Proceedings of the 2018 International Conference on Machine Learning and Cybernetics, Chengdu, China, 15-18 July, 2018

A MISSING DATA IMPUTATION METHOD WITH DISTANCE FUNCTION

KUEN-FANG JEA, CHIH-WEI HSU, LI-YOU TANG

Department of Computer Science and Engineering, National Chung-Hsing University, Taichung 40227, Taiwan, R.O.C.
E-MAIL: kfjea@cs.nchu.edu.tw, s9556001@cs.nchu.edu.tw, g104056048@mail.nchu.edu.tw

Abstract:

"Missing data" is an important research issue in big data analysis. This problem may cause data hard to analyze precisely. In recent research, several imputation-based methods have been proposed to solve the missing data issue without using domain knowledge. Among them, the missing data imputation method based on association rule mining was proposed to determine which value should be filled in the missing data. However, the generated rules may not always be suitable for filling in missing values. For example, some strong rules may fill up different missing values with the same result. We propose here an algorithm named RID (Rule-based Imputation with Distance function) to deal with this shortcoming. RID generates rules for missing data imputation by association rule mining and then uses a distance function to adjust the rule to fill in values appropriately. Experimental results show that the accuracy of RID is approximately 3 to 5 percentage higher than those of C4.5 and kNN, and approximately 6 to 7 percentage higher than that of HMiT.

Keywords:

Big data; Missing data; Association rules; Distance function; Imputation

1. Introduction

In big data analytics, missing data is an important issue that needs to solve. This issue plays an important role in the analysis of big data that has emerged in recent years. There are many reasons why data may be missing, such as incorrect data format, typos, or malfunction of data gathering equipments. This problem can cause data hard to analyze precisely.

For the missing data problem, most solutions can be divided into two categories: deletion methods and imputation-based methods. The deletion methods [1] discard the records containing missing values, and it only analyzes the rest of data. The deletion methods are the simplest and most commonly used techniques in data processing. However, the more the data loses the more inaccurate the analysis is. On the other hand, the imputation-based approaches [1] attempt to interpolate missing data according to the possible relationships among the data in the dataset. The imputation-based methods can be divided into two types depending on whether using a data distribution model or not. Without the distribution model, the relationship of data may be ignored, so the imputation results may not be good. The method based on the distribution model usually requires domain knowledge to accurately establish the model. This type of methods can provide good imputation results when the data matches the distribution model.

With the advance in machine learning research, many machine learning-based imputation methods have been proposed to solve the missing data problem. These methods use machine learning techniques to find rules from the input data for estimating the possible value of missing data. This kind of methods does not require domain knowledge to build the data distribution model, which can reduce the cost of building models and provide good imputation results. Commonly used methods include K-nearest neighbor based imputation methods, decision tree based imputation methods and association rule mining based imputation methods.

Related research on the processing of missing values in association rule mining refers to the processing methods and possible problems. Association rule mining is a method that can find the correlation among data in large databases. Based on this method, missing values can be estimated by looking for association rules between the attributes of the complete data in the input data and choosing a higher confidence rule for the missing data. However, high confidence rules may not suit all missing data. Especially for low-frequency records containing missing data, they are easy to cause misjudgment by applying a strong rule to interpolate the value.

In order to solve the picking inappropriate rules problem of using association rule mining, this study

978-1-5386-5214-5/18/$31.00 ©2018 IEEE

proposes an association rule based imputation method, named Rule-based Imputation with Distance function (abbreviated as RID). RID fills in the missing parts of the data set through the association rule mining of the complete data and uses a lower threshold to allow association rules to cover more relevant relationships of data. In view of the possible selection of wrong rules, we use a distance function to increase the weight of higher similarity data, thereby reducing the impact caused by the rules of lower similarity data. We also compare the proposed method with the existing commonly used machine learning-based missing data imputation methods.

The rest of this paper is organized as follows. In Section 2, some related work about the missing data issue is described. Section 3 proposes an association rule mining based method with distance function for missing data imputation. Section 4 presents experimental results and their analysis. Finally, Section 5 concludes this research.

2. Related work

The current solutions to missing data can be divided into two categories [1]: deletion methods and imputation-based methods. The deletion methods discard the data with missing value and only analyze the data without missing values. The imputation-based methods use the relationship between complete data in the dataset to estimate missing data and replace the missing values with estimations. The imputation-based methods can be further divided into two categories. One is that does not use a preset model to characterize the attribute with missing values, and the other is that uses a preset model.

The imputation-based method that does not use a preset model includes instead some traditional statistics-based methods [1] such as Mean Imputation and Regression Imputation. Mean Imputation fills in the arithmetic mean into the missing value based on the other existent values of the attribute containing missing values. Regression Imputation fills missing values with the predicted values based on a regression analysis of the data. This kind of methods does not ignore the data of missing values as compared to the deletion methods, but it fails to take into account the relationship between other attributes or data, resulting in a lower imputation effect.

A. P. Dempster et al. [2] proposed two imputation-based method using models, including Maximum Likelihood and Multiple Imputation. The Maximum Likelihood method assigns a preset distribution model to the attribute of missing values and performs the Maximum Likelihood estimation of the missing values by observing the distribution of the data. The parameter estimation method used for missing data is Expectation Maximization. It consists of two steps: the first step is to calculate the Maximum Likelihood of the missing attribute values, and the second step is to maximize the estimation and use it for the next iteration. The algorithm continues to iterate between the two steps until the estimation converges. Multiple Imputation creates several reasonable sets of estimated data and then uses standard statistical methods to fit each data set with an adapted model and calculate the estimated values. Finally, the estimated values of each data set are combined for filling up. Since these two methods take into account the distribution of the data, both can produce valid estimates if the data attribute can be assigned a correctly preset model. The disadvantage is that the filling effect depends on whether the choice of the default model is correct. And further the model is not easy to build, since it usually requires the assistance of a professional with domain knowledge to have a good effect.

With the development of machine learning technology, many imputation-based methods [3-7] based on machine learning have been proposed to deal with the missing data problem. Such methods do not require assumptions about the distribution of the attribute data, and they can reduce the impact of inaccurate modeling assumptions. The commonly used methods, for example, fill missing values through Decision Tree, K-Nearest Neighbors, and association rule mining. The Decision Tree-based method recursively branches the input data to construct a tree, divides the data into two or more leaf nodes by identifying the the attribute value through a mathematical function (e.g., Entropy, Information gain or GINI index, etc.). The method repeats on each leaf node until the entire tree is constructed. The classification performance of this method largely depends on the structure of the tree, and due to the continuous identification of each attribute, overfitting may be a critical issue of missing data imputation.

O. Troyanskaya et al. [5] proposed K-Nearest Neighbors imputation to deal with missing values in DNA microarrays. This method uses the distance calculation to select the k-distance data similar to the missing data from the complete data. This k-letter data is called the k nearest neighbors. The average value is calculated in the k nearest neighbors, and the average value is used as the padding value of the missing data. Common distance measures include Pearson correlation and Euclidean distance. D. R. Wilson [8] proposed Heterogeneous Euclidean-Overlap Metric for distance calculation. The first disadvantage of this type of methods is that, whenever looking for the nearest neighbors (i.e., the most similar points), the

algorithm searches the entire data set for the neighbors, which can be very time-consuming for a huge amount of data. The second disadvantage is the difficulty in choosing the right k value. If the value of k is too small, it will overemphasize the data of a few advantages, leading to a decrease in accuracy; on the other hand, the nearest neighbors with too large k will contain data significantly different from the missing data, leading to the accuracy dropped.

R. Agrawal et al. [9] first proposed the Apriori algorithm, a systematic association rule mining algorithm. The Apriori algorithm uses pruning techniques to avoid dealing with infrequent itemsets, increasing the efficiency of candidate itemsets generation. But there are two problems: one is using a large amount of time, space, and memory to form a candidate set of items, and the other is the need of scanning the database multiple times.

A. Ragel et al. [3] mentioned two problems with the association rule mining of datasets containing missing data. One is that it will lead to a reduction in the support of itemsets and the confidence of the rules due to the presence of missing values. Although the available rules can be obtained by lowering the minimum confidence and the minimum support, it is relatively difficult to handle the huge amount of low threshold data. Another problem is that even if the rules have high confidence, the rules still may not be appropriate to fill the missing data. And at low thresholds, many rules may be found. It is difficult to select useful rules to fill the missing data. In order to solve this problem, Ragel et al. [3] proposed a new method to remove the data containing missing values from the input data set, and to redefine the support and confidence of the rules according to the new data set. Ragel et al. also proposed a new parameter to evaluate the availability of rules. Through this method, the impact of missing values on rules can be reduced, and the association rule mining can achieve good results in a data set with missing data.

S. Bashir et al. [10] proposed the Hybrid Missing Values Imputation Technique (HMiT) to overcome the limitations of the prebuilt model in the existing methods for processing missing data and to solve the problem of the K-Nearest Neighbors method that needs a large amount of processing time to find neighbors. The method first finds the rule set through association rule mining. If there are rules that can be used to fill in the missing values, the average of the values derived from all the available rules is filled in the missing values. If there are no available rules, the missing data can be filled by using the K-Nearest Neighbors method.

3. Proposed method

The association rule mining imputation methods sometimes may pick wrong rules to fill missing values. We have the following ideas to improve them. First, from a large number of association rules being discovered, we select the rules with the highest confidence for each attribute to form a special rule set named imputation rule set. We filter out all records in the data set that match the imputation rule set. Then we use a distance function to evaluate the similarity between the rules in the imputation rule set and the data records with missing values. Because the association rules produced through the minimum confidence may only be generated by some attributes, it is easy to ignore the correlation of other attributes. So we use a distance function as the metric for evaluating the similarity between the association rules and the missing data in order to find out the appropriate rules for filling missing values.

We propose an algorithm, called Rule-based Imputation with Distance function (RID), to solve the above problem. RID generates association rules for missing data imputation and adjusts the filling-in values for missing data by using a distance function. RID can generate more appropriate rules for filling the missing values, and it also reduces the impact of other rules that are not applicable to missing data, thus improving the accuracy of filling the missing values. Table 1 lists the definition of notations used in the proposed method. The proposed method is shown below.

TABLE 1. The definition of notations

notation	description
InputDataset	Input dataset which contains missing data.
MA	Attribute with missing values in some data records.
CD_{MA}	Complete dataset without missing data in MA.
MD_{MA}	Dataset with missing values in some MA.
MinSup	Minimum support for mining association rules.
MinConf	Minimum confidence for mining association rules.
RuleSet	All association rules whose right part matches MA.
CRuleTable	Rule set generated from data in CD_{MA}.
MRuleSet	Rule set generated from data in MD_{MA}.
CChooseDataset	Set of data matching CRuleTable and MRuleSet.
OutputDataset	Imputation result of InputDataset.

RID Algorithm
Input: InputDataset、MinSup、MinConf、MA
Output: OutputDataset
Step 1: For each tuple t in InputDataset, if MA of t is missing, then move t to MD_{MA}; otherwise, move t to CD_{MA}.
Step 2: Generate the association rules for CD_{MA} according to MinSup and MinConf.
Step 3: For each rule r generated in Step 2, move r to RuleSet if the right part of r is MA.
Step 4: For each tuple c_i in CD_{MA}, create CChooseRuleSet containing all rules whose left part is c_i.
Step 4.1: For each rule $crule_x$ with highest confidence in CChooseRuleSet, insert tuple (c_i, $crule_x$) into CRuleTable.
Step 5: For each attribute m in MD_{MA}, do steps 6 – 12
Step 6: For each rule r in RuleSet, add r to MChooseRuleSet if the left part of r is equal to m.
Step 6.1: For each rule $mrule_x$ with highest confidence in MChooseRuleSet, add $mrule_x$ to MRuleSet.
Step 7: For each rule $crule_x$ in CRuleTable, if $crule_x$ can be found in MRuleSet, then add the corresponding c_i of $crule_x$ to CChooseDataset.
Step 8: Calculate d_i for each h_i in CChooseDataset, where d_i is the number of different attributes between h_i and m.
Step 9: Let d_{min} be the minimum of all d_i, and d_{max}be the maximum of all d_i.
Step 10: Calculate w_i for each h_i in CChooseDataset;
if d_{max} equals d_{min}, then w_i= 1;
w_i=(d_{max}- d_i)/(d_{max}- d_{min}) , otherwise.
Step 11: Separate all h_i into several groups according to the value in MA.
Step 11.1: Let sum_j be the summation of w_i which is corresponding to h_i in group j.
Step 12: Substitute the missing value of m with the value in MA of group j, which has the maximal sum_j of all groups.
Step 13: Merge MD_{MA} and CD_{MA} as OutputDataset and output.

The RID algorithm proceeds as follows. RID first divides InputDataset into two sets MD_{MA} and CD_{MA} according to whether or not MA has missing data in each tuple of InputDataset. Next RID obtains a rule set by performing the Apriori algorithm on CD_{MA} according to two predefined parameters MinSup and MinConf. RID creates RuleSet of the rules whose right part equals MA and left part does not contain a null value. RID then compares all rules in RuleSet for each c_i of the CD_{MA}. If the left part of a rule matches the attribute of c_i, then this rule is added into CChooseRuleSet for c_i. For each c_i, RID adds all corresponding rules with highest confidence in the CChooseRuleSet into CRuleTable. RID builds MRuleSet of MD_{MA} in a similar way. After MRuleSet and CRuleTable are built, RID stores the rules and corresponding data suitable for imputation in the CChooseDataset. Further RID calculates the distance d_i and the weighted distance w_i for each h_i in the CChooseDataset. The weighted distance formula used in RID is as follows, where d_{max} denotes the maximal one among all d_i, and d_{min} is the minimal one.

$$w_i = \begin{cases} \dfrac{d_{max} - d_i}{d_{max} - d_{min}}, & if\ d_{max} \neq d_{min} \\ 1, & otherwise \end{cases} \quad (1)$$

The RID algorithm selects the maximal value of the weighted results, which is used for filling the missing data. After all data in MD_{MA} has been filled, RID merges MD_{MA} with CD_{MA} into OutputDataset.

4. Experimental result

In this section, we present the experimental results of RID. We implement RID in Java on Microsoft Windows 10. The processor of the hardware is Intel(R) Core(TM) i5-6200U CPU @2.3GHz and memory is 8GB. We use 'Census income data set'[11] on UCI Machine learning repository site for our experiments. This dataset contains 16,280 records with 15 different attributes. We discretize numerical data for association rule mining by using Equal-frequency interval method.

In order to evaluate the accuracy of missing data imputation, we randomly mark data as 'missing' in the dataset and define the percentage of the records containing missing values in the dataset as *data missing rate*. The accuracy of imputation result is defined as follow.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

where TP denotes True Positive and FN denotes False Negative.

The experimental results are shown in Fig. 1. We compare RID with some related work including k-NN based imputation method (denoted as kNN), C4.5 based imputation method (decision tree based, denoted as C4.5), and HMiT (a hybrid of association rule and k-NN).

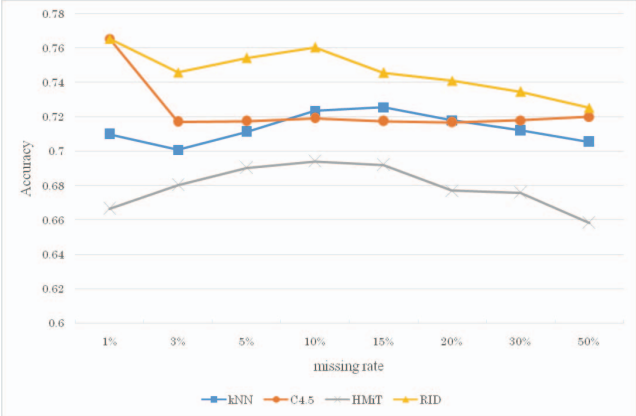

FIGURE 1. Accuracy comparison among RID and other methods under different data missing rate.

It is observed from Figure 1 that RID performs better than other methods in different data missing rates. The accuracy of RID is approximately 3 to 5 percentage higher as compared with C4.5 and kNN, and approximately 6 to 7 percentage higher as compared with HMiT. RID generates a larger number of low confidence rules and uses a distance function to pick appropriate ones for missing values. HMiT just uses high confidence rules to fill missing values and ignoring low confidence rules. However, sometimes the appropriate rules to fill in missing data are composed of low confidence rules rather than high confidence rules.

It can also be observed in Figure 1 that the higher the data missing rate the more the accuracy loses in all methods. Because when the data missing rate is high, the number of complete data records available for generating rules to fill missing data is less. RID uses a distance function to pick the low confidence rules applicable, making the decreasing rate of accuracy slower than other methods.

We also test the accuracy of RID under different minimum confidence thresholds. The experimental results are shown in Figure 2.

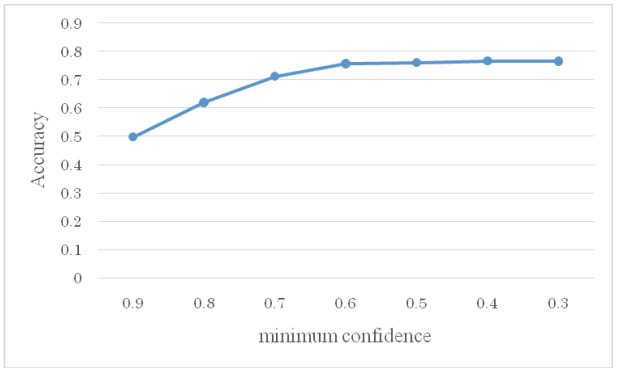

FIGURE 2. The accuracy of RID under different minimum confidence thresholds.

It can be observed from Figure 2 that, when the minimum confidence is high, the accuracy is low, and vice versa. When the minimum confidence is 0.6 or below, the increasing rate of accuracy gets slow down. The reason for this phenomenon is that under high confidence thresholds, there may have no appropriate rules available, resulting in low accuracy of missing values imputation. In this data set, if a certain minimum confidence is reached (i.e. 0.6), the generated rules can cover almost all of the data, so the impact of reducing the confidence level will be small. However, for lower confidence, the accuracy of RID will not continue to rise indefinitely. It is because RID uses a distance function to evaluate the similarity between rules and missing values. The rule set generated under low confidence includes more attributes than that under high confidence. The similarity between data and the rules generated below a certain confidence threshold is too low to be applicable to fill missing values. So it makes the accuracy of RID stop rising when the minimum confidence reaches a certain value.

5. Conclusion

In this study, we aim at the problem of handling missing data by using association rules mining that may pick up wrong rules to fill in missing values. We design the RID algorithm to overcome the problem. RID first mines association rules from the part of data set without missing values and generates a rule set with the highest confidence as the imputation rules for filling missing data. We use a distance function to handle the problem of picking up wrong rules in association rule based imputation algorithms. RID calculates the distance between the data without and with missing data according to the selected rules and gives different weights to measure the similarity

of possible imputation values. By this way, RID can reduce the impact of the data with low similarity and select better rules for imputation, thus improving the accuracy of missing data imputation.

References

[1] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 39, no. 1, pp. 1-38, 1977.

[3] A. Ragel and B. Cremilleuz, "Treatment of missing values for association rules," (in English), *Research and Development in Knowledge Discovery and Data Mining,* vol. 1394, pp. 258-270, 1998.

[4] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of missing data in industrial databases," (in English), *Applied Intelligence,* vol. 11, no. 3, pp. 259-275, Nov 1999.

[5] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," (in English), *Bioinformatics,* vol. 17, no. 6, pp. 520-525, Jun 2001.

[6] C. Anagnostopoulos and P. Triantafillou, "Scaling out big data missing value imputations: pythia vs. godzilla," presented at the Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, New York, USA, 2014.

[7] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning *k* for kNN Classification," *ACM Trans. Intell. Syst. Technol.,* vol. 8, no. 3, pp. 1-19, 2017.

[8] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," (in English), *Journal of Artificial Intelligence Research,* vol. 6, pp. 1-34, 1997.

[9] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, 1994, vol. 1215, pp. 487-499.

[10] S. Bashir, S. Razzaq, U. Maqbool, S. Tahir, and A. R. Baig, "Using association rules for better treatment of missing values," presented at the Proceedings of the 10th WSEAS International Conference on Computers, Athens, Greece, 2006.

[11] UCI Machine learning repository. *Census income data set*. Available: https://archive.ics.uci.edu/ml/datasets/Census+Income