I J C T A, 8(5), 2015, pp. 2119-2126

The Comprehensive Analysis for Mining the Knowledge using Feature Reduction Techniques in Medical Data

R. Prasanka*, M. Balamurugan**

Abstract: Feature reduction is removing unwanted attributes which is also known as variable selection, attribute selection or variable subset selection, and it is the process of selecting a subset of relevant features for use in model construction. The feature selection technique is used in various research fields. Data mining is a computational sequence to determine patterns in hefty data sets. Medical data mining using feature reduction method for medical field contains large number of attributes and information so dimensionality reduction is must now. This paper mainly focuses on work for feature reduction mechanism. The frame work which is proposed in paper tends to reduce the attributes without affecting the classification accuracy which would be the future direction.

Index Terms: Data Mining, Preprocessing, Medical data, Weka Tool, Feature Extraction

1. INTRODUCTION

The term Data mining refers to “extracting” or “mining” the knowledge from large volume of data. Data mining which means discovery of extracting large hidden data, previously unknown patterns and relationships that are difficult to detect with traditional statistics. Mining is the core process involved in Knowledge Discovery Process. Knowledge Discovery is a process of getting high level knowledge from low level data. The major challenge in the medical data mining is imprecision and uncertainty. Nowadays, offering cost worthy service is major problem faced by the health care organizations. It is essential to check more number of diagnose test, to predict a disease. By using intelligent diagnostic tool in the health care, there is possibility to predict the disease. The researchers in the medical field have succeeded in identifying and predicting the disease with the aid of Data mining techniques. This paper proposes a framework for medical data by applying data mining techniques for feature reduction.

In data mining, the hefty data sets are a computational progression to ascertain patterns. Data mining techniques are the result of a long process of research and product development [1]. It is the examination to excerpt concealed and previously mysterious patterns, to perceive the Traditional statistics relationships and knowledge that are difficult for further use of the overall goal of data mining process extract information and transform it into a comprehensible format from the data set. The role of Data Mining health care data is massive. Human decision making it is optimal; when it is poor the amount of data is huge to be classified. It cannot be allowed the enormous stress and overwork load resulted in poor inaccurate decision making which may lead to disastrous consequences in medical field. Based on improper information, it is acquired from medical data the most exorbitant and harmful mistake is performing decision making process.

Since now and then is not declined Institute of Medicine estimated that the effect of medical error. Medical history data, which comprises the essential to diagnose and on particular disease. The benefits of Data mining, it is conceivable to increase in health care by employing it as an intellectual symptomatic tool

* Research Scholar,

** Associate Professor, School of computer science, Engineering and Applications, Bharathidasan University, Tiruchirappalli-23.

[2] [3]. An association rule of Data Mining has been significantly used in health care data prediction. Data mining technique using association rule has been significantly health care data prediction [4][5][19]. To improve the quality and effectiveness of health care system, there is an eventual goal of knowledge discovery is to identify factors [6].

The remaining sections of the paper is formed as follows, in section II, the existing data mining techniques in feature reduction in medical data is discussed. Thereafter, hybridization technique framework is proposed in section III. Section IV presents the simulation results of Trees and Rules techniques and compared. Finally, the conclusion of this paper is presented.

2. RELATED WORKS

Heon Gyu Lee et al [7] proposed a technique to enhance the characteristics of heart rate variability using multi-parametric feature mechanism. M. A. Jabbar et. al [8] developed a heart disease prediction mechanism with the support of associative classification technique in data mining. Sellappan Palaniappan et al. [9] proposed an Intelligent Heart Disease Prediction System (IHDPS) with the help of Neural Network, Decision Trees and Naïve Byes. Liu H et al. [10] discussed the feature reduction technique which is used to find out the accuracy of heart diseases. They also discussed Feature reduction algorithm for classification and clustering.

Niti Guru et al. proposed a mechanism with the support of neural network for predicting the heart disease, blood pressure and sugar[11]. Charly, K used classification and prediction algorithms for decision trees, neural networks; clustering, association rules and regression [12]. From the medical point of view association rules are developed and preliminary results are defendable [13]. Carlos et al, formulate the interrelation technique and decision tree algorithm [14].

Ordonez et al. [15] proposed an association rules based constrained improved mapping algorithm for predicting heart diseases. This algorithm concentrates on finding useful constrains and mapping heart diseases

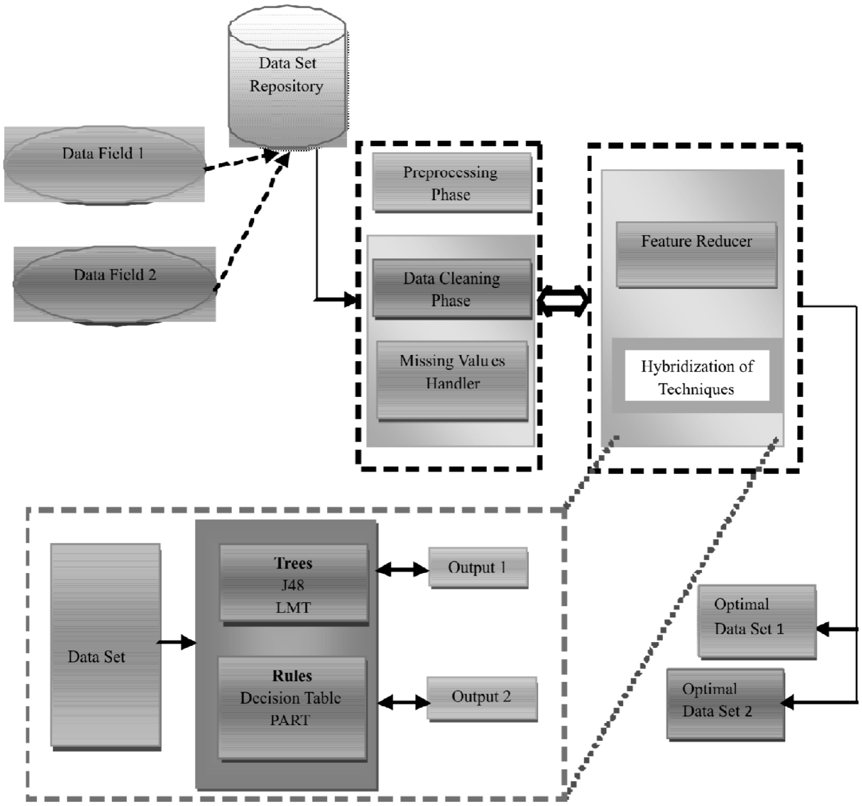


Figure 1: Proposed framework for feature reduction

data to applicable format for interrelation technique. The association rule algorithm employs various significant constraints which are used to reduce the rules count and improve the mining process speed.

3. PROPOSED FRAMEWORK

Preprocessing is an important step involves data mining process. The major task of preprocessing is cleaning, integration and transformation, reduction, discretization. Preprocessing is also the most critical steps in the data mining process, which deals with the preparation and transformation of the initial dataset. Raw data is highly susceptible to noise missing values and inconsistency. The quality of data a affects the data mining results.

3.1. Data Cleaning

Data cleaning is a one step of preprocessing. Collecting the data's for stored in data repository cleaning the dataset. Data cleaning is a real world data tend to be incomplete, noisy and inconsistent. Data cleaning (or data cleansing) routines attempt to full in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Noise data is a random error or variance in a measured variable [16].

3.2. Missing Values

Missing values and its problems are very common in the data cleaning process. Fill in the missing value manually in general, this approach is time-consuming and may be feasible given large data set with many missing value [18]. Many types of disease data set retrieve the particular disease dataset that missing the one type of rows add for the value. The different types of missing mechanisms are stated as below

1. MCAR
2. MAR
3. NAMR

MCAR

The term "Missing Completely At Random" refers to data where the missing value mechanism does not depend on the variable of interest, or any other variable, which is observed in the dataset Here the data are collected and observed arbitrarily and the collected data does not depend on any other variable of the dataset. Such type of missing data is very rarely found and the best method is to ignore such cases.

MAR

The term Missing At Random consider an entry Xi as missing at random if the data meets the requirement that missing values should not depend on the value of Xi after controlling for another variable. As an example, depressed people tend to have less income and thus the reported income now depends on the variable depression. As depressed people have lower income the percentage of missing data among depressed individuals [20].

NAMR

If the data is not missing at random or informatively missing then it is termed as "Not missing at Random". Such a situation occurs when the messing value mechanism depends on the actual value of missing data. Modeling such a condition is a very difficult task to achieve. When have a data with NMAR problem the only way to attain an estimate of parameters is to model the missing data. This means need to write a model for missing data and then integrate it into a more complex model for estimating missing values [19].

3.3. Classification

Classification is a supervised learning. Data mining algorithms follow three different learning approaches supervised, unsupervised, semisupervised. In supervised learning algorithm works with asset of examples whose labels are known. The labels can be nominal values in the case of the classification task or numerical values in the case of the regression task. In unsupervised learning in contrast, the labels of the examples in the data set are unknown and the algorithm typically aims at grouping examples according to the similarity of their attributes values characterizing task [17][20].this paper provides a survey of various feature reduction technique and classification techniques used for mining.

3.4. Filter

In the filter approach the attribute selection method is independent of the data mining algorithm to be applied to the selected attributes. A filter that creates a new dataset with a Boolean attribute replacing a nominal attribute [21].

3.5. Numeric to Nominal

The first step for preprocess and change the CSV file is a comma separate values format. CSV file importing into weka preprocessing step after applying classify convert the numeric to nominal for 10 cross-validation folds using a correct value of results all variables importing as numeric, need to change them to nominal. Before preprocessing kappa statistics value is 0.7 and after the conversion of numeric to nominal the value of kappa statistics is 0.8 which is an accurate one.

4. RESULTS AND DISCUSSIONS

4.1. Data Set

The dataset used for experimentation is taken from data mining repository of the University of California, Irvine (UCI). Data set from Cleveland Data set, Hungary Data set, Switzerland Data set, Long beach and Stat log Data set are collected. Cleveland, Hungary, Switzerland and Val Long Beach data set contain attributes. Among all the attributes, 14 attributes are taken for experimentation, since all the published experiments refer to using subset of 14 attributes. Researchers in the medical domain mostly use Cleveland data set and Stat log data set for testing purpose. This is because all the other data set has more number of missing values than Cleveland data set. This attributes removing noisy data’s. In the data set gathered for mining will contain either numeric attributes or nominal attributes. The data set collected for Heart disease encompasses both the numeric and nominal attributes.

J48

The J48 algorithm is used as data mining tool that tools that various algorithm to be applied on datasets. The J48 algorithm is used to univariate decision tree approach its results are discussed the multivariate approach is introduced as the linear machine approach that makes the use of the absolute error correction and also the thermal perception rules [21]. This algorithm is used to reduce the kappa statistics value accuracy.

LMT

Logistic Model Tree (LMT) predicts a numeric values given an instances that is defined over a fixed set of numeric to nominal attributes. The data set is subject to 10 fold cross validation. The cross validation technique is used to estimate the performance of a feature reduction model. In the 10 fold cross validation; the data sets are divided into 10 sets.

4.2. Decision Table

Weka tool is one of the Decision table. Classifying for rules model this model applies for after preprocessing and before convert the numeric to nominal attributes. Kappa statistics values in reduce accuracy in decision table for this algorithm.

PART

The Part and tree comparison for lower value in accuracy of rules. If the kappa statistic value is 0.6 is lower value. Comparing The TREE and RULES value of higher then tree value 0.6 is greater than 0. 7is after preprocessing then it is said to correct accuracy.

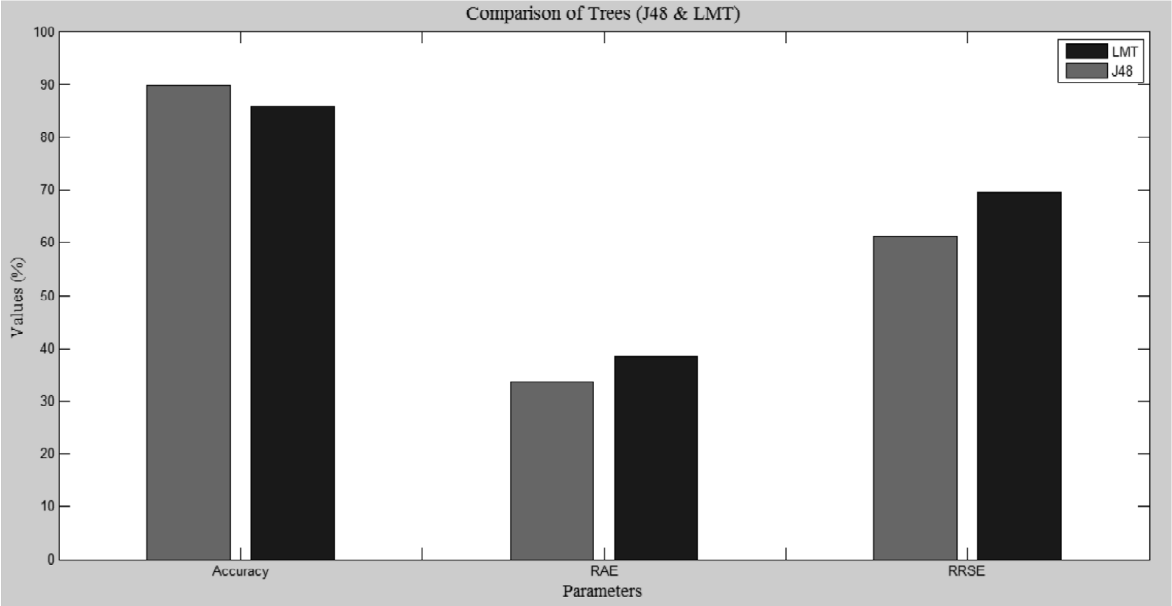
Table 1
Comparision-performance of Trees and Rules Table

Measures	Trees		Rules	
	J48	LMT	Decision Table	PART
Accuracy	89.8333	85.8333	80.5	82.8333
Relative Absolute Error	33.6511	38.316	61.095	41.5519
Root Relative Squared Error	61.2344	69.5173	76.093	61.2344

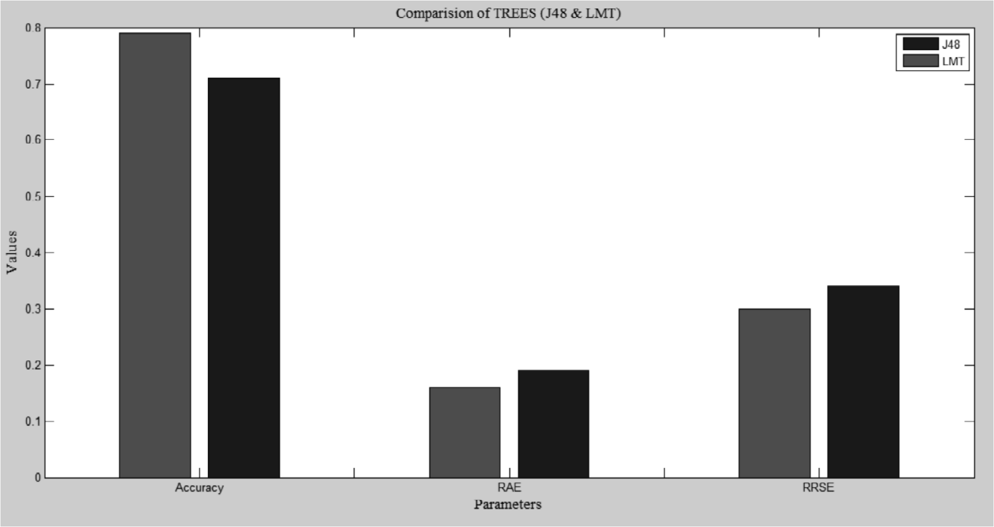
Table 2
Comparison-Performance of Trees and Rules Table

Measures	Trees		Rules	
	J48	LMT	Decision Table	PART
Kappa Statics	0.7942	0.7121	0.6056	0.6538
Mean Absolute Error	0.167	0.1902	0.3082	0.2026
Root Mean Squared Error	0.305	0.3463	0.379	0.3875

Kappa statistic value is higher than tree value for J48 is 0.7942 and LMT value is 0.7121 when the tree and rules compared with each other. By analyzing the value, tree value is higher than the Rules. Weka Tool is used to analyze the performance of trees and rules.

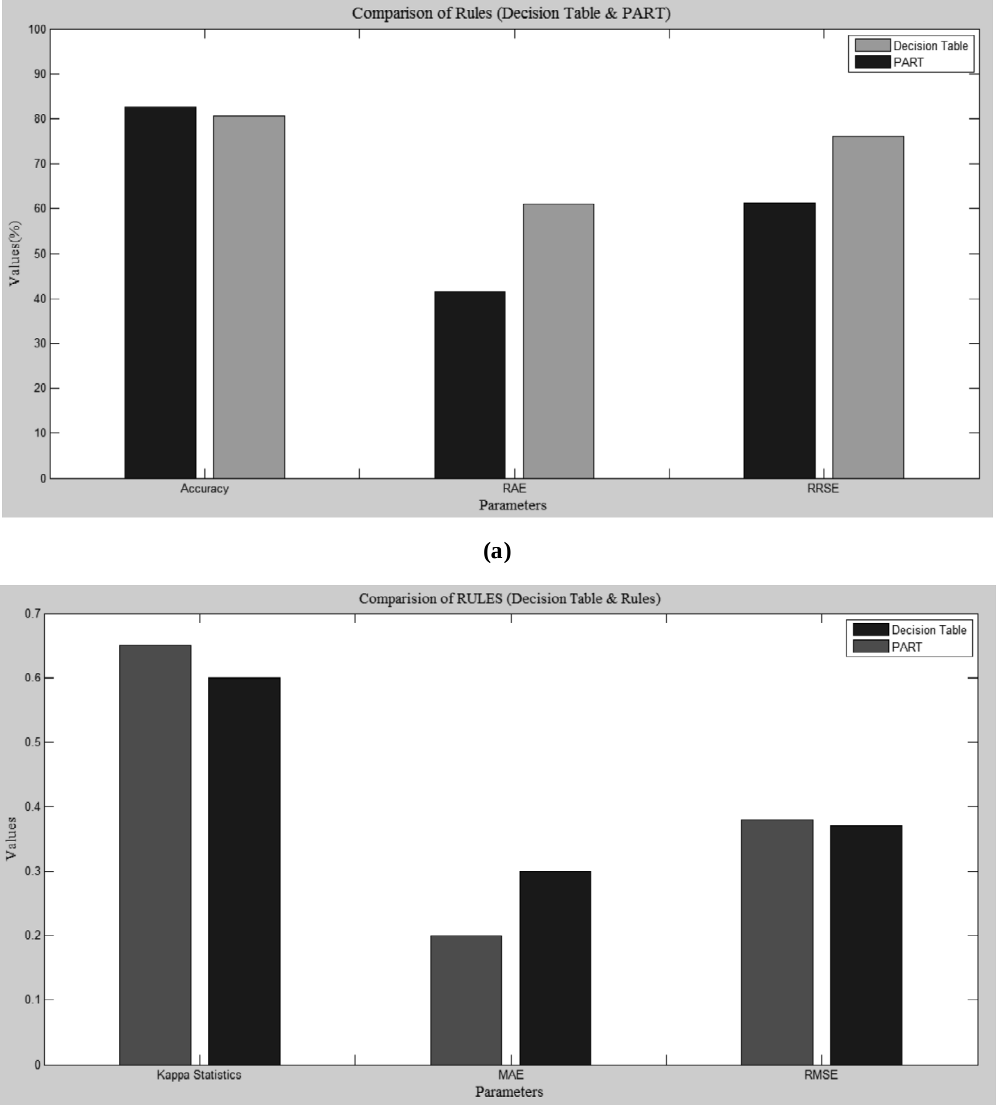


(a)



(b)

Figure 2: (a & b)Comparision of J48&LMT



(a)

(b)

Figure 3: (a & b) comparison of Decision Table PART

5. CONCLUSION

Medical Data Mining is a domain of challenge which involves a lot of imprecision and uncertainty. Provision of quality services at affordable cost is the major challenge faced in the health care organization. This work compares two feature reduction methods namely, Trees and Rules. From the experimentation, it is observed that J48 gives higher accuracy than LMT in trees algorithm. In case of Rules algorithm decision table produces. Higher value when compared with PART. Hence combining the J48 algorithm with Decision Table will produce more accuracy after reducing the attributes without affecting the level of accuracy. In this work, two existing algorithms from rules and trees were compared and the results were tabulated. Results are obtained without reducing the irrelevant features; the classification accuracy can be increased. The frame work which is proposed in paper tends to reduce the attributes without affecting the classification accuracy which would be the feature direction.

Refferences

[1] Bhagyashree Ambulkar and Vaishali Borkar "Data Mining in Cloud Computing", MPGINMC, Recent Trends in Computing, ISSN 0975-8887, pp. 23-26, 2012.

[2] Huan Liu and Hiroshi Motoda, Rudy Setiono and Zheng Zhao. "Feature Selection: An Everlasting Frontier in Data Mining", JMLR: The Workshop on Feature Selection and Data Mining, 2010.

[3] Rafiah Awang and Palaniappan. S. "Web based Heart Disease Decision Support System using Data Mining Classification Modeling techniques", Proceedings of iiWAS, pp. 177-187, 2007.

[4] Carlos Ordonez, Edward Omincenski and Levien de Braal "Mining Constraint Association Rules to Predict Heart Disease", Proceeding of 2001, IEEE International Conference of Data Mining, IEEE Computer Society, ISBN-0-7695-1119-8, pp. 433-440,2001.

[5] Deepika. N., "Association Rule for Classification of Heart Attack patients", IJAEST, Vol 11(2), pp. 253-257, 2011.

[6] Setiawan N.A, "Rule Selection for Coronary Artery Disease Diagnosis Based on Rough Set", International Journal of Recent Trends in Engineering", Vol 2(5), pp 198-202, Dec 2009.

[7] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV", LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2013.

[8] M.A. Jabbar et al., "Knowledge discovery using associative classification for heart disease prediction", AISC Vol. 182, pp. 29-39, 2012.

[9] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol. 8, No. 8, August 2008.

[10] Liu, H. and Yu, L. (2005), "Toward integrating feature selection algorithms for classification", IEEE Trans. Knowledge Data Engineering, vol. 17, pp., 491-502, 2005.

[11] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1, January-June 2007.

[12] Charly, K.: "Data Mining for the Enterprise", 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer", 7, 295-304, 1998.

[13] C. Ordonez, "Comparing Association Rules and Decision Trees for Disease Prediction", pp. 17-24, 2006.

[14] David Cooke, Carlos Ordonez, Ernest V. Garcia, Edward Omiecinski, Elyzabeth Krawczynska, Russell Folks, Cesar Santana, Levien de Braal, and Norberto Ezquerra." Data mining of large myocardial perfusion spect (mps) databases to improve diagnostic decision making", Journal of Nuclear Medicine, 40(5), 1999.

[15] C. Ordonez, "Mining Constrained Association Rules to Predict Heart Disease", In IEEE ICDM Conference, pp. 433–440,2001.

[16] I. Guyon, N. Matic and V.Vapnik,"Discovering Informative patterns and data cleaning", In:Fayyad UM, Piatetsky-Shapiro G, smyth Pand Uthurusamy R.(ed) Advanced in knowledge discovery and data mining, AAAI/MIT Press, California, pp. 353-357,1994.

[17] S. Vijaykumar, M. Balamurugan, S. G. Saravanakumar, Unique Sense: Smart Computing Prototype, Procedia Computer Science, Volume 50, 2015, Pages 223-228, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2015.04.056.

[18] Luengo, J, 2011 "Missing Values in Data Mining [Online] Available at:http://sci2s.ugr.es/MVDM/index.php [Accessed 2 September 2013]

[19] Vijaykumar, S., Saravanakumar, S., & Balamurugan, M. (2015). Unique Sense: Smart Computing Prototype for Industry 4.0 Revolution with IOT and Bigdata Implementation Model. *Indian Journal Of Science And Technology, 8*(35). doi:10.17485/ijst/2015/v8i35/86698.

[20] Heitjan, D.F., & Basu, S. (1996), “Distinguishing `Missing at Random’ and `Missing Completely at Random”. American Statistician, volume 50, issue 3 207,2013.

[21] Vijaykumar S, Dr. M. Balamurugan, Ranjani K, Big Data: Hadoop Cluster Deployment on ARM Architecture, International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 4, Special Issue 1, June 2015, ISSN 2278-1021 & 2319-5940.

[22] Haitovsky, Y,”Missing data in regression analysis”. Journal of the Royal Statistical Society, Series B (Methodological), Vol. 30, No. 1, pp. 67-82.

[23] Sunita Beniwal, Jitender Arora, “Classification and feature selection techniques In data mining”, IJERT Vol. 1Issue 6, August 2012.

[24] Gaganjot Kaur, Amit Chakra,”Improved J48 classification algorithm for the prediction of diabetes”. International Journal of computer applications, (0975–8887) Volume 98–No. 22, July 2014.