MISSING VALUE IMPUTATION USING VOTING BASED OPTIMIZED ASSOCIATIVE RULE MINING ALGORITHM

Dr. S. Hemalatha

Abstract

In real time applications, imputation of missing value is a definite and tough problem confronted by machine learning and data mining. As a result, there are many attempts to missing value imputation. To overcome this respective issue, the experimental study has been carried out using continuous and discrete from UCI repository to shows that the proposed work is well effective than the other existing systems. Voting is made to find out the best candidate having the highest vote is finally chosen as the imputed value and the proposed system increases the accuracy rate and greatly reduces the error rate. The proposed voting based optimized ARM algorithm outperforms with less RMSE values than Robust Associative Rule Mining Algorithm for 80% missing rate respectively.

Keywords : Root Mean Square Error and Accuracy rate, ARM, voting based optimized ARM, Data Mining.

I. Introduction

Missing values usually appears either as null or empty cells in database in the forms of table as .arff files which can be identified. But can also appear as outliers or wrong data [6]. Natural datasets often have missing values in them. The imputation of missing values accurately is an important data preprocessing task. A number of techniques have been proposed in literature for imputing missing values [9, 11]. An early approach is to use the mean of all available values of the attribute, having a missing value [5], as an imputed value. Another initial approach to deal with missing values of a data set is to remove the records having missing value/s [1]. Mean imputation can generate more erroneous outputs than the simple record deletion approach. On the other hand, the deletion of records may also reduce the usefulness of a dataset for a statistical analysis and data mining [3].

A more advanced approach called Expectation–Maximization (EM) [10] relies on correlations between attributes and mean values of the attributes. A record may have some missing values and available values. EM considers that the deviation of a missing value (belonging to an attribute) from the mean value of the attribute is proportional to the deviation of an available value (belonging to another attribute) from the mean value of the second attribute. In order to impute a missing value EM uses the whole dataset, whereas another recent technique called IBLLS [8] divides a dataset in both horizontal and vertical segments by identifying k-most similar records and correlated attributes, respectively. It then

Asst. professor, Department of Computer Science, Karpagam University, Coimbatore

E.mail : drhemashanmugam@gmail.com

locally imputes a missing value by applying a framework [2] within each segment separately. However, EM and IBLLS impute only the numerical missing values.

This paper presents an overview of missing values problem and strategies for dealing data with missing values. Core part of the paper concentrates on hybrid method and the results are discussed briefly. Since the existing ARM algorithm was incapable to impute missing values in case of there were no appropriate association rule. This issue was managed by combining association rules and most common attribute value [4]. First was used in association with the rules approach and if there was no appropriate association rule then the most common attribute value was used. It can be solved by enhancing the missing values imputation accuracy through better combination of association rules and the most common attribute value method. Only association rules with confidence not lower than the relative frequency of occurrence of the most common value of the attribute were involved. To achieve imputation, the rest of missing values and the most common attribute value method was used. This technique was according to the fact that the most common attribute value can be managed as a special zero attribute rules. The antecedent of this rule contains no attributes and consequent contains one attribute. Support and confidence of this rule is equal to the relative frequency of the most common value of the attribute. Proposed voting system is used to find out the best candidate who has the highest vote is finally chosen as the imputed value.

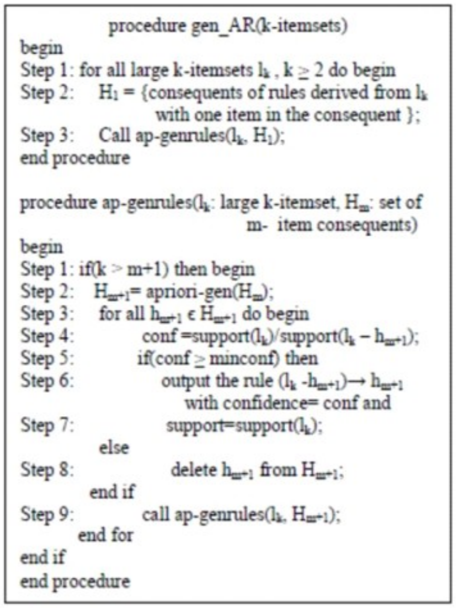
This paper reviews the problems that are caused by missing values and ways to solve the same. Section 2 describes the methodology involved in the heuristic approach to identify and eliminate the missing values in heterogeneous dataset. In section 3 we discuss the comparison made on both the existing and proposed algorithm followed by conclusion in section 4.

II. Methodology :

This paper presents an effective missing value imputation through association rule mining, based on voting method using multiple minimum supports that ensures the filling of missing values found in the dataset before intended analysis has been made.

2.1 Procedure to Construct ARM with Minimum Support

Construction of Association rule is a straight forward method. This technique uses the following procedure which is proposed in [7].



```
procedure gen_AR(k-itemsets)
begin
Step 1: for all large k-itemsets $l_k$ , $k \geq 2$ do begin
Step 2:    $H_1$ = {consequents of rules derived from $l_k$
                  with one item in the consequent };
Step 3:    Call ap-genrules($l_k$, $H_1$);
end procedure

procedure ap-genrules($l_k$: large k-itemset, $H_m$: set of
                                m- item consequents)
begin
Step 1: if(k > m+1) then begin
Step 2:    $H_{m+1}$= apriori-gen($H_m$);
Step 3:    for all $h_{m+1} \in H_{m+1}$ do begin
Step 4:          conf =support($l_k$)/support($l_k - h_{m+1}$);
Step 5:          if(conf $\geq$ minconf) then
Step 6:             output the rule ($l_k$ -$h_{m+1}$)$\rightarrow h_{m+1}$
                      with confidence= conf and
Step 7:             support=support($l_k$);
           else
Step 8:             delete $h_{m+1}$ from $H_{m+1}$;
         end if
Step 9:          call ap-genrules($l_k$, $H_{m+1}$);
       end for
end if
end procedure
```

Figure 1. Construction of ARM with Minimum support

2.2 Proposed missing Value Imputation Using Voting Based Optimized ARM

After constructing the association rules it is necessary to find out the consequent parameter, so as to determine which value is better for imputing missing value. To obtain this, voting system has been involved. Let the record R_i has a missing value in attribute $A_j \in A$, i.e. R_{ij} is missing. Let l be the actual value in the pth attribute $A_p \in A$, i.e. $R_{ip} = l$. Let x, y, z are the consequent values which is determined from association rules and therefore, x, y, and z are the candidates for possible imputation. Let $A_p \in \{l, m, n\}$ use a voting system to attain the best candidate with highest vote and it is finally chosen as the imputed value. Let, C_{xl} be the co-appearance of x and l in the whole data set, and fl be the total number of appearances (frequency) of l in the whole dataset. is the vote in favor of x based on A_p considering only the available value ^{l} and e$V_x^{N,p}$ will be calculated as follows,

$$V_x^{N,p} = \frac{C_{xl}}{f_l} \qquad (1)$$

x is the vote in favor of x based on A_p considering the available value along with its similar values. That is, $V_x^{S,p}$ is calculated considering l, m and n as follows.

$$V_x^{S,p} = \sum_{\forall a \in A_p} \frac{C_{xa}}{f_a} \times M_{la}^{p} \qquad (2)$$

where M_{la}^{p} is the mutual information between l ($R_{ip} = l$) and a of the pth attribute. We then calculate the weighted vote V_x^p in favor of x based on attribute A_p as follow.

$$V_x^p = \{V_x^{N,p} \times \lambda + V_x^{S,p} \times (1- \lambda)\} \times k_{jp} \qquad (3)$$

where k_{jp} is the correlation between the jth and the pth attributes. We now calculate the total vote V_x^T in favor of x by considering all attributes ($A = \{A_1, A_2, ...,A_m\}$) except the jth attribute (since$^{x \notin A_j}$) as follows,

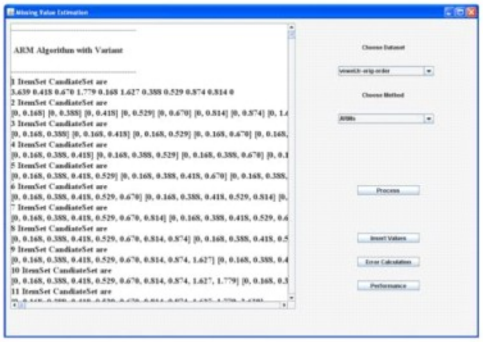
$$V_x^T = \sum_{\forall A_p \in A \setminus A_j} V_x^p \qquad (4)$$

Similarly, the total vote V_Y^T and V_Z^T can be calculated According to this total vote value, it is possible to choose the best value among x, y and z for imputing the missing value in dataset. In other words, the imputed value has highest vote and that value is used for the replacement of missing in the mixed attributes dataset. Finally, a complete dataset will be obtained.

III. Result and Discussion

The proposed system is compared with single minimum support value to build the association rules based on threshold value. Here the proposed system makes uses of multiple minimum support value. Minimum support values are calculated for each item in terms of minimum item support, the frequent item sets are generated. After finding the frequent item sets, the association rules are constructed using two datasets. The first dataset is known as training dataset which consists of data for constructing the association rules. The second dataset is referred as dataset for missing values imputation which includes the data with imputed values. Followed by the consequent association rules, voting system is performed. At last, the best imputed value will be found to acquire the imputed dataset.

Dataset has been taken form UCI repository to carry out this research work. It involves two type of dataset such as continuous data and discrete data. From the continuous data, two dataset are taken, namely Auto-mpg and Housing. From the discrete data, four dataset such as Abalone, Pima, Vowel and Anneal are taken. In the continuous data, Auto-mpg has the attribute as categorical and real attributes. This dataset consist of nearly 398 instances and 8 attributes. And Housing continuous data has attribute as categorical, integer and real attributes with 506 instances and 14 attributes. In the discrete data, Abalone has attributes such as categorical, integer and real attributes. It contains the 4177 instances and 8 attributes. For Anneal data, its attributes type also categorical, integer and real attributes. It consists of 798 instances and 38 attributes. In Pima data, the type of attributes is integer and real attributes and contains the 768 instances and 8 attributes. Vowel discrete data has only real attributes with 640 instances and 12 attributes.

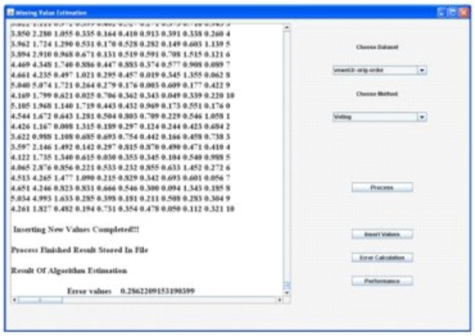


```
ARM Algorithm with Variant

1 ItemSet CandiateSet are
3.639 0.418 0.670 1.779 0.168 1.627 0.388 0.529 0.874 0.814 0
2 ItemSet CandiateSet are
[0, 0.168] [0, 0.388] [0, 0.418] [0, 0.529] [0, 0.670] [0, 0.814] [0, 0.874] [0, 1.4
3 ItemSet CandiateSet are
[0, 0.168, 0.388] [0, 0.168, 0.418] [0, 0.168, 0.529] [0, 0.168, 0.670] [0, 0.168,
4 ItemSet CandiateSet are
[0, 0.168, 0.388, 0.418] [0, 0.168, 0.388, 0.529] [0, 0.168, 0.388, 0.670] [0, 0.1
5 ItemSet CandiateSet are
[0, 0.168, 0.388, 0.418, 0.529] [0, 0.168, 0.388, 0.418, 0.670] [0, 0.168, 0.388,
6 ItemSet CandiateSet are
[0, 0.168, 0.388, 0.418, 0.529, 0.670] [0, 0.168, 0.388, 0.418, 0.529, 0.814] [0,
7 ItemSet CandiateSet are
[0, 0.168, 0.388, 0.418, 0.529, 0.670, 0.814] [0, 0.168, 0.388, 0.418, 0.529, 0.6
8 ItemSet CandiateSet are
[0, 0.168, 0.388, 0.418, 0.529, 0.670, 0.814, 0.874] [0, 0.168, 0.388, 0.418, 0.5
9 ItemSet CandiateSet are
[0, 0.168, 0.388, 0.418, 0.529, 0.670, 0.814, 0.874, 1.627] [0, 0.168, 0.388, 0.4
10 ItemSet CandiateSet are
[0, 0.168, 0.388, 0.418, 0.529, 0.670, 0.814, 0.874, 1.627, 1.779] [0, 0.168, 0.3
11 ItemSet CandiateSet are
```

Figure 2. Estimation of Frequent Item Set Mining using ARM Approach

Figure 2 illustrates the missing value estimation using Association Rule Mining Algorithm method (ARM) satisfies the given constraints and predicts the missing data efficiently, the insertion of missing values into file is followed by possible error generation for the given algorithm respectively.



```
3.850 2.280 1.055 0.335 0.164 0.410 0.913 0.391 0.338 0.260 4
3.962 1.724 1.290 0.531 0.170 0.528 0.282 0.149 0.603 1.139 5
3.894 2.910 0.968 0.671 0.131 0.519 0.591 0.708 1.515 0.121 6
4.469 4.348 1.740 0.886 0.447 0.883 0.374 0.577 0.908 0.089 7
4.661 4.235 0.497 1.021 0.295 0.457 0.019 0.345 1.355 0.062 8
5.040 5.074 1.721 0.264 0.279 0.176 0.003 0.609 0.177 0.422 9
4.169 1.799 0.621 0.025 0.706 0.362 0.343 0.049 0.339 0.220 10
5.105 1.968 1.140 1.719 0.443 0.432 0.969 0.173 0.551 0.176 0
4.544 1.672 0.643 1.281 0.504 0.803 0.709 0.229 0.546 1.058 1
4.426 1.167 0.008 1.315 0.189 0.297 0.124 0.244 0.423 0.684 2
3.622 0.988 1.108 0.685 0.693 0.754 0.442 0.166 0.458 0.738 3
3.597 2.146 1.492 0.142 0.297 0.815 0.870 0.490 0.471 0.410 4
4.122 1.735 1.340 0.615 0.030 0.353 0.345 0.104 0.540 0.988 5
4.065 2.876 0.856 0.221 0.533 0.232 0.855 0.633 1.452 0.272 6
4.513 4.265 1.477 1.090 0.215 0.829 0.342 0.693 0.601 0.056 7
4.651 4.246 0.823 0.831 0.666 0.546 0.300 0.094 1.343 0.185 8
5.034 4.993 1.633 0.285 0.398 0.181 0.211 0.508 0.283 0.304 9
4.261 1.827 0.482 0.194 0.731 0.354 0.478 0.050 0.112 0.321 10

 Inserting New Values Completed!!!

Process Finished Result Stored In File

Result Of Algorithan Estimation

                Error values    0.2862209153190399
```

Choose Dataset: vowel1b-orig-order

Choose Method: Voting

Figure 3. specifies the estimation of missing values for Voting Approach with Error Calculation

Figure 3 specifies the estimation of missing values in the form of matrix using vowel data for voting approach with Error Calculation. After finding the possible missing values from the dataset it would be inserted at the right place followed by error calculation. The error calculation for voting aapproach seems to be less when compared to the missing value imputation method using multiple support respectively.

The RMSE is also called the Root Mean Square Deviation and RMSD commonly used to measure the difference between original attribute values and estimated attribute value. The RMSE of an estimated attribute value with respect to the original attribute value is defined as the square root of the mean squared error. For different missing rate the RMSE rate for two approaches are calculated. While the missing rate is increased, RMSE rate also increases correspondingly. The proposed system has low RMSE rate compared to the other existing system as shown in the below fig. 4.

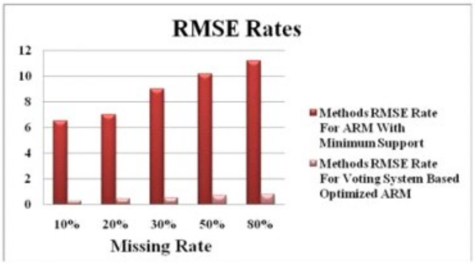


Figure 4. Performance of RMSE Rate of Existing and Proposed Algorithm

Table 1. Comparison of Accuracy Rate for both Existing and Proposed Algorithm

Missing Rate	Methods	
	Accuracy rate for ARM with Minimum support	Accuracy rate for voting system based Optimized ARM
10%	80	86
20%	86	87
30%	87	92
50%	88	93
80%	91	95

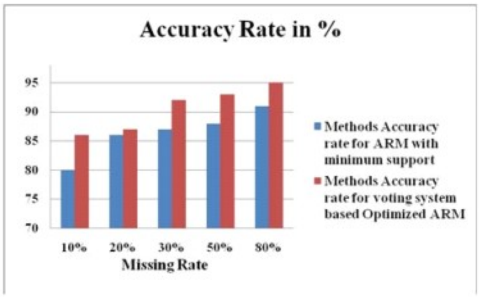


Figure 5. Performance Comparison of Accuracy Rate of Existing and Proposed Algorithm

Table 1 illustrates the accuracy rate of both existing and proposed voting based optimized ARM algorithms. Accuracy rate has been calculated for 10 iterations in terms of accuracy. For 10% missing rate the accuracy rate for missing value imputation using ARM with minimum support algorithm shows 80% and missing value imputation using voting system based Optimized ARM algorithm shows 86% approximately. For 20% missing rate the accuracy rate for missing value imputation using ARM with minimum support algorithm shows 86% and missing value imputation using voting system based Optimized ARM algorithm shows 87% approximately. For 30% missing rate, the accuracy rate for missing value imputation using ARM with minimum support algorithm shows 87% and missing value imputation using voting system based Optimized ARM algorithm shows 92% approximately. For 50% missing rate the accuracy rate for missing value imputation using ARM with minimum support algorithm shows 88%, and missing value imputation using voting system based Optimized ARM algorithm shows 93% approximately. For 80% missing rate the accuracy rate for missing value imputation using ARM with minimum support algorithm shows 91%, and missing value imputation using voting system based Optimized ARM algorithm shows 95% respectively. Thus the proposed voting system based optimized ARM outperforms the existing algorithm as shown in above fig.5.

IV. Conclusions

Handling of imputation involves issues in forms of information loss as consequence or efficiency, secondly data handling leads to computation and analysis due to indiscretion of data structures followed by systematic difference found in data. Since the Associative Rule Mining with single support algorithm results in low accuracy and high error rate. A novel algorithm voting system based Optimized ARM has been presented. The proposed Voting system based Optimized ARM system produces highest accuracy rate and greatly reduces the error rate with high computational complexity. Voting methodology yields a better imputation value and produces a best candidate system with highest vote and eventually reduces the error rate and better accuracy rate compared to the existing algorithm.

References

1. A.Derjani Bayeh, M.J.Smith, Effect of Physical ergonomics on VDT workers health: a longitudinal intervention field study in a service organization, international journal on human computation interaction 11(2):109-135, 2009.

2. H.Kim, G.Golub, H.Park, Missing value estimation for micro array gene expression data:local least square estimation, bioinformatics 21(2):187-198, 2005.

3. J. Osberne, A. Overbay, Best practices in data cleaning, Best Practices Quantitative Methods, 205-213, 2008.

4. J.Han and M.Kamber 2006, Data mining and Concepts and Techniques, Second Edition, Moorgan Kaufmann Publishers.

5. J.L.Schafer, J.W.Graham, Missing Data: Our View of the State of the Art, Psychol, Methods, 7(2), 2002.

6. Jiri Kaiser, Dealing with Missing Values in Data, Journal of Systems Integration,1-10, 2014.

7. K. Rameshkumar, A Novel Algorithm for Association Rule Mining from Data with Incomplete and Missing Values Journal on Soft Computing, 1(4):171- 177, 2011.

8. K.Cheng, N.Law, W.Siu, Iterative Bicluster based Least Square Framework for Estimation of Missing Values in Micro Array Gene Expression Data, Pattern Recognition, 45(4):1281-1289, 2012.

9. M.G.Rahman, M.ZIslam, A decision tree-based missing value imputation technique for data pre-processing, Australian data mining conference, vol.121, ACS, Ballart, Australia, pp-41-50, 2011,

10. T. Schneider, Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and imputation of missing values, Journal of climate, 14(15): 853-871, 2001.

11. Z.Cai, M.Heydari, G. Lin, Iterated Local Least Squares Micro Array Value Imputation Journal of Bioinformatics Computational Biology, 4(5):935-958, 2006.