MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 1, Issue 1, January-June 2016

Comparison of Various Discretization Techniques in Healthcare Data

S.Subhitsha[1], V.P.Sumathi[2]

[1] Student, [2]Assistant Professor,

Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India.

E-mail: *subhitsha96@gmail.com[1],sumathi.vp.cse@kct.ac.in[2]*

Abstract

Data mining is widely used in healthcare applications. However, data in healthcare are continuous in nature and data mining algorithms cannot directly deal with it. Discretization of continuous attributes improves the efficiency of data mining algorithms. This paper compares the performance of different discretization techniques on three different University of California, Irvine (UCI) datasets namely Wisconsin Diagnostic Breast Cancer (WDBC), Breast tissue and PIMA Indian diabetes in terms of classification accuracy. The aim of this study is to identify the discretization technique that helps to improve the accuracy of the classification process. The experimental results have showed that the entropy based discretization gives the highest classification accuracy using Naïve Bayes classifier.

Keywords: *Breast Cancer, Breast Tissue, Indian Diabetes, Classification, Discretization*

INTRODUCTION

Data mining is widely used by many organizations around the world to discover interesting patterns. Healthcare industry generates huge amounts of data and data mining algorithms play a major role in the prediction and diagnosis of the diseases. Proper diagnosis of any human disease is a challenging task for the health care organizations. Data mining techniques facilitate healthcare systems to analyze the data to improve patient care by identifying effective treatments and best practices. Healthcare data mining provides numerous potential for hidden pattern investigation from the clinical data sets. These patterns can be used by medical experts to diagnose and predict treatments

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

for patients. Medical data mining is very much valuable to medical experts.

Classification analysis is the organization of data in given classes. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. Some classification algorithms can only take nominal data as inputs while others discretize numeric data into nominal data during the learning process. Therefore, discretization is needed as a pre-processing step to partition each numeric feature into a finite set of adjacent distinct intervals/items. Discretization methods can be classified as supervised or unsupervised depending on whether it uses class information of data sets. Supervised methods make use of the class label when partitioning the continuous features whereas unsupervised discretization methods do not require the class information to discretize continuous attributes. Supervised discretization can be further characterized as error-based, entropy-based or statistics based. Equal- width and equal-frequency are unsupervised methods. A good discretization algorithm should not only describe the original data to produce but also help to improve the classification performance.

This paper compares the performance of various discretization algorithms on three UCI medical datasets [9] using R language. Different preprocessing methods, discretization, feature selection and classification algorithms are examined with three different datasets with Breast cancer, breast tissue and diabetes (PIMA) and compared based on their accuracy.

The main objective of this work is to improve the accuracy of classification by using various discretization techniques. Most of the existing work in this area has concentrated on improving the accuracy of the classifiers without discretizing the data and have been implemented using Weka, Tanagara and Rapid Miner. The algorithms in this work have been implemented using R language. R is a powerful statistical language widely used for data analysis.

This paper is organized as follows. Section 2 gives a study of the related works. Section 3 explains about the various discretization and classification algorithms. Section 4 gives a detailed description of the data sets. The experimental results are discussed in section 5 and section 6 gives the conclusion.

RELATED WORK

Johana et al.[1] have used natural language processing techniques and have generated a training set consisting of 53 rules to predict the various stages of cancer. Different supervised

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

classification algorithms were used to analyze the PIMA Indian diabetes data set using Tanagra tool. The performance of various data mining algorithms using 10 fold cross validation were also compared [2]. The performance of conventional supervised learning algorithms were used to analyze the breast cancer data set based on precision and recall parameters. The results have shown that random tree outperforms all other algorithms with highest accuracy rate [3].

Senturk et al.[4] in their work have analyzed the breast cancer data set using Rapid Miner data mining tool. They have shown that SVM has better performance accuracy than other algorithms. Kumar et al. [5] have done a comparison of different data mining classifiers on breast cancer data set and have found that SVM has the highest accuracy. Kaur et al. [6] proposed a new approach for effectively predicting diabetes from the Pima Indian Diabetes data set. They have used modified J48 classifier in WEKA and have achieved accuracy of 99.87%.

Rahman et al.[7] have investigated the performance of classification algorithms based on WEKA, Tanagra and MATLAB. They have found that Naïve Bayes in Tanagra gives highest accuracy of 100%. Jacob et al. [8] have highlighted the importance of feature selection in improving the performance of classification algorithm. Their study has shown that Rnd classification algorithm gives 100% accuracy in classifying the data sets.

METHODOLOGY

Discretization is an important step in data preprocessing. It reduces the number of distinct values for a given continuous variable by dividing its range into a finite set of disjoint intervals, and then relates these intervals with meaningful labels. Discretization methods can be supervised or unsupervised depending on the usage of class labels. . Supervised methods make use of the class label when partitioning the continuous features while unsupervised discretization methods do not require the class information to discretize continuous attributes. The different discretization methods used in this study are given below:

Equal Width Discretization

Equal-width interval discretization is a simplest discretization method. The algorithm divides the range of observed values for an attribute into k equal sized bins, where k is a parameter provided by the user. The process involves sorting the observed values of a continuous attribute and finding the minimum, Vmin and maximum, Vmax, values. The interval can be computed by dividing the range of observed values for the variable into k equally sized bins using the following

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

formula, where k is a parameter supplied by the user. The steps are given in figure 1.

Equal Frequency Discretization

The equal-frequency algorithm sorts all values in ascending order, determines the minimum and maximum values of the attributes and divides the sorted continuous values into k intervals such that each interval contains approximately n/k data instances with adjacent values where n is the total number of instances. The steps are shown in figure 2.

Algorithm: Equal Width Discretization

Input: Attribute a, interval k

Output: Discretized value of attribute a

Step1: Identify minimum and maximum value of the attribute a, store it in V_{min} and V_{max}.

Step2: Divides the attribute range into k intervals of equal size, where k is a parameter provided by the user.

Step3: Compute the width of intervals $W = (V_{max} - V_{min}) / k$.

Fig. 1. Equal Width Discretization

Algorithm: Equal Frequency Discretization

Step 1: Sort the values of the attribute X to be discretized in ascending order. Step 2:

Find the number of all possible values for X.

Step 3: Divide the values of X into the user-specified k number of intervals, where each interval contains the same number of sorted sequential values.

Fig. 2. Equal Frequency Discretization

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

Entropy Based Discretization

Entropy-based discretization is a supervised method that will use the class information entropy of candidate partitions to select boundaries for discretization. The steps in Entropy based discretization are shown in figure 3.

Feature Selection

Feature selection is another important step in the classification process and is used for dimensionality reduction. It is a process of selecting the subset of attributes from the data set that are most significant .It is used to reduce the number of attributes in a high dimensional data set which will improve the classification accuracy at reduced costs.

Classification Algorithms

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome. Two classification algorithms have been used in this paper to test the accuracy of discretization techniques.

Decision Tree (J48)

Decision trees are powerful tools for classification and prediction. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the attributes. The algorithm constructs the tree starting from a training set. It uses divide and conquer strategy to create a tree structure which represents the rules. It calculates the information gain for each attribute and then selects the attribute with the highest information gain. The algorithm is given in figure 4.

Algorithm: Entropy based Discretization

Step 1: Given a set of samples S, if S is partitioned into two intervals S1 and S2 using boundary T, the entropy after partitioning is

$$E(S,T) = \frac{|s_1|}{|s|} Ent(s_1) + \frac{|s_2|}{|s|} Ent(s_2)$$

Step 2: The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization.

Step 3: The process is recursively applied to partitions obtained until some stopping criterion is met, e.g., $Ent(S) - E(T, S) > \delta$

Fig. 3 Entropy Based Discretization

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

Algorithm: Decision Tree

Step 1:Select an attribute and formulate a logical test on attribute

Step 2:Branch on each outcome of test, move subset of examples (training data) satisfying that outcome to the corresponding child node.

Step 3:Run recursively on each child node.

Step 4:Termination rule specifies when to declare a leaf node.

Fig. 4 Decision Tree Construction

Naïve Bayes:

Naïve Bayes classifier is a variant of Bayesian network with the simple structure that has the class node as the parent node of all other attributes. Naive Bayes classifier estimates the class- conditional probability by assuming that the attributes are conditionally independent, given the class label Ck. Suppose that there are n classes, C1,C2, . . . ,Cn, the conditional independence assumption can be formally stated in Equation(1)

$$P(X \mid C_k) = \prod_{i=1}^{m} P(X_i \mid C_k) \quad (1)$$

where every attribute set X = $\{x_1, x_2, \ldots, x_m\}$ consists of *m* attributes.

DATA SET DESCRIPTION

A series of experiments with different datasets from UCI Machine Learning Repository was carried out. These datasets vary extensively in the number of instances and the dimension of the attribute space. The different datasets used for the study is described below.

Wisconsin Diagnostic Breast Cancer (Wdbc)

The WDBC datasets are made at the University of Wisconsin Hospital for the diagnosis and prognosis of breast tumors solely based on Fine Needle Aspiration(FNA) test. This dataset contains a total of 699 clinical cases, with 458 benign (non-cancerous) and 241 malignant (cancerous) cases. Each and every clinical case has 9 attributes with assigned integer values and one class output with a binary value of either 2 or 4, showing benign and malignant breast cancer diagnoses, separately .

Breast Tissue

This is a dataset with electrical impedance measurements in samples of freshly excised tissue from the Breast. This dataset contains a total of 106 clinical cases, with

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

54 pathological and 52 normal cases and there are 10 attributes.

Pima Indian Diabetes

A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases is a collection of medical diagnostic reports of 768 with 10 attributes each.

Each dataset consists of some classification patterns or instances with a set of numerical features or attributes and is shown in Table 1.

Table 1. Description of the data sets

Data Set	No. of Attributes	No. of Instances	No. of Classes
WDBC	11	699	2
Breast Tissue	10	106	2
PIMA Indian Diabetes	10	768	2

EXPERIMENTAL RESULTS

The experiment was carried out in R language on the three different data sets using 2,5,and 10 fold cross validation using entropy based discretization, equal width discretization (4 bins) and equal frequency discretization (4 bins). The resulted datasets after feature selection were used for classification using decision tree (J48) and Naïve Bayes algorithms to estimate the performance of the datasets. The performance measure used to evaluate the classification effectiveness is classification accuracy. The classification accuracy refers to the percentage of correct predictions made by the model, compared to the actual classifications in the test data. The accuracy percentage obtained using J48 and Naïve Bayes algorithms for 3, 4 and 5 attributes selected using feature selection are given in Tables 2, 3 and 4 respectively. Figure 5 shows the comparison of accuracies for the two classifiers with different folds of cross validation. Accuracy is the proportion of

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

instances whose class the classifier can correctly predict.

Accuracy = (TP + TN) / (TP + TN + FP + FN) (2)

- True Positive (TP) is the number of correct predictions that an instance is positive.
- False Positive (FP) is the number of incorrect predictions that an instance is positive.
- False Negative (FN) is the number of incorrect predictions that an instance is negative.

True Negative (TN) is the number of correct predictions that an instance is negative.

From Table 2, it can be seen that in the WDBC, entropy based discretization gives the highest average accuracy of 95.73% using J48 and 98.9% using Naïve Bayes. In the case of breast tissue data set, equal width discretization gives average accuracy of 81.08% using J48, whereas it gives an average accuracy of 84.91% for entropy based discretization technique in Naïve Bayes. In PIMA Indian Diabetes dataset, the entropy based discretization with J48 gives 75.3% average accuracy whereas it is 75.87% using entropy with Naïve Bayes algorithm.

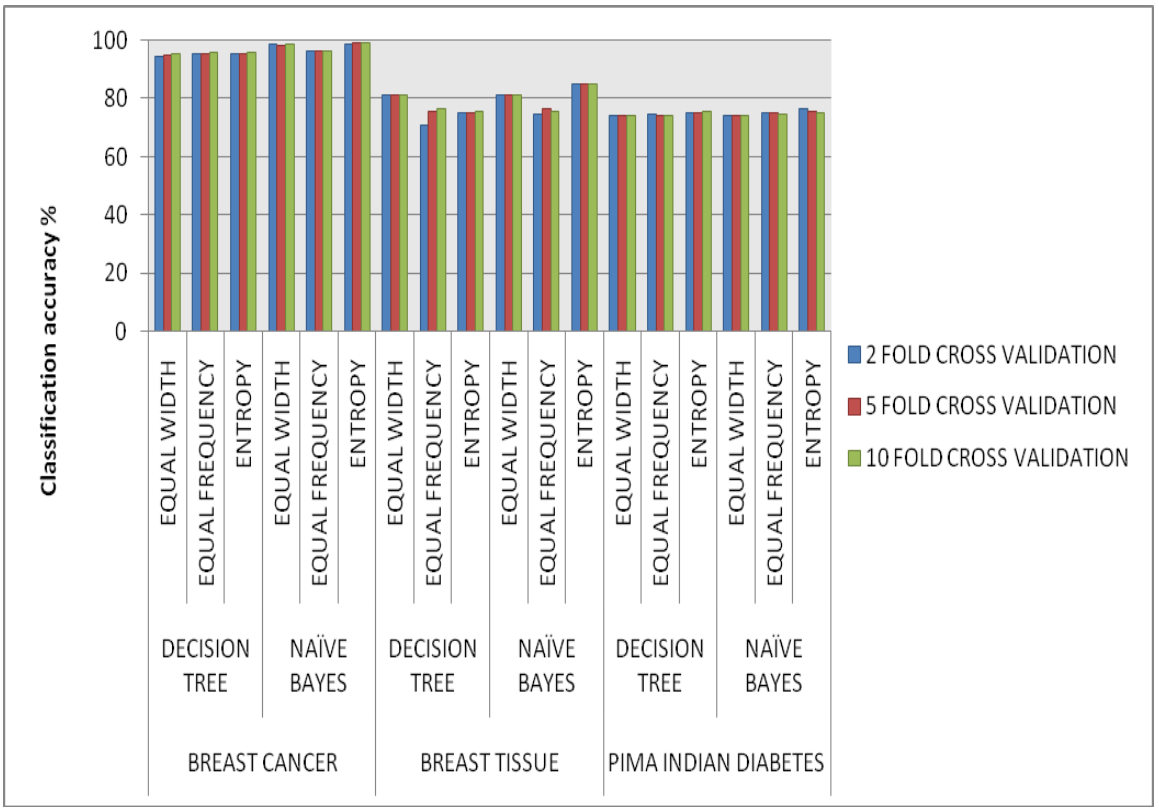


Figure 5. Comparison of Discretization Techniques (3 attributes)

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

Table 2. Classifier accuracy percentage for J48 and Naïve Bayes algorithms (3 Attributes)

DATA SETS	CLASSIFICATION ALGORITHM	DISCRETIZATION TECHNIQUE	ACCURACY IN %			
			2 FOLD CROSS VALIDATION	5 FOLD CROSS VALIDATION	10 FOLD CROSS VALIDATION	AVERAGE ACCURACY
WDBC	DECISION TREE	EQUAL WIDTH	94.56	94.7	95.56	94.94
		EQUAL FREQUENCY	95.57	95.42	95.7	95.56
		ENTROPY	95.14	95.27	**95.71**	**95.73**
	NAÏVE BAYES	EQUAL WIDTH	98.57	98.42	98.57	98.52
		EQUAL FREQUENCY	96.28	96.28	96.28	96.28
		ENTROPY	98.57	98.99	**99.14**	**98.9**
BREAST TISSUE	DECISION TREE	EQUAL WIDTH	**81.13**	81.11	81.01	**81.08**
		EQUAL FREQUENCY	70.75	75.47	76.41	74.21
		ENTROPY	75.13	75.25	75.52	75.3
		EQUAL WIDTH	81.13	81.13	81.13	81.13

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

	NAÏVE BAYES	EQUAL FREQUENCY	74.53	76.42	75.47	75.47
		ENTROPY	**84.91**	**84.91**	**84.91**	**84.91**
PIMA INDIAN DIABETES	DECISION TREE	EQUAL WIDTH	74.09	74.09	74.08	74.09
		EQUAL FREQUENCY	74.61	73.96	73.96	74.18
		ENTROPY	75.13	75.25	**75.52**	**75.3**
	NAÏVE BAYES	EQUAL WIDTH	74.35	74.22	74.35	74.31
		EQUAL FREQUENCY	75	75	74.74	74.91
		ENTROPY	**76.69**	75.65	75.26	**75.87**

Table 3. Classifier accuracy percentage for J48 and Naïve Bayes algorithms (4 Attributes)

DATA SETS	**CLASSIFICATION ALGORITHM**	**DISCRETIZATION TECHNIQUE**	**ACCURACY IN %**			
			2 FOLD CROSS VALIDATION	**5 FOLD CROSS VALIDATION**	**10 FOLD CROSS VALIDATION**	**AVERAGE ACCURACY**
	DECISION TREE	EQUAL WIDTH	94.13	95.28	95.27	94.89
		EQUAL FREQUENCY	95.57	95.42	**95.71**	**95.57**
		ENTROPY	94.85	94.27	95.13	94.75
	NAÏVE BAYES	EQUAL WIDTH	98.56	98.57	98.71	98.61

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

BREAST CANCER		EQUAL FREQUENCY	96.42	96.42	96.42	96.42
		ENTROPY	98.57	98.71	**98.85**	**98.71**
BREAST TISSUE	**DECISION TREE**	EQUAL WIDTH	81.13	81.21	81	81.11
		EQUAL FREQUENCY	70.75	76.41	76.41	74.52
		ENTROPY	**84.91**	84.79	84.79	**84.83**
	NAÏVE BAYES	EQUAL WIDTH	81.13	81.13	81.13	81.13
		EQUAL FREQUENCY	72.64	77.34	76.41	75.46
		ENTROPY	**84.91**	**84.91**	**84.91**	**84.91**
PIMA INDIAN DIABETES	**DECISION TREE**	EQUAL WIDTH	72.4	74.09	74.09	73.53
		EQUAL FREQUENCY	74.61	73.7	74.61	74.31
		ENTROPY	**76.17**	75.13	74.48	**75.26**
	NAÏVE BAYES	EQUAL WIDTH	73.95	73.69	73.7	73.78
		EQUAL FREQUENCY	74.61	73.94	74.09	74.21
		ENTROPY	76.17	76.43	**77.08**	**76.56**

In Table 3, it can be seen that the in WDBC data set, equal frequency discretization gives the highest average accuracy of 95.57% using J48 and 98.71% using entropy based discretization with Naïve Bayes. In the case of breast tissue data set, entropy based discretization gives the highest average accuracy of 84.83% using J48 and 84.91% in Naïve Bayes. In PIMA Indian Diabetes dataset, the entropy based discretization with J48 gives 75.26% average accuracy whereas it is 76.56% using entropy with Naïve Bayes algorithm.

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

Table 4. Classifier accuracy percentage for J48 and Naïve Bayes algorithms (5 Attributes)

DATA SETS	CLASSIFICATION ALGORITHM	DISCRETIZATION TECHNIQUE	ACCURACY IN %			
			2 FOLD CROSS VALIDATION	5 FOLD CROSS VALIDATION	10 FOLD CROSS VALIDATION	AVERAGE ACCURACY
BREAST CANCER	DECISION TREE	EQUAL WIDTH	93.99	94.28	95.28	94.52
		EQUAL FREQUENCY	95.84	95.14	**95.85**	**95.61**
		ENTROPY	94.56	95.13	95.71	95.13
	NAÏVE BAYES	EQUAL WIDTH	97.71	98.14	98.28	98.04
		EQUAL FREQUENCY	96.56	96.42	96.42	96.47
		ENTROPY	97.99	98.14	**98.3**	**98.14**
BREAST TISSUE	DECISION TREE	EQUAL WIDTH	81.13	81.02	80.88	81.01
		EQUAL FREQUENCY	83.96	**85.84**	84.91	84.91
		ENTROPY	84.91	84.97	85.05	**84.98**
	NAÏVE BAYES	EQUAL WIDTH	81.13	81.13	81.13	81.13
		EQUAL FREQUENCY	83.96	83.96	83.96	83.96
		ENTROPY	**84.91**	**84.91**	**84.91**	**84.91**
PIMA INDIAN DIABETES	DECISION TREE	EQUAL WIDTH	71.48	74.88	76.04	74.13
		EQUAL FREQUENCY	74.61	74.35	75.26	74.74
		ENTROPY	76.69	77.33	**78.52**	**77.51**
	NAÏVE BAYES	EQUAL WIDTH	73.44	75.13	74.09	74.22
		EQUAL FREQUENCY	76.3	76.3	76.3	76.3
		ENTROPY	**78.78**	78.65	78.65	**78.69**

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

In Table 4, it can be seen that the in WDBC data set, equal frequency discretization gives the highest average accuracy of 95.61% using J48 and entropy based discretization gives the highest accuracy of 98.14% using Naïve Bayes. In the case of breast tissue data set, equal frequency discretization gives an average accuracy of 84.98% using J48 and 84.91% using Naïve Bayes with entropy based discretization. In PIMA Indian Diabetes dataset, the entropy based discretization with J48 gives 77.51% accuracy whereas the average accuracy is 78.69% using entropy with Naïve Bayes algorithm.

Comparing all the above results, it is seen that the entropy based discretization with Naïve Bayes classifier performs well on all the three datasets.

CONCLUSION

This paper compares the performance of various discretization techniques on three different UCI datasets. It has been found that the performance of classification techniques vary based on the discretization technique used. The results showed that the entropy based discretization outperforms equal width and equal frequency discretization. In terms of classification Naïve Bayes algorithm gives better classification results. The results reveal that when feature selection is used, discretization greatly enhances the accuracy of classification.

REFERENCES

1. Rani, G., Dennis Gladis, and Joy Mammen. "Classification and Prediction of Breast Cancer Data derived Using Natural Language Processing."Proceedings of the Third International Symposium on Women in Computing and Informatics. ACM, 2015.

2. Karthikeyani, V., and I. Parvin Begum. "Comparison a performance of data mining algorithms (CPDMA) in prediction of diabetes disease." International Journal on Computer Science and Engineering 5.3 (2013): 205.

3. Shajahaan, S. Syed, S. Shanthi, and V. M. Chitra. "Application of Data Mining techniques to model breast cancer data." International Journal of Emerging Technology and Advanced Engineering 3.11 (2013): 362-369.

4. Senturk, Zehra Karapinar, and Resul Kara. "Breast Cancer Diagnosis via Data Mining: Performance Analysis of Seven different algorithms." Computer Science & Engineering 4.1 (2014): 35.

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved

5. Kumar, G. Ravi, Dr GA Ramachandra, and K. Nagamani. "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques."International Journal of Innovations in Engineering and Technology (IJIET) 2.4 (2013): 139.

6. Jacob, Shomona Gracia, and R. Geetha Ramani. "Discovery of knowledge patterns in clinical data through data mining algorithms: multi-class categorization of breast tissue data." International Journal of Computer Applications (IJCA) 32.7 (2011): 46-53.

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

8. Rahman, Rashedur M., and Farhana Afroz. "Comparison of various classification techniques using different data mining tools for diabetes diagnosis." Journal of Software Engineering and Applications 6.03 (2013): 85.

9. Kaur, Gaganjot, and Amit Chhabra. "Improved J48 classification algorithm for the prediction of diabetes." International Journal of Computer Applications 98.22 (2014).

Page 1-14 © MANTECH PUBLIATIONS 2016. All Rights Reserved