International Journal of Computer Science & Mechatronics
A peer reviewed international journal | Article Available at www.ijcsm.in | sjif-4.454
©smsamspublications.com | Vol.3.Issue.1.2017

A Survey on Data Preprocessing Techniques-Missing Value Imputation-Outlier Detection and Attribute Subset Selection

Nagavali Saka[1], Daveedu Raju Adidela[2], Radha M.S[3], Dr Srinivas Rao. G[4]
[1]Associate Professor, [2]Professor, [3]Assistant Professor, [4]Professor,
Department of Computer Science and Engineering, Ramachandra College of Engineering, Eluru, India,
Email: *vali214@gmail.com, 123.davidjoy@gmail.com, radha.renukesh@gmail.com, sr.goteti@gmail.com*

Abstract: With the advent of the rapid improvement in technology, huge spectrum of data has been accumulated among the databases around the globe. Most of the databases interest a, highly prioritized concept, the quality of the data. This quality of the data is achieved by various preprocessing techniques that implement prior to the application of any data mining modules through which, required knowledge is abstracted. This paper promotes various views of different methodologies adopted by researchers on three preprocessing techniques such as missing value imputation, outlier detection and attribute subset selection. The major objective is to provide technical aspects of the methodologies that the researchers have implemented. So far this work is adopted to facilitate the concept of preprocessing techniques for the new researchers in the field of mining to have a broader view of the various methods adopted by various researchers to avoid ambiguity in choosing the suitable method. This improves the quality of the knowledge discovery among databases by procedures such as classification, clustering techniques. It also provides various techniques with their advantages and disadvantages. An attempt is also made to connect one process of technique to other in each of the three techniques. Further various tools that can support for these procedures are briefed in the tabular form for easy conclusions. It concluded that the best choice of the method to be implemented for preprocessing the data varies with respect to the opted database.

Keywords: *Preprocessing, missing value imputation, Outlier detection, Attribute subset selection, Classification, Clustering.*

I. Introduction on Missing Value Imputation:

Some of the tuples in the database table have no recorded value for several attributes; those empty locations are called as missing values. Missing values lead to the difficulty of extracting useful information the database. The presence of missing values in a dataset can affect the performance of a classifier. Several methods have been proposed to treat missing data. The following are the different methods used by various researches.

1. List wise detection/Case Detection (CD)
2. Mean/Mode Imputation (MMI)
3. All Possible Value Imputation (APV)
4. Regression Method (RM)
5. Hot Deck Imputation (HDI)
6. K-Nearest Neighbor Imputation (KNN)
7. Multiple Imputation (MI)
8. Maximization Likelihood method (ML)
9. Internal Treatment Method for C4.5 (C4.5)
10. Bayesian Iteration Imputation (BII)

List Wise Detection/ Case Deletion (CD):

List wise detection method omits the cases/instances with missing data and does analysis on the remaining data values [1]. Though it is the most common used method, it has two obvious disadvantages: a) A substantial decrease in the size of dataset available for the analysis. b) Data are not always missing completely at random. This method will bias the data distribution and statistical analysis. A variation of this method is to delete the attributes with high missing rate. But before deleting any attribute, it is necessary to run relevance analysis, especially on the attributes with high levels of missing data.

1. Mean/Mode Imputations (MMI)

In this method replace a missing data with the mean (numeric attribute) or mode (nominal attribute) of all cases observed [2]. To reduce the effect of exceptional data, median can also be used. This is one of the most common used methods. But there are some problems. Using constants to replace missing data will change the characteristic of the original dataset; ignoring the relationship among attributes will bias the data mining algorithms. A variation of this method is to replace the missing data for a given attribute by the mean or mode of all known values of that attribute in the class where the instance with missing data belongs.

2. All Possible Value Imputation (APV)

All Possible Value Imputation method consists of replacing the missing data for a given attribute by all possible values of that attribute [3]. In this method, an instance with a missing data will be replaced by a set of new instances. If there are more than one attribute with missing data, the substitution for one attribute will be done first, then the next attribute be done, etc., until all attributes with missing data are replaced. This method also has a variation. All possible values of the attribute in the class are used to replace the missing data of the given attribute. That is restricting the method to the class.

3. Regression Method (RM)

Regression imputation assumes that the value of one variable changes in some linear way with other variables [4]. The missing data are replaced by a linear regression function instead of replacing all missing data with a statistics. This method depends on the assumption of linear relationship between attributes. But in the most case, the relationship is not linear. Predict the missing data in a linear way will bias the model.

4. Hot Deck Imputation (HDI)

In Hot Deck Imputation method, missing attribute value is filled in with a value from an estimated distribution for the missing value from the current data [5]. It is typically implemented in two stages. a) Data are partitioned into clusters. b) Missing data are replaced within a cluster. This can be done by calculating the mean or mode of the attribute within a cluster. In Random Hot deck, a missing value of attribute is replaced by an observed value of the attribute chosen randomly.

5. K-Nearest Neighbor Imputation (KNN)

This method uses k-nearest neighbor algorithms to estimate and replace missing data. The main advantages of this method are that a) it can estimate both qualitative attributes i.e., the most frequent value among the k nearest neighbors and quantitative attributes i.e., the mean of the k nearest neighbors, b) It is not necessary to build a predictive model for each attribute with missing data. Efficiency is the biggest trouble for this method [6]. While the k-nearest neighbor algorithms look for the most similar instances, the whole dataset should be searched. However, the dataset is usually very huge for searching. On the other hand, the procedure of the selection of the value “k” and the measure of similar will impact the result more.

6. Multiple Imputation (MI)

The basic idea of Multiple Imputation method is that: a) A model which incorporates random variation is used to impute missing data. b) To do this M times, producing M complete datasets; c) Run the analysis on each complete dataset and average the results of M cases to produce a single one [6]. For Multiple Imputation, the dataset must be missing at random. In general, multiple methods are performing the deterministic imputation methods. They can introduce random error in the imputation process and get approximately unbiased estimates of all parameters. But the cost of calculating is too high for this method to implement in practice.

7. Maximization Likelihood Method (ML)

Maximization Likelihood use all data observed in a database to construct the best possible first and second order moment estimates. It does not impute any data, but rather a vector of means and a covariance matrix among the variables in a database. This method is a development of expectation maximization (EM) approach [7]. One advantage is that it has well-known statistical foundation. Disadvantages include the assumption of original data distribution and the assumption of incomplete missing at random.

8. Internal Treatment Method for C4.5 (C4.5)

Decision tree C4.5 is a widely accepted classifier. One of the improvements for C4.5 is the development of the internal algorithms for missing data treatment [7]. It uses probability approaches to handle missing data a) Select attribute according to the correctional information gain ratio and the correctional generation depends on the proportion of missing data on the attribute. b) All the instances with missing data are distributed into all the subsets according to the probability and the probability depends on the size of the subset they belonged to. c) While the decision tree is used to classify the new instance, all the possible paths are searched and then give a classification result in the form of probability, if the instance have missing data on the training attribute.

9. Bayesian Iteration Imputation (BII)

Naive Bayesian Classifier is a popular classifier, not only for its good performance, but also for its simple form. It is not sensitive to missing data and the efficiency of calculation is very high. Bayesian Iteration Imputation uses Naive Bayesian Classifier to impute the missing data [8]. It is consisted of two phases: a) Decide the order of the attribute to be treated according to some measurements such as information gain, missing rate, weighted index, etc.; b) Using the Naive Bayesian Classifier to estimate missing data. It is an iterative and repeating process. The algorithms replace missing data in the first attribute defined in phase one, and then turn to the next attribute on the base of those attributes which have be filled in. Generally, it is not necessary to replace all the missing data (usually 3~4 attributes) and the times for iterative can be reduced.

The following table, Table 1 provides the information of various missing data manipulation techniques and the corresponding tools which supports the imputation techniques. The supporting tools indicated in the tables are described in the section IV.

Table 1. Techniques for dealing with missing data

S.No	Method for Imputation	Supporting tool
1	List wise detection/Case Detection(CD)	KEEL,R
2	Mean/Mode Imputation(MMI)	WEKA,R
3	All Possible Value Imputation (APV)	KEEL
4	Regression Method(RM)	R
5	Hot Deck Imputation(HDI)	WEKA
6	K-Nearest Neighbor Imputation (KNN)	WEKA
7	Multiple Imputation(MI)	R,WEKA,KEEL
8	Maximization Likelihood method(ML)	KEEL
9	Internal Treatment Method for C4.5(C4.5)	KEEL,WEKA
10	Bayesian Iteration Imputation(BII)	KEEL

II. Introduction on Outlier Detection

There exist data objects that do not comply with the general behavior or model of the data, such data objects, which are grossly different from or inconsistent with the remaining set of data are called outliers. Outlier detection has been used for centuries to detect and, where appropriate, remove anomalous observations from data. Outliers arise due to mechanical faults, changes in system behavior, fraudulent behavior, human error, instrument error or simply through natural deviations in populations. Their detection can identify system faults and fraud before they escalate with potentially catastrophic consequences. It can identify errors and remove their contaminating effect on the data set and as such to purify the data for processing. The original outlier detection methods were arbitrary but now, principled and systematic techniques are used, drawn from the full gamut of Computer Science and Statistics.

The following are the different methods used to detect outlier by various researches.

1. Statistical approach
2. The Distance based approach
3. Density based local outlier
4. Deviation based approach
5. Depth based approach
6. High dimensional approach
7. Bounded maximum asymptotic bias
8. Connectivity based outlier detection scheme
9. Clustering
10. Wavelet based outlier detection approach

1. Statistical Approach

Given a certain kind of statistical distribution (e.g., Gaussian) Compute the parameters assuming all data points have been Compute the parameters assuming all data points have been generated by such a statistical distribution (e.g., mean and standard deviation) Outliers are points that have a low probability to be generated by the overall distribution (e.g., deviate more than 3 times the standard deviation from the mean) [8].

A huge number of different tests are available differing in , type of data distribution (e.g., Gaussian) ,number of variables i e dimensions of the data objects Number of variables, dimensions of the data objects (univariate/multivariate), number of distributions (mixture models), parametric versus non-parametric (e.g., histogram-based).

2. **The Distance Based Approach**

The notion of distance-based (DB) outlier is been defined by Knorr and Ng. An object O in a dataset T is a DB(p,D)-outlier if at least fraction p of the objects in T lie greater than distance D from Other concept of DB-outlier is well defined for any dimensional dataset [8]. The parameter p is the minimum fraction of objects in a data space that must be outside an outlier D-neighborhood. This notion generalizes many concepts from distribution-based approach and better faces computational complexity. It is further extended based on the distance of a point from its k-th nearest neighbor.

3. **Density Based Local Outlier**

This method assigns to each object a degree to be an outlier. This degree is called the local outlier factor (LOF) of an object [8]. It is "local" in the sense that the degree depends on how isolated the object is with respect to the surrounding neighborhood. In LOF algorithm, outliers are data objects with high LOF values whereas data objects with low LOF values are likely to be normal with respect to their neighborhood. High LOF is an indication of low-density neighborhood and hence high potential of being outlier (Mansur & Noor,). In order to understand LOF method, it is necessary to define some auxiliary notions. (Breunig et al,) for Distribution Based Approach

These methods are typically found in statistics textbooks. They deploy some standard distribution model (Normal, Poisson, etc.) and flag as outliers those data which deviate from the model. However, most distribution models typically apply directly to the future space and are univariate i.e., having very few degrees of freedom. Thus, they are unsuitable even for moderately high-dimensional data sets [9].

4. **Depth Based Approach**

This approach is based on computational geometry and computes different layers of k-d convex hulls (Johnson et al., 1998). Based on some definition of depth, data objects are organized in convex hull layers in data space according to peeling depth and outliers are expected to be found from data objects with shallow depth values. In theory, depth-based methods could work in high dimensional data space [9].

5. **Subspace Method for High Dimensional Approach**

Main aim of the subspace outliers detection methods are finding outliers in relevant subspaces that are not outliers in the full dimensional space (where they are covered by irrelevant) attributes. To identify which subspace is relevant there are many techniques developed [10]. This approach resembles a grid based subspace clustering approach but not searching dense but sparse grid cells. Report objects contained within sparse grid cells as outliers evolutionary search for those grid cells (Apriority-like search not possible, complete search not feasible) problems with this approach is increasing dimensionality, the expected value of a grid cell quickly becomes too low to find significantly sparse grid cells

6. **Bounded Maximum Asymptotic Bias**

The maximum asymptotic bias, is carried over here is to multivariate outlier identifiers [11]. This method show how this term depends on the respective biases of estimators which are used to construct the identifier. It turns out that the use of high-breakdown robust estimators is not sufficient to achieve outlier identifiers with bounded maximum asymptotic bias.

7. **Connectivity Based Outlier Detection Scheme**

In this study, we propose an incremental connectivity outlier factor (COF) [11]. The main idea of the static COF algorithm is to assign to each data record a degree of being outlier. This degree is called the connectivity-based outlier factor (COF) of a data record. Data records (points) with a high COF have average local connectivity smaller than their neighborhood and typically represent strong outliers, unlike data records belonging to uniform clusters that usually tend to have lower COF values.

8. **Clustering**

Clustering is a basic method to detect potential outliers. From the viewpoint of a clustering algorithm, potential outliers are data which are not located in any cluster [12]. Furthermore, if a cluster significantly differs from other clusters, the objects in this cluster might be discovery of clusters with arbitrary shape; Good efficiency on large databases.

9. **Wavelet Based Outlier Detection Approach**

The wavelet transform or wavelet analysis is probably one of the most recent solutions to overcome the shortcomings of the Fourier transform. In wavelet analysis the use of a fully scalable modulated window solves the signal-cutting problem [13]. The window is shifted along the signal and for every position the spectrum is calculated. This process is repeated many times with a slightly shorter (or longer) window for every new cycle. The processing result in detecting the outlier based on wavelet.

The following Table 2, summarizes the various outlier detection techniques and corresponding which handles the techniques.

Table 2. Techniques for dealing with outlier detection

S.No	**Method for outlier detection**	**Supporting tool**
1	Statistical approach	Rapid miner
2	The Distance based approach	Rapidminer,Weka
3	Density based local outlier	Rapidminer,Weka
4	Deviation based approach	Rapid miner
5	Depth based approach	Rapid miner
6	High dimensional approach	Weka
7	Bounded maximum asymptotic bias	Weka
8	Connectivity based outlier detection scheme	R
9	Clustering	Rapid miner, Weka
10	Wavelet based outlier detection approach	Rapid miner

III. Introduction on Attribute Subset Selection

Attribute subset selection reduces the data set by removing irrelevant or redundant attributes or dimensions. The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. There are several aims to feature selection.

- To reduce the size of the problem reducing compute time and space required to run our algorithms.
- To improve classifiers. Firstly by removing noisy or irrelevant features. Secondly by reducing the likelihood of over fitting to noisy data.
- To identify which features may be relevant to a specific problem. For example, to demonstrate which gene expressions are relevant in a certain disease

The following are the different Attribute subset selection methods used by various researches.

1. Person's correlation coefficient
2. Relief
3. Ensemble with data permutation
4. Mutual Information
5. Greedy Forward Search
6. Exhaustive search
7. Ranked Forward search
8. Refined exhaustive search
9. Feature ranking and feature subset selection(FR&FS)
10. The sampling and Feature ranking(S&FR)

1. **Person's Correlation Coefficient**

The correlation of two sets of data is measured by this method. That is, how much does a variation in one data set affect the variation in another [14]. We might want to use this as if there is a high correlation between one feature and the class of the data we are using then this will most likely be good at separating the data and will be useful for classification purposes.

2. **Relief**

This method (first proposed in works through a data set looking at the separation capabilities of randomly selected instances. It does this by selecting for each instance both the nearest same-class instance and the nearest opposite-class instance [15]. These are then used to calculate a weighting for each feature which is iteratively updated with each stochastically chosen data point. This method can be highly useful when there is a large amount of data. Time complexity is not an is-sue as a constant number of trials is performed. This means that the relief algorithm may complete quicker than other methods which require all the data to be taken into account.

3. **Ensemble with Data Permutation**

This approach interacts with learning algorithm at a lower computational cost than the wrapper approach. It also captures feature dependencies. It considers not only relations between one input features and the output feature, but also searches locally for features that allow better local discrimination. It uses the independent criteria to decide the optimal subsets for a known cardinality. And then, the learning algorithm is used to select the final optimal subset among the optimal subsets across different cardinality [16].

4. **Mutual Information**

This is an indicator of the shared information between two variables and \measures how much knowing one of these variables reduces uncertainty about the other" [17]. This works similar to person's coefficient and is included to see how well it performs in comparison. Again, if there is a lot of shared information between a feature and the class of our labeled data then this is a good indicator that this feature is important and useful in distinguishing members of one class from another.

5. **Greedy Forward Search**

This method is a greedy finite difference calculation. It works by making changes to the set of features and only keeping the new set if there is an increase in accuracy [17]. Greedy Forward Search works by starting with just one feature and incrementally adding in all the other features. As each feature is added in, the classifier is evaluated with the feature set and the new feature is only kept if there is a notable increase in accuracy.

This is a greedy solution and may not find the absolute optimum feature set, however by looking at which features cause an increase in accuracy it will pick out useful features to the classification scheme. Also, because the accuracy of the classier is evaluated with all the features in a set, this method will pick out features which work well together for classification. Features are not assumed to be independent and so advantages may be gained from looking at their combined effect.

6. **Exhaustive Search**

In this method, we explore a brute force approach as mentioned in. This means looking at every possible combination of features to find which one gives the best result [18]. It is of course only possible to do this with a small number of features and so some simplification of this problem must be done.

7. **Ranked Forward Search**

In this method introduce a ranking to the attribute features and then perform a greedy forward search over the given ranking. It is hoped that only including the variables which increase the accuracy will lead to a better result than thresholding (as in the filter approach). For this to work we are assuming that there will be some co-dependence between features which are highly ranked [18]. The ranking methods we have used so far have all assumed that features are independent and so the ranking of each feature is not a affected by its correlation with other features. Evaluating the ranked set of features with a greedy algorithm should discover some correlator behavior between the highly ranked features

8. **Refined Exhaustive Search**

As previously suggested it is an incomprehensibly massive task to explore the entire feature space of any problem big enough for feature selection to be worthwhile. That is to say that if a problem had a low number of features then it would be relatively easy to look at every possible combination of those and determine which combination was best. But if a problem has a low number of features, it will possibly be obvious which to choose or more likely be the case that all the features are relevant and necessary to classification. Ironically, it is only when the feature space gets large that we want to start employing expensive methods such as a brute force search [19].

The aim of this method is to constrain the feature space to some useful subset over which an exhaustive search may be performed. To do this, we first rank the variables according to some alter method criterion. We then select the best n variables to be our new feature space. We search in this feature space for every subset of r variables. We can work out the number of possible combinations we will compute as:

$$nCr = n! / r!(n-r)!$$

Using this we can select sensible values for n and r. It should be noted that this does not limit n and r to small values.

9. Feature Ranking and Feature Subset Selection (FR&FS)

Filter-based feature ranking method (ranker) rank features independently without involving any learning algorithm (learner or classifier), and then the best features are selected from the ranking list. Researchers have developed a large number of rankers to rank features [20].

The FR&FS hybrid approach is a combination of feature ranking and feature subset selection methods. There are two steps for generating feature subset. First, a feature ranking list is generated using corresponding filter-based rankers and input into the next step. Second, the top k features from the ranking list are selected and then the feature subset selection method is applied to the reduced dataset.

10. The Sampling and Feature Ranking (S&FR):

The S&FR hybrid approach is a combination of sampling and feature ranking methods [16]. Through sampling methods, class distributions are updated by utilizing different sample sizes, and then ranking list is created by using corresponding rankers. After that, the top k features are selected. The following table, Table 3 summarizes the various Attribute subset selection techniques and corresponding which handle the techniques.

Table 3 demonstrates the techniques for dealing with Attribute subset selection

S.No	Method for outlier detection	Supported Tool
1	Person's correlation coefficient	Weka ,R
2	Relief	KEEL,WEKA
3	Ensemble with data permutation	Yale, WEKA, KEEL
4	Mutual Information	KEEL
5	Greedy Forward Search	KEEL,WEKA
6	Exhaustive search	WEKA
7	Ranked Forward search	KEEL,WEKA
8	Refined exhaustive search	WEKA
9	Feature ranking and feature subset selection(FR&FS)	WEKA
10	The sampling and Feature ranking(S&FR)	WEKA

IV. Tools:

Three effective free software tools are described and its supporting methods also described in the three tables Table 1, Table 2, and Table 3.

1. YALE:

It supports the process of experiment design and evaluation. YALE is an environment for machine learning experiments. A modular operator concept allows the design of complex nested operator chains for a huge number of learning problems. The data handling is transparent to the operators. YALE is widely used by researchers and data mining companies [21].

2. WEKA:

Weka is a data mining package which includes a wide variety of methods. It's easy to use interface makes it accessible for general use, while its flexibility and extensibility make it suitable for academic use.

Weka is written in Java and released under the GNU General Public License (GPL). It has an API which allows the algorithms to be called from other programs. Weka can be run on Windows, Linux, Mac and other platforms.

Other programs such as Rapid Miner and R can access Weka routines [22].

3. KEEL:

KEEL is a software tool to assess evolutionary algorithms for Data Mining problems including regression, classification, clustering, and pattern mining and so on. It contains a big collection of classical knowledge extraction algorithms, preprocessing techniques (instance selection, feature selection, discretization, imputation methods for missing values, etc.), Computational Intelligence based learning algorithms, including evolutionary rule learning algorithms based on different approaches (Pittsburgh, Michigan and IRL), and hybrid models such as genetic fuzzy systems, evolutionary neural networks, etc.

It allows us to perform a complete analysis of any learning model in comparison to existing ones, including a statistical test module for comparison. Moreover, KEEL has been designed with a double goal: research and educational [23].

R: R is a well-supported, open source, command line driven, statistics package. There are hundreds of extra "packages" available free, which provide all sorts of data mining, machine learning and statistical techniques. It has a large number of users, particularly in the areas of bio-informatics and social science.

The package is well designed; John Chambers received the ACM 1998 Software System Award for "S" (which R is based on). The citation said that Dr. Chambers' work "will forever alter the way people analyze, visualize, and manipulate data" [24].

V. Conclusions:

Three major preprocessing techniques such as missing value imputation, outlier detection, and attribute subset selection are explained which accompanied with its corresponding implementations. These are precisely described to be handy for the new researchers who concentrate on the preprocessing techniques specifically on structured data. The free data mining tools such as KEEL, WEKA, YALE (formerly Rapid Miner), and R are described and its supporting methods are demonstrated in the tables. This precise study will provide the budding data mining scientists an insight of the various preprocessing techniques and its supporting tools for easy management.

REFERENCES

[1] Roderick J., A. Little, "Regression With Missing X's: A Review", Journal of the American Statistical Association vol. 87, no. 420, pp. 1227-1237, 1992.

[2] Han J., and Kamber M., DataMining, "Concepts and Techniques", Morgan Kaufmann Publishers, Second Edition, 2006.

[3] Kurgan, L., Cios, K.J., Tadeusiewicz, R., Ogiela, M. & Goodenday, L.S., "Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis", Artificial Intelligence in Medicine, vol. 23, no. 2, pp. 149-169, 2001.

[4] Cios KJ, Kurgan LA., "Trends in data mining and knowledge discovery", Knowledge discovery in advanced information systems, Berlin: Springer series, 2002.

[5] Rubin B., "Multiple Imputation for Nonresponsive in Surveys", New York: John Wiley & Sons, vol. 12, no. 1, pp. 37-47, 1986.

[6] Kim H., and S. Yates., "Missing value algorithms in decision trees", Statistical Data Mining and Knowledge Discovery, pp. 155-172, 2004.

[7]. Batista G.E.A.P.A., and M.C. Monard, "An analysis of four Missing Data treatment methods for supervised learning", Applied Artificial Intelligence, vol. 17, pp. 519-33, 2001.

[8] Ester M., Kriegel H.-P., Sander J., Xu X.:"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Proc. 2nd International Conference. on Knowledge Discovery and Data Mining, Portland, pp. 226-231, 2002.

[9] Ester M, H.-P. Kriegel, J. Sander, and X. Xu. "A density-based algorithm for dis-covering clusters in large spatial databases with noise", IEEE Transactions on Knowledge and Data Engineering, vol. 5, no. 6, pp. 903-913, 1996.

[10].Berkhin P, "A survey of clustering data mining techniques; Grouping multidimensional data", Springer Berlin Heidelberg, pp. 25-71, 2006.

[11] Kaur, R., Kumar, G., & Kumar, K. A, "Comparative Study of Feature Selection Techniques for Intrusion Detection", International Journal of Computer Applications, International Conference on Advances in Emerging Technology, vol. 5, no. 7, pp. 340-350, 2013.

[12] I. H. Witten and E. Frank, "Data mining – Practical Machine Learning tools and Techniques (2nd Ed.) San Francisco: Morgan Kaufmann Publishers; 2005.

[13] Shardlow, Matthew. "An Analysis of Feature Selection Techniques." International Journal of Database Theory and Application. vol. 9, no. 9, pp. 75-82, 2016.

[14] Kira, K., & Rendell, L. A. "A practical approach to feature selection". International workshop on Machine learning", International Journal of Computer Applications, vol. 9, no. 6, pp. 67-77, 2012.

[15] Berkhin P, "A survey of clustering data mining techniques; Grouping multidimensional data", Springer Berlin Heidelberg, pp. 25-71, 2006.

[16] Ester M., Kriegel H.-P., Sander J., Xu X., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Proc. 2nd International Conference on Knowledge Discovery

and Data Mining, Portland, AAAI Press, pp. 226-231, 2002.

[17] Divya Tomar, Sonali Agarwal, "Data Pre-processing & Mining Algorithm, Knowledge & Data Mining & Preprocessing", International Journal of Database Theory and Application vol. 7, no. 4, pp. 99-128, 2014.

[18] Rubin, D.B. "Multiple Imputation for Nonresponsive in Surveys", Journal of the American Statistical Association, vol. 91, pp. 473-489, 1987.

[19] Kim H and S. Yates, "Missing value algorithms in decision trees", Statistical Data Mining and Knowledge Discovery, vol. 7, no. 4, pp. 155–172, 2003.

[20] Batista and M.C. Monard, "An analysis of four Missing Data treatment methods for supervised learning", Applied Artifcial Intelligence, vol. 17, pp. 519–533, 2003.

[21] http://www.the-data-mine.com/Software/YALE.

[22] http://www.the-data-mine.com/Software/weka

[23] http://www.the-data mine.com/Software/KeelDataMining

[24] http://www.the-data-mine.com/Software/R