks, responsible for degrading data quality and accuracy of analyses. In this context, the area has shown a lack of standardization for experiments to treat missing values, introducing difficulties to the evaluation process among different researches due to the absence in the use of common parameters. This paper proposes a testbed intended to facilitate the experiments implementation and provide unbiased parameters using available datasets and suited performance metrics in order to optimize the evaluation and comparison between the state of art missing values treatments.

***Keywords*—**Data imputation, data mining, missing values treatment, testbed.

I. Introduction

THE occurrence of Missing Values (MV) in databases is considered a serious problem for the tasks of analyzing and mining the data; its incurrence usually causes problems of efficiency loss, and complications for data analysis [1]. In addition, the inappropriate treatment of MVs may also affect the classifier generalization [2], a crucial point in the process of knowledge discovery in database (KDD).

Specifically during classification tasks, the database learning process with MV becomes even more important [3], since the existence of such values in the training, validation or testing datasets could affect the accuracy of the classifiers modeled. Moreover, most classification algorithms cannot deal with incomplete databases directly [3]. This fact highlights the importance to treat such problem during the pre-processing phase.

The possible causes characterized for data absence, called Missingness Mechanism, i.e. what kind of event originates the MV, are [4]:

Missing Completely At Random (MCAR): occurs when the event is random so the missing values are independent from the observed values;

Missing At Random (MAR): the missing data depend on some observed and available value to be analyzed, so the missingness cause of those MV might be measured;

Missing Not At Random (MNAR): the most difficult to treat, since that the absence may depend on the observed and on the missing values found in the database.

Given the problems generated by the MV existence in databases, the following approaches have been used in literature, particularly in the area of data mining, as possible solutions [5], [6]:

Deletion of examples that contains MV in its attributes and/or complete removal of some attribute in case of a large amount of MV;

Maximum Likelihood procedures, this treatment performs parameters estimation of a model using a database without MV then, an imputation occurs through values sampling; missing value imputation. These methods aim to fill in the incomplete values with estimates using machine learning or statistical methods. They are divided in simple and multiple imputations.

During the present research, a wide bibliographic review was performed on MVT for KDD. The review, while not exhaustive fully demonstrated the lack of standardization of experiments; which brings difficulties to evaluation process of the approaches. Also, in this scope, a deficiency of replicable datasets was noticed as a critical factor.

From a sample of 40 recent articles on the subject of MVT, it was observed that only 12 of them used replicable dataset (i.e. available) while the others used synthetic or proprietary datasets (referred here as non-available). The list of these articles and the graphics of its can be found on http://linc.ufpa.br/liliandias/mvtestbed/.

The available datasets are known as the ones found in open repositories, such as [7]-[9], or through authors' supplementary materials.

On the other hand, the non-available datasets could be separated into proprietary and synthetic. The first are those originated from private or public companies with restriction of its usage, making the comparison of the results obtained from the experiments difficult.

The synthetics are those created from the available datasets, but their examples are randomly deleted in order to simulate the missing values. Therefore, it does not allow for new studies to use or compare these materials, since there is no assurance that a new simulation, using the same dataset, would remove the exactly same values. Using this method to insert missing values, the database generated is classified as MCAR.

Along with the previous problems, it was observed that the evaluation metrics used during the experiments for MVT (e.g. accuracy, error rate) may not indicate the real classifier generalization capability [10]. For example, when there are imbalanced classes in the database, one class can have more examples than others; which, while it could provide for a better recognition, by the classifier, for this particular class, it would also disguise the results of its generalization and

Dias de J. C. Lilian, Lobato M. F. Fábio, and de Santana L. Ádamo are with the Technological Institute of Federal University of Para, Brazil (phone: +55(91)8136-0158; e-mail: lilianchavesdias@gmail.com, lobato.fabio@ufpa.br, adamo@ufpa.br).

produce overall poor quality analyses. Some datasets containing such problem are yeast, page-blocks, glass and balance-scale [11], available on UCI repository.

Considering the MVT problems presented: the lack of availability of replicable databases and metrics that do not correctly describes the classifier performance, the present study proposes a methodology or "testbed", in order to facilitate the evaluation and replication of researches involving MVT.

This paper is organized as follows. In Section II are presented the related works that show the difficulties mentioned, as well as a justification to a testbed usage. Section III describes how this testbed present itself as a methodology. The final remarks are presented in Section IV.

II. Related Work

The main motivation of this study was the need of a guideline for the experiments performed during the MVT proposals. Therefore, the related works mentioned in this section are articles that stimulated this writing. Beyond these, are also described works focused on testbed as a solution for problems similar to those found in the MVT literature.

Some of the works on MVT show difficulties for evaluating its results, since there is a lack of availability of replicable databases and inappropriate use of performance metrics.

Considering the first MVT problem of lack of replicable databases, the works [12]-[14] can be mentioned. There, the authors used datasets without MV. However, to simulate data missingness in their tests they randomly delete samples; incapacitating, in this process, a replication of those datasets for other researchers. The other side of this problem is also when using proprietary/private datasets for simulations, as in [15], [16], since their use are restrict.

Regarding performance metrics, most works aim only to optimize classification systems, using the accuracy value or error rate as evaluation parameters, [17] and [18]. However, as mentioned previously, they may not efficiently describe the classifier's performance, since the database may have an imbalanced class distribution, causing distortion in its analyses.

In [19], other problems to replicate experiments are recognized. There are no explanations about the datasets generation or the used metrics becoming unpractical to compare its results with other approaches.

On the other hand, in [7] 15 MVT were evaluated in different classifiers types, with the objective of finding the best method for each. They used databases from UCI repository that already have MV. For methods comparison the Wilcoxon Signed Rank was used, based on the classifiers accuracy. Through these procedures, the study demonstrated experiments capable of replication and also showed how to realize comparison between recent treatments, becoming the basis for the formulation of this work.

The following section presents the proposal of a testbed [20], [21] also as a way to standardize MVT problems found during the comparison of different proposals and test execution.

III. Proposal

The present approach is composed of a methodology to viable the comparison among state of art treatments and, in addition, facilitate replications for new researches and general academic works.

Fig. 1 shows the testbed stages, highlighting its capacity of using different treatments and classifiers over the same dataset. The items compose the KDD process that a MV dataset must go through, in order to generate knowledge. To make a comparison between different MVT, the experiments must use the same input and generate information in the same format, characterized as the datasets and performance measures, respectively.

These two points comprises this methodology proposal. As showed on previous sections, both of them can make researches replication harder and are the causes of difficulties to direct comparison among different studies.

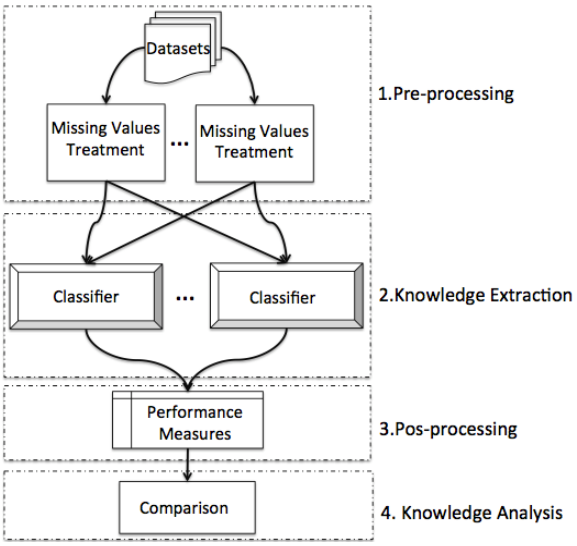

Fig. 1 Testbed Stages

Through the testbed it is perceived its wide applicability as well, enabling a performance comparison of different treatments and also its behavior in different classifiers and datasets.

A. Datasets

As observed in Section I, the major databases used in MVT are not replicable. For this reason, they are hardly used as input in other works preventing a comparison of the obtained results.

In this context, the methodology suggests public domain datasets, initially originated from the UCI repository for the classification task, due to its availability; encouraging the generation of common material to posterior comparisons. Table I lists some bases and its definitions: amount of instances; attributes; classes; and MV percentage. Emphasizing that these databases are reference to classification tasks, their data have domain missingness classified as MCAR [3] and a diversity of its characteristics as

well.

TABLE I
CHARACTERISTICS OF THE MISSING VALUE DATABASES

Name	Instances	Attributes	Classes	% M.V.
Audiology	226	71	24	1,98
Autos	205	26	6	1,11
Bands	540	40	2	4,63
Echocardiogram	132	12	4	4,73
Hepatitis	155	20	2	5,39
Horse-colic	368	24	2	21,82
House-votes-84	434	17	2	5,3
Lung-cancer	32	57	3	0,27
Mushroom	8124	23	2	1,33
Ozone	2534	73	2	8,07
Post-operative	90	9	3	0,37
Primary tumor	339	18	21	3,69
Soybean	307	36	19	6,44

By using these datasets, the methodology offers inputs to compare experiments between different MVT. Other point favored by the present proposal is its applicability to different classifiers, representing an implementation of the entire KDD process.

B. Performance Measures

During the post-processing phase in KDD, the evaluation values for each generated model are obtained. In MVT for classification systems, this estimation can be reached by the accuracy or error rate calculation, which cannot describe the model generalization capacity recognizing very well one class over others when the dataset is imbalanced.

In order to optimize this evaluation, the methodology proposes other more suited metrics for use, such as: confusion matrix analysis, sensitivity (1) and specificity (2). Calculated for each model applied to a given database, i.e. evaluation and performance metrics for classification tasks, describing the classes' generalization characteristics present in the database. The accuracy can be further calculated based on the results of sensitivity and specificity (3).

$$sensitivity = \frac{t\_pos}{pos} \quad (1)$$

$$specificity = \frac{t\_neg}{neg} \quad (2)$$

$$accuracy = sensitivity \frac{pos}{(pos+neg)} + specificity \frac{neg}{(pos+neg)} \quad (3)$$

where, *t_pos* and *t_neg* are the amount of examples classified as positive and negative respectively; *pos* and *neg* are the amount of examples that represents the positive and negative classes.

From those values, it is possible to extract parameters to evaluate and compare the treatment behavior for each class of a given database resulting in suited generalization analysis of the classifier. Therefore, the methodology proposes this metrics for results comparison of general academic works about MVT.

IV. CONCLUSION

In this study, difficulties to evaluate and replicate different approaches of missing values treatments were described, either due the usage of non-available bases or metrics that do not correctly describes the classifier performance.

In this context, the contributions of this study are twofold: the facilitation of the experiments execution by the databases showed, and an unbiased comparison between the state of art missing values treatments by the metrics mentioned.

Thus, the proposed testbed allows the evaluation and replication of experiments for missing values treatment and also covers the entire simulation process avoiding the problems and facilitating future academic researches about MVT.

REFERENCES

[1] Alireza Farhangfar, Lukasz Kurgan, and Witold Pedrycz, "A Novel Frameworl for Imputation of Missing Values in Databases," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, pp. 692-709, 2007.

[2] P. E. Mcknight, K. M. Mcknight, S. Sidani, A. J. Figueredo. *Missing data: a gentleintroduction. New York: The Guilford Press, 2007.*

[3] Julián Luengo, Salvador García, and Francisco Herrera, "On the choice of the best impuation methods for missing values considering three groups of classification methods," *Knowledge Information Systems*, pp. 1-32, 2011.

[4] R. Little and D. Rubin, *Statistical Analisys with Missing Data*, 2nd ed.: John Wiley and Sons, 2002.

[5] Alireza Farhangfar, Lukasz Kurgan, and Witold Pedrycz, "Experimental analysis of methos for imputation of missing values in databases," *Intelligent Computing: Theory and Applications II*, vol. 5421, pp. 172-182, 2004.

[6] Kamakshi Lakshminarayan, Steven A. Harp, and Tariq Samad, "Impuation of Missing Data in Industrial Databases," *Applied Intelligence*, pp. 259-275, 1999.

[7] A. Frank and A. Asuncion, UCI Machine Learning Repository, 2010, University of California, Irvine, School of Information and Computer Sciences.

[8] ACM Special Interest Group on Knowledge Discovery and Data Mining. ACM KDD CUP. [Online]. http://www.sigkdd.org/kddcup/index.php.

[9] Government of Canada. Open Data - Open Data Portal. [Online]. http://www.data.gc.ca/.

[10] Russel G. Congalton, "A Review of Assessing the Accuracy of Classification of Remotely Sensed Data," *Retome Sensing of Enviroment*, vol. 37, no. 1, pp. 35-46, 1991.

[11] Satyam Maheshwari, Jitendar Agrawal, and Sanjeev Sharma, "A new approach for Classification of Highly Imbalancade Datasets using Evolutionary Algorithms," *International Journal of Scientific & Engineering Research*, vol. 2, no. 7, pp. 1-5, 2011.

[12] Xiaofeng Zhu, Shichao Zhang, Zhi Jin, Zili Zhang, and Zhoumin Xu, "Missing Value Estimation for Mixed-Attribute Data Sets," *IEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 1, pp. 110-121, 2011.

[13] Chih-Feng Liu, Thao-Tsen Chen, and Shie-Jue Lee, "A Comparison of Approaches for Dealing With Missing Values," in *International Conference on Machine Learning and Cybernetics*, Xian, 2012.

[14] Phimmarin Keering, Werasak Kurutach, and Tossapon Boongoen, "Cluster-based KNN Missing Value Imputation for DNA Microarray Data," in *IEEE International Conference on Systems, Man, and Cybernetics*, Seoul, Korea, 2012, pp. 445-450.

[15] Y. Zhang, C. Kambhampati, D. N, Davis, K. Goode, and G. F. Cleland, "A Comparative Study of Missing Value Imputation with Multiclass Classification for Clinical Heart Failure Data," in *9th International Conference on Fuzzy Systems and Knowledge Discovery*, 2012, pp. 2840-2844.

[16] Xiaoling Lu, Jeisheng Si, Lanfeng Pan, and Yanyun Zhao, "Imputation of Missing Data Using Ensemble Algorithms," in *8th International*

Conference on Fuzzy Systems and Knowledge Discovery, 2011, pp. 1312-1315.

[17] Ludmila Himmelspach and Stefan Conrad, "Clustering Approaches for Data with Missing Values: Comparison and Evaluation," in *International Conference on Digital Information Management*, 2010, pp. 19-28.

[18] Lars Wohlrab and Johannes Fürnkranz, "A review and comparison of strategies for handling missing values in separate-and-conquer rule learning," *Intelligent Information Systems*, vol. 36, pp. 73-98, 2011.

[19] Dipak V. Patil and R. S. Bichkar, "Multiple Imputation of Missing Data with Genetic Algorithm based Techniques," *IJCA Special Issue on "Evolutionary Computation for Optimization Techniques"*, pp. 74-78, 2010.

[20] Jeff Struckman and James Purtilo, "A testbed for evaluation of web intrusion prevention systems," in *3rd International Workshop on Security Measurements and Metrics*, 2011.

[21] Uttam Adhikari et al., "Development of Power System Tes Bed for Data Mining of Synchrophasors Data, Cyber-Attack and Relay Testing in RTDS," in *IEEE Power and Energy Society General Meeting*, 2012, pp. 1-7.