# Missing Value Imputation using Hybrid K-Means and Association Rules

Geeta Chhabra, Research Scholar
Amity School of Information Technology
Amity University,
Noida, Uttar Pradesh, India
geeta_chhabra@rediffmail.com

Vasudha Vashisht, Assistant Professor
Department of Computer Science & Engineering
Amity School of Engineering
Noida, Uttar Pradesh, India
vvashisht@amity.edu

Jayanthi Ranjan, Professor
Institute of Management Technology
Ghaziabad, Uttar Pradesh, India
jranjan@imt.edu

***Abstract*—Association rules is an important and well researched database mining function for discovering interesting relationship between variables in large database. The data mining architecture works on facts and figures which are used for any type of decision making. To perform any analysis and decision making, these facts must be complete so that the analyst can make a strategy for decision making. In fact the most important problem in knowledge discovery is the missing values of the attributes of the dataset. If such imperfections are there in database, it is cleaned during pre-processing and is prepared in order to be functional. Considering, the importance of handling missing instances in data mining and knowledge discovery, we proposed a hybrid algorithm for using a combination of association rules mining and k-nearest neighbour approach. We performed a detail experimental result on UCI datasets to check the effectiveness of our technique. For this, we have discretize the data using K-means technique and to generate rules in large volume of data using apriori association rule mining. The measurement metrics such as confidence, support, lift & coverage are then used to evaluate the discovered rules.**

***Keywords*—** Data Mining, K-means Clustering, Apriori Algorithm, Association rules, Missing Data.

## I. INTRODUCTION

The problem with missing values in the data occurs at the time of preparation of data for analyses which needs to be addressed. Missing data, that is, fields for which data is unavailable or incomplete, is particularly important problem, since it can lead the analysts to draw inaccurate conclusions. The easiest possible solution for this is reducing the data set. This is commonly used in practice but may cause significant loss of usable data. The other possible solution is missing values imputation. Missing values imputation must be done carefully to avoid biasness in dataset. Commonly used methods, namely mean substitution, neural networks, nearest neighbour and linear regression etc can be used for missing values imputation. The major drawback of these methods is that it does not consider attributes dependencies. This paper describes hybrid technique using K-means technique and apriori rule mining algorithm for association. One attribute is discretize using K-mean technique and rules are then generated using apriori rule mining algorithm for association and these rules are then evaluated with metrics such as confidence, support, lift & coverage. Thus, association rule mining can effectively establish the relationship among attributes in databases and are usually applied to recover the missing data in databases.

## II. IMPORTANCE OF DATA MINING WITH MISSING VALUES

Many contemporary data from industrial and research are incomplete due to several causes such as faulty equipment, incorrect measurements & incorrect entry of data. Thus, in most of the information used, it is common to find missing values. Looking at Null values in data, it is easy to reveal the missing values. However, this does not work in certain situation as the incomplete data can be incorrect or can appear as outliers.

To perform the data analysis with missing values is difficult. The main problems associated with this are;

1. decreased efficiency;
2. problem with analyzing and handling data;
3. biasness results due the difference between missing and complete data.

Missing values can be dealt in many ways. The simplest is to discard the sample with missing data. This method is used when number of samples with missing values in the data is small and the analysis after discarding samples with missing value does not cause any serious biasness during inference. One can also replace the missing values with new value but this type of imputation may cause serious inference problem. It is desirable to perform missing value imputation, if the number of samples that have incomplete data values is significant for relatively less number of variables.

ISBN: 978-1-5386-4119-4/18/$31.00 ©2018 IEEE

The main benefit of imputation method it does not depend on the learning algorithm which makes it easy to select the most suitable method for each situation. A considerable number of these methods are available from simple mean imputation to the complex one which defines the relationships among variables.

Categorization of various mechanisms that leads to presence of missing values in data set in order to determine the appropriate imputation method in general are [23,36];

1. Missing at random (MAR). In this, sample having incomplete data does not depend on missing value but it is dependent on observed value.
2. Missing completely at random (MCAR). In this mechanism, missing data does not depend either on the missing data or on observed data.
3. Not missing at random (NMAR). In this mechanism, probability distribution of incomplete data for an attribute is dependent only on the incomplete data.

The two types of mechanisms viz MCAR & MAR are assumed to be treated by the imputation methods. It is assumed, in case of MCAR that the distribution of complete & missing data is similar while it is different for MAR. The incomplete values can be predicted by using the complete values.

## III. IMPUTATION METHODS

Methods from related fields are used as imputation methods which are briefly described as:

- Do Not Impute (DNI). In this method, missing value remains unchanged so that one must use their default missing value strategies. If one has enough data, a good strategy is to just remove the rows with missing values and work with the subsample of data which is complete. It should be ensured that data mining perform better than original data when imputation is applied. However, after the imputation process, no machine learning method is applied.
- Ignore Missing or Case deletion. All attributes with not less than one missing value is deleted from the data & analysis is performed on rest of complete data set in this method.
- Global Average Value for Numerical variables and Global Most Common Variable Value for nominal variables. It is one of the simplest procedure. Here, missing value for numerical variables are filled in with the average of the corresponding variables and for the nominal variables is filled in using the most common variable.
- Concept Average Value for Numerical Variables and Concept Most Common Value for nominal Variables. Here, substitute the missing value by mean in case of numerical variables and the most frequent one in case of nominal variable by considering the variables of the same class as the related variable.
- K-Nearest Neighbour Imputation. This is an instance based algorithm, whenever a missing value in an instance is found; impute a value from the k nearest neighbours. Among all most common neighbours value is used in case of nominal values and average is used in case of numerical values. Using euclidean distance one needs to define a proximity measure between instances.
- Imputation with Weighted K-Nearest Neighbour. In this method each neighbour is equipped with weights. This method imputes K-Nearest Neighbour with similar values to be one. Using a most repeated value or weighted mean, according to the distance, from neighbours, the estimated value takes into account the different distances [14].
- Imputation with K-means Clustering. Using similarity of objects, the aim is to minimize the dissimilarity between cluster by dividing data into groups. It is the distance among the cluster centroid to which it is assigned and the objects. The cluster centroid is the average of the objects in that cluster. Calculate k revised clusters centroids based on the results from the previous step [29]. Built up the binding between the revised nearest centroid and the original data points, after getting these revised k centroids. Thus, a loop is generated. As a result of this, step by step, k centroids change their location till no more changes are done which means that the centroids do not move any more. Thus nearest neighbour clustering algorithm will impute missing values in same way like k-Nearest Neighbour Imputation [30].
- Imputation with Fuzzy K-means Clustering. In this method, degree with which data object is assigned to certain cluster is described by the member function. In this, process of updating member function, only complete attributes are taken into account. The data object of a particular cluster cannot be assigned to kth clusters with different membership degrees. The kth cluster is represented by a cluster centroid. Based on the membership degree and the value of centroids cluster, one can impute the incomplete data object.
- Support Vector Machines Imputation. It is supervised machine learning model to fill in missing values based on regression method. In this method, using the missing condition variable values, the output or classes can be predicted. First, select the samples which are complete. In the next step, select one input variable as the decision variable and one of the decision variables as the input variable to use Support Vector Machines regression to predict the output variables [16].
- Event Covering. It is a mixed-mode discrete probability model. In this method, discretize one of the continuous variables based on minimum information loss criteria. A mixed-mode attribute is treated as discrete one and apply cluster analysis on it. It does not require ordering discrete values or scale normalization. The conversion of data into statistical knowledge has three steps: 1) detect & analyze data patterns which are statistically dependent 2) based on the detected interdependency group the data into clusters and 3) for

each clusters interpret the underlying patterns. With this developed inference method, missing values can be estimated.

- Regularized Expectation-Maximization. In this method, in each of the three steps, each iteration gives the revised estimates of the mean and covariance matrix. In first step, from the estimates of mean and covariance matrix, the regression parameters of the variables with missing values on the variables with available values are computed. The missing values are then imputed with the estimates of the mean and the covariance matrix & the conditional expectation values given the available values. The conditional expectation is the product of the estimated regression coefficients & the available values. Lastly, the mean & covariance matrix are revised using sample mean & sample covariance matrix of the complete data set and estimate of the conditional covariance matrix of the imputation error. It is started with initial estimates. The iterations are made till the imputed values & the estimates of the mean & covariance matrix stop changing from one iteration to the next iteration.
- Imputation with Singular Value Decomposition (SVD). In this, obtain a set of patterns which are mutually orthogonal expression such that they can linearly be combined to approximate the values of attributes. For this, first estimate, the missing value with the expectation maximization algorithm, and to obtain the eigen values, compute the singular value decomposition. To estimate missing value, this eigen values is applied to regression over the complete attributes of the instances [28].
- Bayesian Principal Component Analysis. This method is based on bayesian principal component analysis for estimating missing values. Using the framework of bayes inference, this method is based on latent variables and probabilistic model which are estimated together. Thus, in terms of statistical methodology, it makes possible to estimate arbitrary missing values. For estimating, incomplete values, it follows three basic steps, namely; (1) regression based on principal component, (2) estimation using bayesian, and (3) an iterative algorithm such as expectation maximization [25].
- Imputation with Local Least Squares. The value to be predicted is a linear combination of similar variables in this method. Based on a similarity measure similar genes are only used instead of using all available genes in this method. This method has two steps. They are (1) use L2-norm to select k genes (2) estimation and regression irrespective of selected k genes.
- Imputation with Neural Networks. It is a predictive modelling that works on iterative parameter adjustment. It includes the neural framework which has number of neurons, number of layers, type of neuron model, etc and structure of interconnection is known as topology or architecture. Only an input and output layer is there in single-layer network. One or more hidden layers are inserted between the input and output layer, in case of multilayer network.
- Multiple Imputations. It is a technique based on statistics for analyzing missing data sets due to various reasons such as non response surveys etc. It has three stages: imputation, analysis and pooling. In imputation, the missing entries are filled in, not once, but m times. Then, theses m completed data set are analysed and pooled by integrating the m analyzed results into a final result.

## IV. Literature Review

Chaudhary V., Choudhary A, Nilambari N. & Tyagi M. (2014) in their article "Review on Association Rule Mining: A Survey" has described different types of approaches on association rule mining and have designed algorithms. They have also described appropriate approach for association rule. They presented in different domains, a complete survey with various algorithms and approaches for association rule mining.

Kaiser Jiří (2011) in their article, "Algorithm for Missing Values Imputation in Categorical Data with Use of Association Rules" has presented new algorithm for categorical data for missing values imputation. The algorithm is based on three variants using association rules. Experiments shows better accuracy in new algorithm than using most common attribute value for missing values imputation.

Jing T., Bing Y., Dan Y. and Shilong M. (2013) in their article "Clustering-Based Multiple Imputation via Gray Relational Analysis for Missing Data and Its Application to Aerospace Field" proposed a hybrid missing data imputation using entropy based and Gray System theory on University of California(UCI) dataset. In the first step, it groups the non-missing data attributes into clusters and computes the imputed value by utilizing the information entropy for each missing attribute of the proximal category based on Gray System theory in terms of the similarity metric.

Sabthami J., Thirumoorthy K. and Muneeswaran K. (2016) in their article "Mining Association Rules for Early Diagnosis of Diseases from Electronic Health Records" classified patients based on the presence of disease using K-Nearest Neighbour (KNN) and Naïve Bayesian algorithm. The clinical documents of 1235 patients are collected from the site www.i2b2.org. The association rule mining is used to find information on frequently occurring diseases.

Jianhua W., Qinbao S. and Junyi S.(2008) in their paper "Missing nominal data imputation using association rule based on weighted voting method" discovered association rules to predict the missing data. They presented a novel approach to impute missing data using association rules based on weighted voting. To evaluate the performance of proposed method, they experimented on UCI Machine Learning datasets repository. From the experimental results it is indicated that the accuracy of classification is increased when the proposed method is utilized for missing values imputation.

Nuntawut K., Phatcharawan C., Kittisak K., Nittaya K. (2013) in their article "Discretization and Imputation Techniques for Quantitative Data Mining" based on the Chi2 algorithm they proposed the discretization technique to categorize numeric values. The discovered association rules are then evaluated using measurement metrics such as support, confidence, lift and coverage. The tree-based data classification method has been used to impute dataset with missing values to assess predictive accuracy. UCI Machine Learning Repository Hepatitis dataset have been used in this paper.

Shah J. M. & Shahu L. (2015) in their article "Recommendation based on Clustering and Association Rules" has made a recommender systems based on most adopted techniques such as hybrid filtering, collaborative filtering, and content-based filtering. The paper mainly describe about the issues of recommender system. The basic aim is to recommend the suitable items to users for better rule extraction using association mining. The clustering method is also applied to cluster the data based on similar characteristics. To overcome the certain problem such as sparsity, cold-start problem association mining over clustering is used.

Zeng Y., Shiqun Y., Jiangyue L. and Miao Z. (2015) in their article "Research of Improved FP-Growth Algorithm in Association Rules Mining" has used frequent-pattern growth algorithm. It is an improved version of the apriori algorithm. It compresses data sets to a frequent-pattern tree and the data set is scanned two times. It does not produce the candidate item sets in this process and greatly improves the efficiency of data mining. But frequent-pattern growth algorithm needs to create a frequent-pattern tree which contains all the data sets. This frequent-pattern tree requires high memory space. Scanning of the database two times improves the efficiency of frequent-pattern growth algorithm.

In this paper, they have worked on two kinds of improved algorithms namely N Painting-Growth algorithm which builds two-item permutation sets to find association sets of frequent item sets and Painting-Growth algorithm which digs up frequent item sets as per the association sets. Painting-Growth algorithm builds an association picture based on the two-item permutation sets to find association sets of all frequent items. It then digs up all the frequent item sets according to the association sets. Both of these algorithms developed by them scans the database only once instead of scanning database twice in traditional frequent-pattern growth algorithm. It completes the data mining only as per the two-item permutation and thus it is faster, takes small memory space, less complex and is easy to handle.

## V. ALGORITHM

### A. *K-means*

K-mean clustering is an unsupervised machine learning algorithm which is used to cluster related data into groups without any prior knowledge of those classes. For these following steps are performed [4]:

Let the data set points are X = {x1, x2, ……..,xn} and centres points are V = {v1, v2,…….,vc}.

1. Randomly select cluster centres.
2. Measure the euclidean distance between cluster centres and each data point.
3. Data item is placed in the cluster having minimum distance from centre.
4. Using vi =1/ci∑xi, calculate the revised cluster centre, where 'ci' is data points in ith cluster.
5. Measure again, distance between every data point & revised cluster centres.
6. Stop the process, if there is no data point to be, else go to step 3.

### B. *Association Rule Mining*

It is one of the popular methods for discovering relationship among attributes of large databases [5, 24]. It is defined as the association among attributes such as A⟶ C [6] where A is an antecedent means if statement and C is consequent means then statement with evaluating measures such as confidence and support. It is necessary to define APRIORI algorithm, before discussing association rule. To generate rules and mine all frequent item sets in database this algorithm is used. Some of the important terms which are helpful to understand the concept of APRIORI algorithm are described as:

I. *Item sets*

In a single transaction, group of items is item set. There are 2n item sets, if the dataset contains n items.

II. *Support*

The measure to know how frequently an item set appears in the dataset is support. It is number of transactions having both antecedent and consequent. For example, support of A⟶ C is the number of transaction items that contains both A and C [10].

III. *Confidence*

This is ratio of number of transaction set containing both consequent and antecedent to the number of transaction containing antecedent. For example, ratio of support of A⟶ C which is the number of transaction that contain both A & C and number of transaction which has A [11, 25].

IV. *Frequent item sets*

If an item set has support equal to or greater than minimum user-specified limit, then the item set is frequent else it is known as infrequent [12,31].

The main requirement is that it should satisfy a minimum user-specified support and confidence for association rule [27]. It is divided into two steps;

1. A minimum threshold support is applied to select an item set as frequent item set in a database.

2. A minimum threshold confidence is applied to these item sets which are frequent in order to have valid rules.

It involves searching all possible item combinations in database to have item sets which are frequent means the subsets of a frequent item sets are also frequent [25]. Thus, a frequent item set cannot have an infrequent subset. The apriori algorithm can be used to find all frequent item sets by using this property of an efficient algorithm [ 2, 13,34].

C. *Apriori Algorithm*

Following steps are performed for this algorithm [1,18]:

1. A set of k + 1 item are generated using k-item sets in the previous pass to find frequent item sets.
2. For an item set to be frequent each k-item set must be equal to or greater than minimum threshold support defined by user otherwise it is candidate item set. Using candidate generation, it finds frequent item set.
3. Here, k + 1 item sets of level 2 are searched using k item sets of level 1 i.e. L1 is used to search level L2, L2 is used to search level L3 and so on. Thus, it is an iterative approach known as level wise search.
4. From the frequent item sets, valid rules are generated.

Multiple passes over dataset are made. In first pass, it estimates the support and confidence for each level of rules. Thus, determines interesting rules decided by minimum support and confidence requirement. Start with a set of interesting rules and with every subsequent pass over the data available in the previous step. This is then utilized to produce advanced rules called candidate rules which are based on the actual support and confidence of these candidates during the scan of the data. At end, the rules are kept that are deemed interesting for use in the next iteration. When in the latest iteration, no new valid rules were found, the process ends.

## VI. PROPOSED SYSTEM

One of the attribute i.e. petal.width has been discretize using K-means and apriori algorithm for association rule mining is applied to generate rules in large volumes of data. Apriori algorithm is best for small datasets where the minimum number of candidate have been generated. A threshold or limit can be set. The measurement metrics such as confidence, support, lift & coverage are then used to evaluate the discovered rules [3,7, 20, 32, 33].

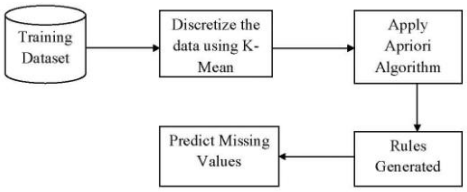


Fig. 1. Proposed System

## VII. EXPERIMENTAL ANALYSIS

For implementing hybrid algorithm for missing values imputation using combination of k-nearest neighbour and association rules mining approach [19], we've used the R Environment with various user defined and system defined libraries. We have used iris dataset consisting of 150 records with five attribute for implementing the algorithms which follows MAR data mechanism. The dataset iris [21] have three classes each of 50 instances of the plant type. All instances like petal width, sepal width, petal length andsepal length, are in cms, three classes are Virginica, Versicolour and Setosa. We have used arules[8] and arulesViz [9] packages. This function discretize has been used to convert one of continuous variables into a factor variable which is required for association rule mining using k mean clustering for three types of flowers.

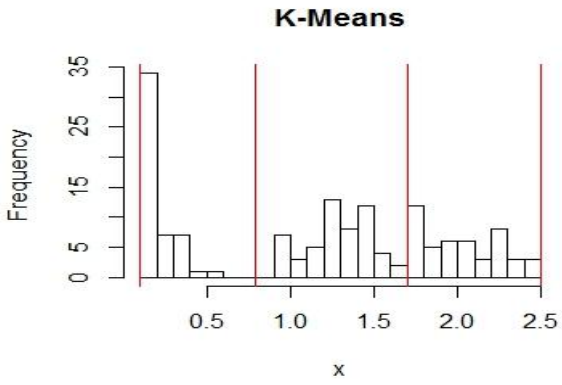


Fig 2: Discretization of Petal Width using K-Means

## VIII. OBSERVATIONS

- After discretize the data, we have 50 rows in petal width in the interval [0.100,0.792), 54 in the interval [0.792,1.705) and 46 in the interval [1.705,2.500].

- The transaction item Matrix will have 150 rows with 15 colunms each having a density of 0.3333 means 33.33% are non zero matrix cell.
- The matrix has 150 times 15 i.e. 2250 cells. Hence, 150 times 15 times 0.3333, so 750 items are there.
- Sepal Width = [2.75,3.29) appeared 74 times out of 150 rows, means 0.49 percent of rows.
- Average transaction contained 750/150 = 5 items.
- A total of 150 transactions contained 5 items,
- The first quartile and median size are 5 and 5 items respectively, means that 25 percent of transactions have 5 or fewer items and about half contained around 5 items.
- The mean of 5 matches the value we calculated manually.

TABLE I. ITEM FREQUENCY OF TOP 5 ITEMS

| | |
|---|---|
| Sepal.Length = [4.30,5.42) | 0.3466667 |
| Sepal.Length = [5.42,6.39) | 0.3733333 |
| Sepal.Length = [6.39,7.90] | 0.2800000 |
| Sepal.Width = [2.00,2.75) | 0.2200000 |
| Sepal.Width = [2.75,3.29) | 0.4933333 |

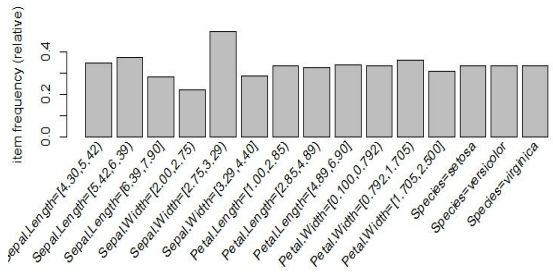


Fig 2: Visualize item support - item frequency plots

## IX. IMPLEMENTATION OF APRIORI ALGORITHM

We will implement apriori algorithm to find the associations among the attributes[11,20].

### A. *Support*

- To set a support by thinking the minimum number of transactions we would need.
- For an example, if an item appeared 15 times then it may be worth taking a look at.
- Then support will be 15 out of 150 transactions, i.e. 0.1

### B. *Confidence*

- We have set confidence threshold as 0.80 that the rule is correct at least 80 percent of time in order to be included in the results.
- This will eliminate the most unreliable rules while allowing some space to modify behaviour with targeted promotions.

In addition, to eliminate rules that contain fewer than two items we have also set minlen to 2.

With the support is 0.1, an item must have appeared at least 0.1*150=15 times. Only15 items appeared those many times, so 168 rules were generated.

A common approach is to take the result of learning association rules and divide them into three categories;

I. Actionable - The goal is to find actionable associations or rules that are clear and useful. Some rules are clear; others are useful and is difficult to find a combination of both of these.
II. Trivial - Some rules are so obvious which need not to mention, they are clear, but not useful.
III. Inexplicable - If the association between the attribute is so unclear that figuring out information for action will require additional research.

TABLE II. TOP TEN RULES

| Rules | Support | Confidence | Lift | Count |
|---|---|---|---|---|
| {Sepal.Length=[6.39,7.90], Species=virginica} => {Petal.Width=[1.705,2.500]} | 0.2000 | 0.9677 | 3.1557 | 30 |
| {Sepal.Length=[6.39,7.90], Petal.Length=[4.89,6.90], Species=virginica} => {Petal.Width=[1.705,2.500]} | 0.2000 | 0.9677 | 3.1557 | 30 |
| {Sepal.Length=[6.39,7.90], Sepal.Width=[2.75,3.29), Species=virginica} => {Petal.Width=[1.705,2.500]} | 0.1467 | 0.9565 | 3.1191 | 22 |
| {Sepal.Length=[6.39,7.90], Sepal.Width=[2.75,3.29), Petal.Length=[4.89,6.90], Species=virginica}=> {Petal.Width=[1.705,2.500]} | 0.1467 | 0.9565 | 3.1191 | 22 |
| {Sepal.Length=[5.42,6.39), Sepal.Width=[2.75,3.29), Species=versicolor} => | 0.1000 | 1.0000 | 3.0612 | 15 |

| | | | | |
|---|---|---|---|---|
| {Petal.Length=[2.85,4.89)} | | | | |
| {Sepal.Width=[2.75,3.29), Species=virginica}=> {Petal.Width=[1.705,2.500]} | 0.1934 | 0.9355 | 3.0505 | 29 |
| {Sepal.Width=[2.75,3.29), Petal.Length=[4.89,6.90], Species=virginica}=> {Petal.Width=[1.705,2.500]} | 0.1800 | 0.9310 | 3.0360 | 27 |
| {Petal.Length=[1.00,2.85)} => {Petal.Width=[0.100,0.792)} | 0.3334 | 1.0000 | 3.0000 | 50 |
| {Petal.Width=[0.100,0.792)}=> {Petal.Length=[1.00,2.85)} | 0.3334 | 1.0000 | 3.0000 | 50 |
| {Petal.Length=[1.00,2.85)}=> {Species=setosa} | 0.333333 | 1.0000000 | 3.0000 | 50 |

The first rule, with a lift of 3.155680, implies that if Sepal.Length=[6.39,7.90] and Species=virginica, then it is nearly three times more likely that Petal Width will fall between 1.705 and 2.5.

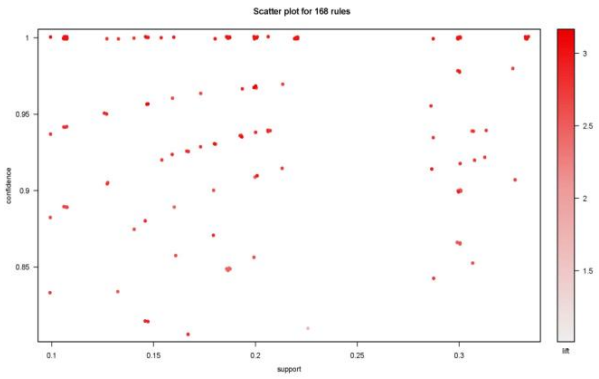


Fig 3: Association Rule-Scatter Plot

## I. CONCLUSION & FUTURE WORK

Data mining methodology has an extraordinary contribution to extract the hidden information and knowledge in large dataset. This paper identifies significant attribute association rules using apriori which is used to estimate the incomplete data in data set. Apriori algorithm is helpful in finding best association rules.

The tests results confirmed expectation that the missing values imputation algorithm using association rules has improved accuracy with increased dependency between variables. In data set with independent attributes, missing values imputation accuracy is similar to accuracy of the most common attribute value method. As a future work, it is required to search for the way of setting minimum required support of association rules to get the best accuracy. Another goal should be an optimization of association rules generation for missing values imputation purpose. A Fuzzy association rule or hierarchical clustering with association rule can also be used for finding missing values for better accuracy.

## REFERENCES

[1] Agrawal Pooja, Kashyap Suresh, Pandey Vikas Chandra, Keshri Suraj Prasad, "A Review Approach on various form of Apriori with Association Rule Mining,International Journal on Recent and Innovation Trends in Computing and Communication", Volume 1 Issue 5, May 2013.

[2] Alsalama Ahmed Mohammed K., "A Hybrid Recommendation System Based On Association Rules, International Journal of Computer, Electrical, Automation, Control and Information Engineering", Vol:9, No:1, 2015.

[3] Baiwal Sweety, Raghuvanshi Abhishek, "Imputation of Missing Values using Association Rule Mining & K-Mean Clustering", IJSDR, Volume 1, Issue 8, August 2016,.

[4] Charliepaul C.Kumar, Gnanadurai G.Immanual, "Comparison of K-Mean Algorithm & Apriori Algorithm-An Analysis",International Journal On Engineering Technology and Sciences, Volume I, Issue III, July 2014.

[5] Chaudhary Vintee, Choudhary Aparna, Nilambari Nawagata, Tyagi Manika, "Review on Association Rule Mining: A Survey", International Journal of Engineering Research & Technology (IJERT), Vol. 3, Issue 4, April 2014.

[6] Dixit Jyotsana, Abha Choubey, "Optimization of Association Rules using Hybrid BPSO", International Journal of Computer Techniques, Volume 2, Issue 3, May 2015.

[7] Gandhi Monali, "An Enhanced Approach towards Tourism Recommendation System with Hybrid Filtering and Association", SRIMCA, VOL. 8, NO. 1, June 2015,.

[8] Hahsler Michael, Package 'arules' September 1, 2017, http://mhahsler.github.io/arules/, CRAN Repository.

[9] Hahsler Michael and Chelluboina Sudheer, "Visualizing Association Rules: Introduction to the R-extension Package arulesViz", Southern Methodist University, Sudheer Chelluboina, Southern Methodist University, https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf.

[10] Hsu P.L, Lai R, Chiu C.C, Hsu C.I, "The hybrid of association rule algorithms and genetic algorithms for tree induction: An example of predicting the student course performanc", Expert Systems with Applications, Volume 25, Issue 1, July 2003.

[11] Isakki P. alias Devi, Rajagopalan S.P., "Analysis of Customer Behavior using Clustering and Association Rules", International Journal of Computer Applications, Volume 43, No.23, April 2012.

[12] Jau-Ji Shen, Chin-Chen ChangandYu-Chiang Li, "Combined association rules for dealing with missing values", Journal of Information Science, 2007.

[13] Jeetha B. Rosiline, "Approches on Future Request Prediction in Web Usage Mining using Data Mining Techniques", ARPN Journal of Engineering and Applied Sciences, VOL. 9, NO. 10. October 2014

[14] Jianhua Wu, Qinbao Song and Junyi Shen,"Missing nominal data imputation using association rule based on weighted voting method", IEEE International Joint Conference on Neural Networks, 2008.

[15] Jing Tian, Bing Yu, Dan Yu, and Shilong Ma, "Clustering-Based Multiple Imputation via Gray Relational Analysis for Missing Data and Its Application to Aerospace Field", The Scientific World Journal, 2013.

[16] Jinubala, V, Lawrance, R, "Analysis of Missing Data and Imputation on Agriculture Data With Predictive Mean Matching Method", International Journal of Science and Applied Information Technology (IJSAIT), Vol.5 , No.1, 2016.

[17] Kaiser Jiří, "Algorithm for Missing Values Imputation in Categorical Data with Use of Association Rules", Int. J. on Recent Trends in Engineering and Technology, Vol 6, No. 1, Nov. 2011.

[18] Kaur Daljeet, Kaur Jagroop, "Analysis of Super Market using Association Rule Mining", International Journal of Advanced Research in Computer Science, Volume 8, No 7, August 2017.

[19] Khurana, K., and Sharma, S. "A comparative analysis of association rule mining algorithms". International Journal of Scientific and Research Publications, Volume 3, Issue 5, May 2013,

[20] Lavanya D., Rani K.Usha, "A Hybrid Approach to Improve Classification with Cascading of Data Mining Tasks", International Journal of Application or Innovation in Engineering & Management, Volume 2, Issue 1, January 2013.

[21] Lichman, M., "UCI Machine Learning Repository", [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2013.

[22] Nuntawut Kaoungku, Phatcharawan Chinthaisong, Kittisak Kerdprasop, Nittaya Kerdprasop, "Discretization and Imputation Techniques for Quantitative Data Mining", Proceedings of the International Multi Conference of Engineers and Computer Scientists, March 13 - 15, 2013.

[23] Pigott T D, "A review of missing data treatment methods", Educational Research and Evaluation, 2001.

[24] Rameshkumar K., "A Novel Algorithm For Association Rule Mining From Data With Incomplete And Missing Values", ICTACT Journal On Soft Computing, Volume 01, Issue 04, April 2011.

[25] Riyanarto Sarno, Dewandono Rahadian Dustrial, Tohari Ahmad, Farid Naufal Mohammad and Fernandes Sinaga, "Hybrid Association Rule Learning and Process Mining for Fraud Detection", IAENG International Journal of Computer Science, April 2015.

[26] Sabthami J., Thirumoorthy K. and Muneeswaran K., "Mining Association Rules for Early Diagnosis of Diseases from Electronic Health Records", Middle-East Journal of Scientific Research, 2016.

[27] Sahaaya S.A., Mary Arul and Malarvizhi M., "A New Improved Weighted Association Rule Mining With Dyanmic Programming Approach For Predicting", A User's Next Access, Computer Science & Information Technology, 2012.

[28] Schmitt Peter, Mandel Jonas and Guedj Mickael,"A Comparison of Six Methods for Missing Data Imputation", Jornal of Biometrics and Biostatistics 2015.

[29] Shah Jaimeel M., Sahu Lokesh, "Recommendation based on Clustering and Association Rules", IJARIIE, Vol-1, Issue-2, 2015,

[30] Shrival Neelesh, Sahu Kapil, "Imputation of Missing Values using Hybrid Approach with Association Rule and K-Mean Clustering Algorithm", International Journal for Rapid Research in Engineering Technology & Applied Science, , Vol 2, Issue 7. August, 2016.

[31] Singh Gurdeep, Aggarwal Shruti, "Audio Classification based on Association and Hybrid Optimization Technique", International Journal of Computer Applications, Volume 120, June 2015.

[32] Singh Parvinder, Dahiya Vijay,"A Hybrid Algorithm Using Apriori Growth and Fp-Split Tree For Web Usage Mining", IOSR Journal of Computer Engineering, Volume 17, Issue 6, Ver. III, Nov – Dec., 2015.

[33] Soni Richa, Kaur Gurpreet, "Hybrid Recommendation Using Association Rule Mining By Partial Evaluation of Web Personalization for Retrival Effectiveness", International Journal of Advanced Research in Computer Engineering & Technology, Volume 3, Issue 2., February 2014.

[34] Vinjamuri Swathi, Reddy Sunitha M.," Music Recommendation System Using Association Rules", International Journal of Technology Enhancements and Emerging Engineering Research, Vol 2, Issue 7. 2014,

[35] Yi Zeng, Shiqun Yin, Jiangyue Liu, and Miao Zhang, "Research of Improved FP-Growth Algorithm in Association Rules Mining", Hindawi Scientific Programming, 2015.

[36] Young W, Weckman G, Holland W., "A survey of method¬ologies for the treatment of missing values within datasets: Limitations and benefits", Taylor and Francis. Jun; 12(1):15–43, 2010.