Treatment of Missing Values for Association Rules: A Recent Survey

Anupama A Chavan
M.Tech (2nd Year)
Department of Computer Science & Engineering
Lord Krishna College of Technology

Vijay Kumar Verma
Asst. Professor M.Tech (CSE)
Department of Computer Science & Engineering
Lord Krishna College of Technology

ABSTRACT

Missing values and incomplete data are usual occurrences in real datasets [1]. The problem of recovering missing values from a dataset has become an important research issue in the field of data mining and machine learning [2]. With the speedy increase in the use of databases, the difficulty of missing values unavoidably arises. The techniques developed to effectively recover these missing values should be highly accurate in order to remove the missing values completely. The association rules are the popular method that is effectively used to establish the relationship among items in databases. The discovered association rules are useful to recover the missing values in databases. There are several methods proposed to surmount the problem of missing value [4]. In this paper we present a study over the existing methods.

Keywords

Missing, Association, Incomplete, Recovering, Relationship

1. INTRODUCTION

Techniques developed for data mining or Knowledge Discovery in Databases (KDD) try to extract hidden and useful information from large databases. Although current technologies can handle massive amounts of data, the rapid growth of databases causes some attribute values to be missed or causes inconsistencies in the data gathering process [5]. Missing data are the absence of data items for a subject who may hide useful and important information and it may be difficult for data analysis. In practice, missing data have been one major factor affecting data quality. The presence of missing data is a general and challenging problem in the data analysis field [3].

In general Missing values comprise an important and unavoidable problem in data mining, data management and analysis. Conventional mining techniques are not capable of working with missingness directly, and require some form of workaround or preprocessing [6].

Before data analysis begins, the data cleaning step deals with errors and inconsistencies from raw data to improve the quality of the discovered information. The problem of missing values play an important role in data mining field, to get to the bottom of it has become a top priority.

One simple method of dealing with missing values is to delete all tuples with missing values. The result is reduced databases, thus small data is only available to analyze and may not give appropriate results. An alternative option to above technique, users can apply some simple statistical methods such as using mean or median to predict missing values. However, the predicted values are still inaccurate, become noise and influence the quality of the information [4].

Consider a simple example of incomplete dataset in table there are twelve tuples and three attributes, A1, A2, and A3. The values of the attributes are shown as follows: A1 = {1, 2}, A2 = {1, 2, 3}, A3 = {3, 4}. Each attribute has 50% missing values [2].

Table 1. Example of an incomplete dataset

	Attribute		
TID	**A1**	**A2**	**A3**
T1	1	2	3
T2	1	3	?
T3	?	?	3
T4	?	2	4
T5	?	3	?
T6	?	1	?
T7	?	2	3
T8	?	?	?
T9	2	?	3
T10	2	?	?
T11	2	?	?
T12	2	?	4

To fill in the missing values for attributes former strategy are: [3, 5]

1. To ignore the tuple

When the class label is missing then this method can be used. If tuple contains several attributes with missing values this method is not successful. It is poor when the percentage of missing values per attribute varies considerably.

2. To fill the missing value manually

In this missing values are filled manually but in a large data set with many missing values this approach is not feasible as it is time consuming.

3. To use a global constant

In this method missing values is replaced by the same constant, such as a label like “Unknown” or - ∞ . If missing values are replaced by “Unknown” then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common. Hence, although this method is simple, it is not recommended.

4. To use attribute mean

In this method missing value is filled by mean or average. For example suppose if average age of graduate student is 20, use this value to fill missing value for attribute age.

5. To use attribute mean for all samples belonging to the same class

This method uses mean or average of attribute for the all samples of the same class.

6. To use the most probable value

This method is most popular and widely used. It uses maximum information available from the present data to predict the missing value. Each missing value is replaced by a probability distribution. This probability distribution represents the likelihood of possible values for the missing data, calculated using frequency counts from the entries that do contain data for the corresponding field.

7. To use the most probable value for all the samples belonging to the same class

This method uses probability of attribute for the all samples of the same class.

2. BACKGROUND

Missing values lead to the complexity of extracting constructive information from that data set. Missing data occurs in many real-world datasets due to it the complexity of analysis increases, and the results can be influenced considerably. Ideally, we would want to make the same inferences about the data, as if we had the complete dataset. A key component here is the mechanism at which this missingness occurs. The missingness mechanism must be considered while analyzing a dataset with missing values. Missingness can be classified into three categories. Table gives a very simple database with no missing values, which will serve as a reference [6].

Table 2. Database without missing values

Age	Gender	Income
30	female	40000
37	male	45000
19	male	50000
26	female	55000

1. MCAR (Missing Completely At Random) – A value is missing completely at random if it is independent of its own value, or the value of any other attribute. For example Missing value (p) neither depends on p nor q. The probability that a value of some attribute is missing is always the same. This nice feature of data that are MCAR **is** that the analysis remains unbiased. Table shows missing values for attribute ’Income’.

Table 3. Database with missing values, MCAR

Age	Gender	Income
30	female	?
37	male	45000
19	male	?
26	female	55000

2. MAR (Missing At Random) - A value is called missing at random, if it is dependent of one or more other attributes i.e. their values in the same tuple. For example Missing value (p) neither depends on q not on p. Table shows missing values in which ‘Income’ depends on ‘Gender’**.**

Table 4. Database with missing values, MAR

Age	Gender	Income
30	female	?
37	male	45000
19	male	50000
26	female	?

3. MNAR (Missing Not At Random) - A value is called missing not at random, if it is dependent of itself. For example the probability of missing value depends on the variable that is missing. This is the most difficult mechanism to deal with, since it cannot be derived from the data. Table shows ‘Income’ is missing.

Table 5. Database with missing values, MNAR

Age	Gender	Income
30	female	40000
37	male	45000
19	male	?
26	female	?

3. RELATED WORK

The problem of deriving associations from data is necessary in data mining. An association rule is an expression X→Y, where X is a set of attributes and Y is usually a single attribute. It means in the set of tuples, if all the attributes in X exist in a transaction, then Y is also in the tuple with a high probability. To achieve this purpose, Agrawal and his co-workers formulated a solution which is referred to as market-basket-analysis. [2, 14]. For example examining the dataset of a supermarket we may get an association rule like:

Milk→ColdDrinks

(support = 0.01%, confidence = 5%)

This means that 0.01% of all transactions contain both milk and cold drinks, and 5% of all transactions which contain milk also contain cold drinks.

To effectively deal with missing values, several researchers have proposed several methods. The methods described below uses association rule to treat the missing values. The basic flow is shown below figure.

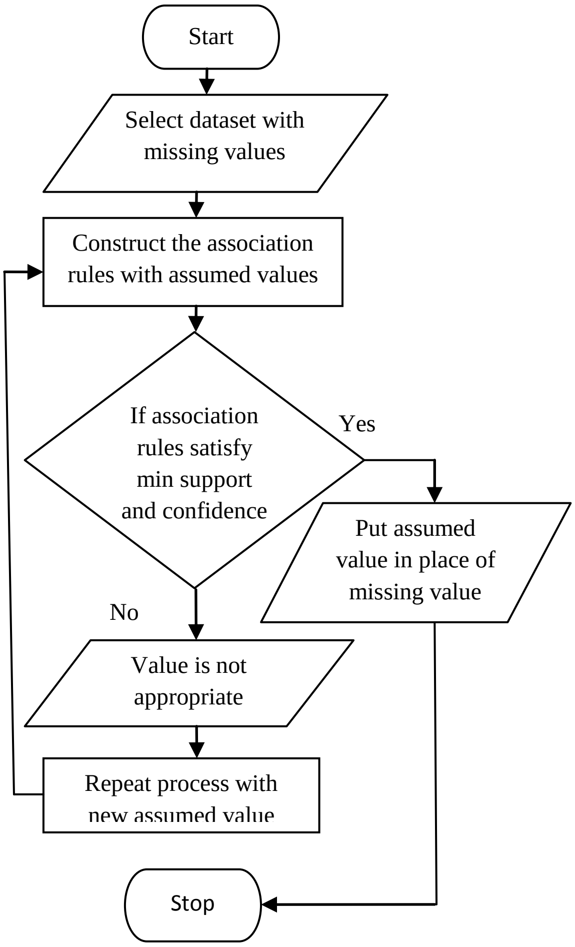
Start
Select dataset with missing values
Construct the association rules with assumed values
If association rules satisfy min support and confidence
Yes
Put assumed value in place of missing value
No
Value is not appropriate
Repeat process with new assumed value
Stop

Fig 1: Flow chart to show Filling of Missing Values

3.1 Robust Association Rules (RAR)

RAR approach to mine association rules in an incomplete dataset was proposed by Ragel and Cremillex. RAR partially disabled the tuples with missing attribute value instead of deleting it to ease the issue of lost rules. Then association rules discovered by redefining the support and confidence are mined with RAR method to recover the missing values in a dataset. Deleting tuples with missing values often generates too few useful rules to be applied effectively [2, 10]. The RAR approach cuts a database into several valid databases (vdb), which contain no missing values, in order to discover rules. A valid database (vdb) is defined for an item set X as the largest sub database without missing values for this itemset [9, 10].

RAR is the simplest and the oldest method to recover missing but the it does not recover multiple missing values.

3.2 Missing Value Completion (MVC)

RAR only partially disables the victim tuples to passively discover the association rules. To overcome, Ragel and Cremilleux have also proposed the Missing Values Completion (MVC) approach, which is based on the RAR method, to recover multiple missing values in a database. First, MVC applies RAR to discover all association rules. Then, MVC applies the most appropriate rule to fill in a single missing value in a tuple. If a tuple has multiple missing values, MVC runs the process repeatedly. To keep away the propagation of the wrong value, MVC uses the rules, which have a high confidence value (more than 95%), to recover a tuple with multiple missing values [4, 9].

Thus this a simple method to recover multiple missing values in a database.

3.3 Combined Association Rules Using Recycling Technique

Give two pre-defined thresholds minSup and minRSup. If the support value of an itemset, *X*, is less than minSup and not less than minRSup, the itemset is called a sub-frequent itemset. That is minRSup ≤ Sup(X) < minSup. The discovered association rules can be used to recover corresponding missing values. Therefore, increasing the discovered rules can increase the recovery rate for missing values in a relational database. To increase the identified rules, this study combines at least two sub-frequent itemsets to become a frequent combined itemset. The support value of each combinable sub-frequent itemset requires archiving the pre-defined recycle support threshold value, minRSup, where minRSup < minSup. The method of recycling some sub-frequent itemsets is called the Recycle Combined Association Rules (RCAR) method [4].

Thus combining method increases the discovered rules which ultimately increases the recovery rate for missing values

3.4 An array based algorithm for RCAR

To speed up the mining process of RCAR this study establishes the bit-arrays of a relational database table using simple Boolean AND/OR operations, called the Fast Recycle Combined Association Rules (FRCAR), on the arrays. Given a relational database table, performing the array-based algorithm involves three matrices, the Itemset Matrix (IM), the Missing Value Matrix (MVM), and the Recycling Itemset Matrix (RIM). IM and MVM are used to calculate the representative and support value of each corresponding itemset. RIM stores the sub frequent itemsets, which are prepared to be recycled. [4]

Thus FRCAR speed up's the process of RCAR. Since this method not only generates association rules but combined association rules so its performance and accuracy is better and faster than pervious methods.

3.5 Iterative Missing Value Completion Method

The iterative missing-value completion method with the RAR support to extract the association rules from an incomplete dataset with a high missing rate. It consists of three phases. The first phase uses the association rules which are mined from the original incomplete dataset to roughly complete missing values. The second phase uses the reduced minimum supports to gather more association rules from the originally incomplete dataset to complete the rest of missing values in an iterative way until no missing values exist. The third phase uses the association rules from the completed dataset to correct the missing values that has been filled into predicted values until convergence [2].

As seen in above methods the rate of missing values was less but for high missing rate iterative method is the best option. Thus results will have higher accuracy and increased recovery rate of missing values

3.6 ARDM Algorithm

ARDM (Association Rule mining from Data with Missing Values) is a novel technique to mine association rules from data with missing values. The proposed algorithm adopts Apriori approach, and uses partitioned databases. It consists of two phases: database scrutinizing phase and construction of association rule phase. The association rule generation phase consists of the following two processes: frequent itemset generation and Association rule construction. Database scrutinizing phase generates database partition and deletes unnecessary attributes which means attributes are not satisfied for user specified minimum representativity threshold from database. The association rules are mined by using database partition which is created by the database scrutinizing phase. This phase adopts Apriori technique using transaction reduction based approach. This procedure is divided into two processes as: Generation of frequent itemset and Construction of association rules [1].

The most improved part in this method is that it avoids unwanted database partition process. The method gives good performance even if the missing rate is high or low.

4. COMPARITIVE STUDY

Following table shows the comparative study between some of the methods shown above

Table 6. Comparison table between various methods

Name of Method	Used Concept	Type of Missing Value handled
RAR	Partial disables the victim tuple	Handles only one missing value in a tuple
MVC	Recursive RAR to predict multiple missing values	Handles multiple missing values in a tuple
RACR	Combines at least two sub-frequent itemsets to become a frequent combined itemset	Handles many missing values at a time compared to previous
FRCAR	Establishes the bit-arrays of a relational database table using simple Boolean AND/OR operations	Handles maximum missing values at high speed
Iterative MVC	Three phases used Roughly assigning values, filling in remaining missing values & adjusting the assigned missing values	High Missing Rates
ARDM	Two phases used Database Securitizing & Construction of association rules	High & Low Missing Rates

5. CONCLUSION

The recovery of missing values is an important issue in data preprocessing. In the methods they reduce the minimum support and there are no parameter for reducing it, since it is not fixed such rules discovered are useless. There are no algorithms for temporal dataset. In general, there is no best, universal method of handling missing attribute values. On the basis of study of research on comparison such methods we may conclude that for every specific data set the best method of handling missing attribute values should be chosen individually, using as the criterion of optimality of multi-fold cross validation experiments.

6. REFERENCES

[1] K. Rameshkumar, "A Novel algorithm for association rule mining from data with incomplete and missing values," ICTACT Journal on Soft Computing, Vol. 01, No, 4, 2011.

[2] Tzung Pei Hong, Chih Wei Wu, "Mining rules from an incomplete dataset with a high missing rate," Expert Systems with Application An International Journal, Vol. , No, , 2011.

[3] Dinesh Prajapate,Jagruti Prajapate, "Handling Missing Values: Application to University Data Set," International Journal of emerging trends in engineering and development, Vol.01 , 2011.

[4] J. J. Shen, C. C. Chang and Y. C. Li, "Combined association rules for dealing with missing values," Journal of Information Science, Vol. 33, No, 4, 2007.

[5] Jiawei Han, Micheline Kamber, Data Mining Concepts & Techniques, Morgan Kaufmann Publishers

[6] Michael Mampaey, "Association Rule Mining met Missing Values," UNIVERSITENT ANTWERPWN, Department Wiskunde en Informatica, Acdamiejaar, 2005-2006.

[7] **]** J. R. Nayak and D. J. Cook, "Approximate association rule mining," The Submitted manuscript 26 / 27 Fourteenth International Florida Artificial Intelligence Research Society Conference, 2001,

[8] M. Kryszkiewicz, "Probabilistic Approach to Association Rules in Incomplete Databases," Lecture Notes in Computer Science, Vol. 1846, 2000

[9] A. Ragel and B. Cremilleux, "MVC – a preprocessing method to deal with missing values," Knowledge-Based Systems, Vol. 12, No. 5, 1999

[10] A. Ragel and B. Cremilleux, "Treatment of missing values for association rules," Lecture Notes in Computer Science, Vol. 1394, 1998, pp. 258-270

[11] R. Srikant, Q. Vu and R. Agrawal,, "Mining association rules with item constraints," The Third International Conference on Knowledge Discovery and Data Mining, 1997

[12] R. Agrawal and R. Srikant, "Fast algorithm for mining association rules," The International Conference on Very Large Data Bases, 1994

[13] R. Agrawal, T. Imielinksi and A. Swami, "Database mining: a performance perspective," IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, 1993

[14] R. Agrawal, T. Imielinksi and A. Swami, "Mining association rules between sets of items in large database, " The 1993 ACM SIGMOD Conference on Management of Data, 1993