ISSN:2231-0711

A. Rajalakshmi et al | IJCSET(www.ijcset.net) | August 2016 | Vol 6, Issue 8, 293-298

Data Discretization Technique Using WEKA Tool

A. Rajalakshmi[1], R. Vinodhini[2], K. Fathima Bibi[3]

M.Phil Research Scholar[1, 2], Assistant professor[3]

Department of Computer Science

Rajah Serfoji Govt. College (Autonomous), Thanjavur-5.

Abstract - Knowledge Discovery from Data defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" data pre-processing is an essential step in the process of Knowledge Discovery. The goal of pre-processing is to help improve the quality of data, consequently the mining results. Real-world data sets consist of continuous attributes. Many algorithms related to data mining require the continuous attributes need to be transformed into discrete. Discretization is a process of dividing a continuous attribute into a finite set of intervals to generate an attribute with small number of distinct values. In this paper handle continuous values of iris data set taken from UCI machine learning repository. Discretization filter applied in iris data set using WEKA Tool and also data set used in various classification algorithms namely J48, Random Forest, RepTree, Naïve Bayes, RBF network, OneR, BF Tree, and Decision Table. The performance measures are Accuracy and Error Rate noted both before and after discretization, and it shows that discretization improves the classification accuracy in iris data set.

Keywords: ***Classification, Discretization, Pre-processing, WEKA Tool, Decision Tree***

1. INTRODUCTION

Data Mining (DM), the extraction of hidden predictive facts from huge databases is a potent novel technology with great potential to study vital facts in the data warehouse. DM searches databases for unseen patterns, discovering predictive information that professionals may miss, as it goes away from their outlooks. Several individuals treat DM as a replacement for alternative widespread used term, Knowledge Discovery from Data, or KDD. Knowledge discovery as a process consists of an iterative classification of the subsequent steps: Data cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation, and Knowledge Presentation. Discretization is needed to change from continuous attributes to discrete attributes in order to increase the accurateness in prediction.

1.1 Problem Definition

This research concentrates on the problem of discretization. It is one of the pre-processing techniques. Most of the classification tasks require the data to be in the discrete form to be able to perform the mining process. Discretization filter is used in the iris dataset. The filter changes the continuous values into discrete values. The research aims to achieve higher accuracy in classification and reduced error rate.

1.2 DATA PRE-PROCESSING

Data preparation and filtering steps can take considerable amount of processing time. Pre-processing is to transform the data set in order to remove inconsistencies, noise and redundancies There are many pre-processing techniques [6. The major tasks in data pre-processing are: data cleaning, data reduction, data integration, and data transformation. This paper organized as follows: in section 2 we review some related works, the methodology and tools used for the experiment are covered in section 3, results and discussion described in section 4, finally the paper ends with conclusion and knowledge gained in section 5.

2. RELATED WORK

[Mangesh metkari et al., (2015)] proposed a system to classify medical data to help the doctors while making the decision in cases disease of patients. This system employs two classification techniques Genetic Algorithm (GA) and Artificial Neural Network (ANN) to predict heart disease of a patient. ANN and GA used for the classification of heart disease data set. Finally, accuracy results of both ANN and GA of heart disease data set with and without discretization are compared. Experimental results are carried out on Heart disease data set using the four approaches.

[Hemada et al., (2013)] presented a new discretization technique, which considers the maximum frequent value in each class as initial cut-points and applies the Entropy-MDLP method between the initial cut-points to find the final cut-points. Since the technique is essentially pre-processing (all the cut points are found prior to learning), it does not have to use the binary splitting approach necessary to reduce complexity during learning. As a result, the discretization is multi-interval, which is the optimal choice for maximum possible discrimination between the classes.

[Elsayad radwan et al., (2013)] described rough sets to classify Thyroid in the presence of missing bases and build the Modified Similarity Relations that is dependent on the number of missing bases with respect to the number of the whole defined attributes for each rule. The Thyroid relation attributes are converted to suitable representation for rough set analysis by discretization and then constructing a matrix where each row corresponding to the similarity score between Thyroid attributes and each column corresponding to a defined attribute that describe the position of bases inside the rule.

[Tajun han et al., (2015)] proposed a post-processing method that can improve the quality of discretization. After the normal discretization process, the boundary point of the discretization for each attribute was adjusted and then after evaluating the group effect of the adjusted point. The results of the empirical experiments show that the adjusted data set improves the classification accuracy. The proposed method can be used with any discretization algorithms, and improve their discretization power.

3. MATERIAL AND METHODOLOGY

Data Source

1. Iris Plant Data Set

Number of Features: 4 numeric, predictive features and the class

Feature Information:

1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal width in cm
5. Class:
   -- Iris Setosa
   -- Iris Versicolour
   -- Iris Virginica

Number of Instances: 150 (50 in each of three classes)
Missing Feature Values: None
Class Distribution: 33.3% for each of 3 classes.

2. Weka Tool

The Waikato Environment for Knowledge Analysis (WEKA) is a machine learning toolkit introduced by Waikato University, New Zealand. At the time of the project's inception in 1992. WEKA would not only provide a toolbox of learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation. It can be run on Windows, Linux and Mac. It consists of collection of machine learning algorithms for implementing data mining tasks. Data can be loaded from various sources, including files, URLs and databases. Supported file formats include WEKA"s own ARFF format, CSV, Lib SVM"s format, and C4.5"s format.

The second panel in the Explorer gives access to WEKA"s classification and regression algorithms [12]. The corresponding panel is called "Classify" because regression techniques are viewed as predictors of "continuous classes". By default, the panel runs a cross validation for a selected learning algorithm on the dataset that has been prepared in the Pre-process panel to estimate predictive performance.

METHODOLOGY

Many real-world data sets predominately consist of continuous attributes also called quantitative attributes. These types of data sets are unsuitable for certain data mining algorithms that deals only nominal attributes. So that we need to transform continuous attributes into nominal attributes, this process known as discretization. Discretization filter is applied in Fisher's iris data set [14]. The performance measures namely accuracy and error rate will be noted both before and after discretization using various classification algorithms. The methodology of the research work is as follows:

1. Discretization
2. Binning
   2.1 Equal width binning
   2.2 Equal frequency binning
3. Classification Algorithms
   3.1 Tree
   3.2 Bayes
   3.3 Rules
   3.4 Function

1. Discretization

Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values [5]. This leads to a concise, easy-to-use, knowledge-level representation of mining results. Data discretization can perform before or while doing data mining. Most of the real data set usually contains continuous attributes. Some machine learning algorithms that can handle both continuous and discrete attributes perform better with discrete-valued attributes. Discretization involves:

- Divide the ranges of continuous attribute into intervals
- Some classification algorithms only accept categorical attributes
- Reduce data size by discretization
- Prepare for further analysis

Discretization techniques are often used by the classification algorithms. Unsupervised discretization algorithms that do not use class information that divides continuous ranges into sub-ranges [8]. Discretization involves several advantages. Some of them are given below:

- Discretization will reduce the number of continuous features values, which brings smaller demands on system's storage.
- Discretization makes learning more accurate and faster.
- In addition to many advantages of having discrete data over continuous one, a suite of classification learning algorithms can only deal with discrete data.
- Data can also be reduced and simplified through discretization. For both users and experts, discrete features are easier to understand, use, and explain.

TABLE 1
DISCRETIZATION

S.No	Feature	Subset	Range
1	Sepal length	S1	{5.0 to 5.9}
		S2	{6.0 to 6.9}
		S3	{7.0 to 7.9}
2	Sepal width	S1	{2.0 to 2.9}
		S2	{3.0 to 3.9}
3	Petal length	S1	{1.0 to 1.9}
		S2	{4.0 to 4.9}
		S3	{5.0 to 5.9}
		S4	{6.0 to 6.9}
4	Petal width	S1	{0.0 to 0.9}
		S2	{1.0 to 1.9}
		S3	{2.0 to 2.9}

Discretization performed manually in the iris data set, and the values divided into subsets.s Results are shown in Table1, and also WEKA Discretization shown in Table 2.

TABLE 2
WEKA DISCRETIZATION

S.No	Feature	Subset	Range
1	Sepal length	S1	{-∞ to 5.5}
		S2	{5.5 to 6.1}
		S3	{6.1 to ∞}
2	Sepal width	S1	{-∞ to2.9}
		S2	{2.9 to 3.3}
		S3	{3.3 to ∞}
3	Petal length	S1	{-∞ to 2.4}
		S2	{2.4 to 4.7}
		S3	{4.7 to ∞}
4	Petal width	S1	{-∞ to 0.8}
		S2	{0.8 to 1.7}
		S3	{1.7 to ∞}

2. BINNING

In the unsupervised methods, continuous ranges are divided into sub-ranges by the user specified parameter – for instance, equal width (specifying range of values), equal frequency (number of instances in each interval)

2.1 EQUAL WIDTH BINNING (EWB)

The simplest unsupervised discretization method, which determines the minimum and maximum values of the discretized attribute and then divides the range into the user-defined number of equal width discrete intervals [8]. There is no "best" number of bins, and different bin sizes can reveal different features of the data. The following table 3 containing the values of accuracy and error rate which depending the number of bins used. There are five number of bins were used and they are 2,4,5,10,40.

TABLE 3
EQUAL WIDTH BINNING

No.of bins	Acc in %	E.R in %
2	78.66	21.33
4	90.66	9.33
5	93.33	6.66
10	96.00	4.00
40	95.33	4.66

2.2 EQUAL FREQUENCY BINNING (EFB)

The unsupervised method, which divides the sorted values into k intervals so that each interval contains approximately the same number of training instances. Thus each interval contains n/k (possibly duplicated) adjacent values. k is a user predefined parameter. Here k represents the bin value. In equal frequency binning an equal number of continuous values are placed in each bin [4].

TABLE 4
EQUAL FREQUENCY BINNING

No.of bins	Acc in %	E.R in %
2	74.66	25.33
4	88.00	12.00
5	94.00	6.00
10	91.33	8.66
40	90.00	10.00

3. CLASSIFICATION ALGORITHMS

Classification is a form of data analysis that can be used to extract models describing important data classes. Whereas classification predicts categorical (discrete, unordered) labels.

3.1 TREE BASED CLASSIFICATION

A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Decision trees represent a supervised approach to classification. Decision trees are trees that classify instances by sorting them based on feature values.

3.1.1 J48 TREE

C4.5 (J48) algorithm is an improvement of IDE3 algorithm, developed by Quinlan Ross (1993) based on Hunt's algorithm is serially implemented like ID3. C4.5 has an enhanced method of tree pruning by replacing the internal node with a leaf node thereby reducing misclassification errors due to noise or too many details in the training data set. WEKA implements decision tree C4.5 algorithm using "J48 decision tree classifier".The explanation of the C4.5 algorithm as well as the J48 [9] implementation is as follows:

- Whenever a set of items (training set) is encountered, the algorithm identifies the attribute that discriminates the various instances most clearly.
- Among the possible values of this feature, if there is any value for which there is no ambiguity
- For all other cases, another set of attributes are looked at that gives the highest information gain.

3.1.2 RANDOM FOREST

The random forest algorithm was developed by Leo Breiman, a statistician at the University of California, Berkeley. Random forests, a meta-learner comprised of many individual trees, was designed to operate quickly over large datasets and more importantly to be diverse by using random samples to build each tree in the forest[13].

3.1.3 REP TREE

REP Tree (reduced error pruning tree) algorithm is a fast decision tree learner. It builds a decision/ regression tree using information gain/variance and prunes it using reduced-error pruning (with back-fitting). The algorithm only once sorts the values for numeric attributes. Missing values are dealt with by splitting the corresponding instances into pieces [9].

3.1.4 BF TREE

In BF tree learners the "best" node is expanded first as compared to standard DT learners such as C4.5 and CART which expand nodes in depth-first order. The "best" node is the node whose split leads to maximum reduction of impurity among all nodes available for splitting. The resulting tree will be the same when fully grown; just the order in which it is built is different. BF tree constructs binary trees, i.e., each internal node has exactly two outgoing edges. This method adds the "best" split node to the tree in each step.

3.2 BAYES CLASSIFICATION

Bayesian classifier is statistical classifier based on bayes theorem. It can be used to predict class membership probabilities. The probability that the tuple that belongs to the particular class or not. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets.

3.2.1 NAIVE-BAYES

Naive-Bayes classifiers are generally easy to understand and the induction of these classifiers is extremely fast, requiring only a single pass through the data if all the attributes are discrete. Naive-Bayes classifiers are also very simple and easy to understand. Naïve Bayesian classifiers are very robust to irrelevant attributes and classification takes into account evidence from many attributes to make the final prediction.

3.3 RULES

Rule based classification algorithm also known as separate-and-conquer method is an iterative process consisting in first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover.

3.3.1 ONE R

OneR or “One Rule” is a simple algorithm proposed by Holt. The OneR builds one rule for each attribute in the training data and then selects the rule with the smallest error rate as its one rule. The algorithm is based on ranking all the attributes based on the error rate [1]. To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to their majority class.

3.3.2 DECISION TABLE

Decision rules can be generated for each class. Typically, a decision table is used to represent the rules. Rough sets can also be used for attribute subset selection. The algorithm decision table is found using Weka classifiers under Rules [15].The simplest way of representing the output from machine learning is to put it in the same form as the input.

3.4 FUNCTION BASED CLASSIFICATION

Radial functions are simply a class of functions. In principle, they could be employed in any sort of model (linear or nonlinear) and any sort of network (single-Layer or Multi-Layer).

3.4.1 RBF NETWORK

Radial basis function networks (RBF Networks) have traditionally been associated with radial functions in a single–layer network. In WEKA RBF works as Class that implements a normalized Gaussian radial basis function network. It uses the k-means clustering algorithm to provide the basis functions.

4. RESULTS AND DISCUSSIONS

Theoretical studies of the discretization technique used in classification algorithms and binning methods are performed. For analyzing algorithms WEKA tool is used with tenfold cross validation. Two performance parameters have been considered for experimental evaluation. Following parameters examined both before and after discretization:

- Accuracy
- Error Rate

ACCURACY

In machine learning methods, the classification accuracy is often predicted by Tenfold cross-validation. In the process, the whole data set is split into ten parts, nine parts of the data set is used for learning and one for testing. This procedure is repeated ten times. Here eight classification algorithms were experimented using WEKA Tool [13].The accuracy of a classifier on a given data set is the percentage of data set tuples that are correctly classified by the classifier.

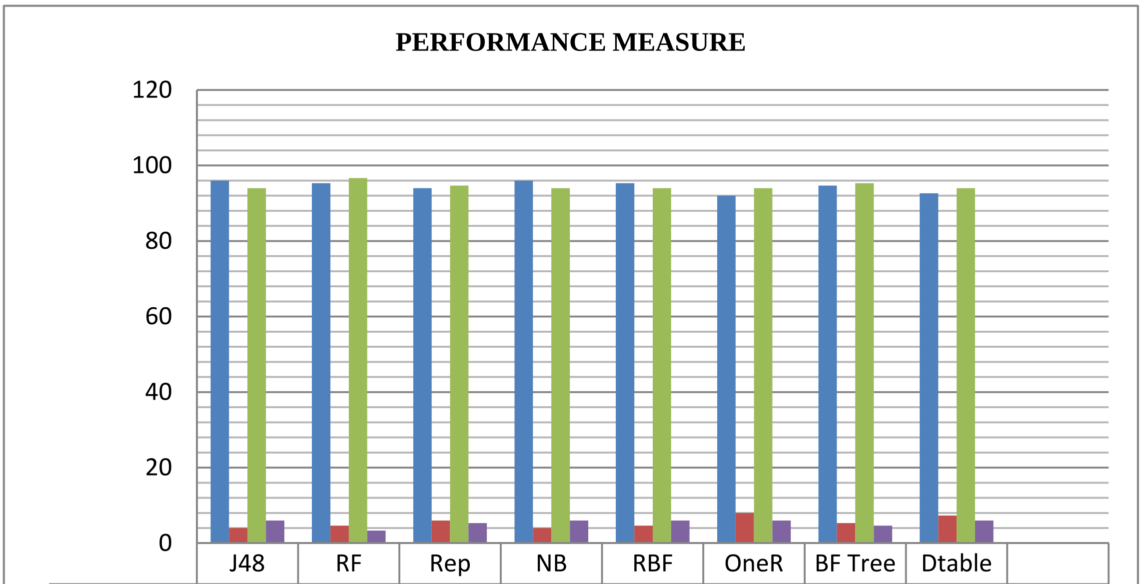
ERROR RATE

The error rate estimation can be computed as total loss from the k iterations, divided by the total number of initial tuples. In other words number of incorrectly classified instances during classification process determines the error rate. Table 5 consist the error rate in percentage. The table gives the accuracy, error rate for each model build by Tenfold cross-validation [6] for the individual classifier respectively.

TABLE 5
PERFORMANCE MEASURES

Classifiers	Before Disc.		After Disc.	
	Acc in %	ER in %	Acc in %	ER in %
J48	96.00	4.00	94.00	6.00
RF	95.33	4.66	96.66	3.33
RepTree	94.00	6.00	94.66	5.33
NBayes	96.00	4.00	94.00	6.00
RBF	95.33	4.66	94.00	6.00
OneR	92.00	8.00	94.00	6.00
BFTree	94.66	5.33	95.33	4.66
DTable	92.66	7.33	94.00	6.00

The following figure is a graphical representation of the performance measures namely accuracy and error rate of different classification methods. The performance measures are represented, and according to this figure each classification algorithms performance was measured. Rule based classification namely OneR, Decision Table experimented, and these two algorithms provide good classification accuracy. OneR improves its accuracy from 92% into 94%, and also decision table algorithm improves its accuracy from 92.66% into 94%. Random Forest classification also improves its classification accuracy with the help of discretization.



	J48	RF	Rep	NB	RBF	OneR	BF Tree	Dtable
Accuracy	96	95.33	94	96	95.33	92	94.66	92.66
Error Rate	4	4.66	6	4	4.66	8	5.33	7.33
Disc+Acc	94	96.66	94.66	94	94	94	95.33	94
Disc+ER	6	3.33	5.33	6	6	6	4.66	6

Fig 1: Graphical representation of performance measures on various classification algorithms.

WEKA RESULTS

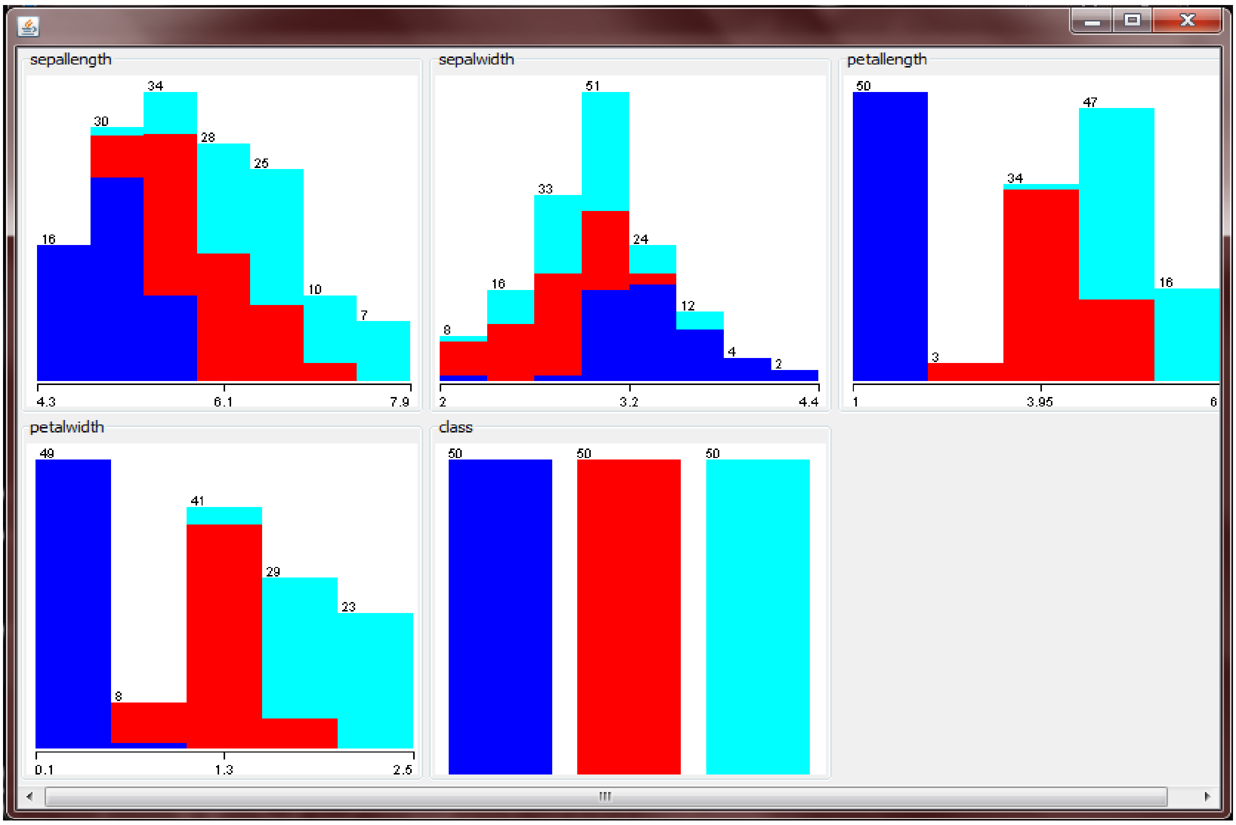


Fig 2: Before Discretization values of iris data set in WEKA Tool.

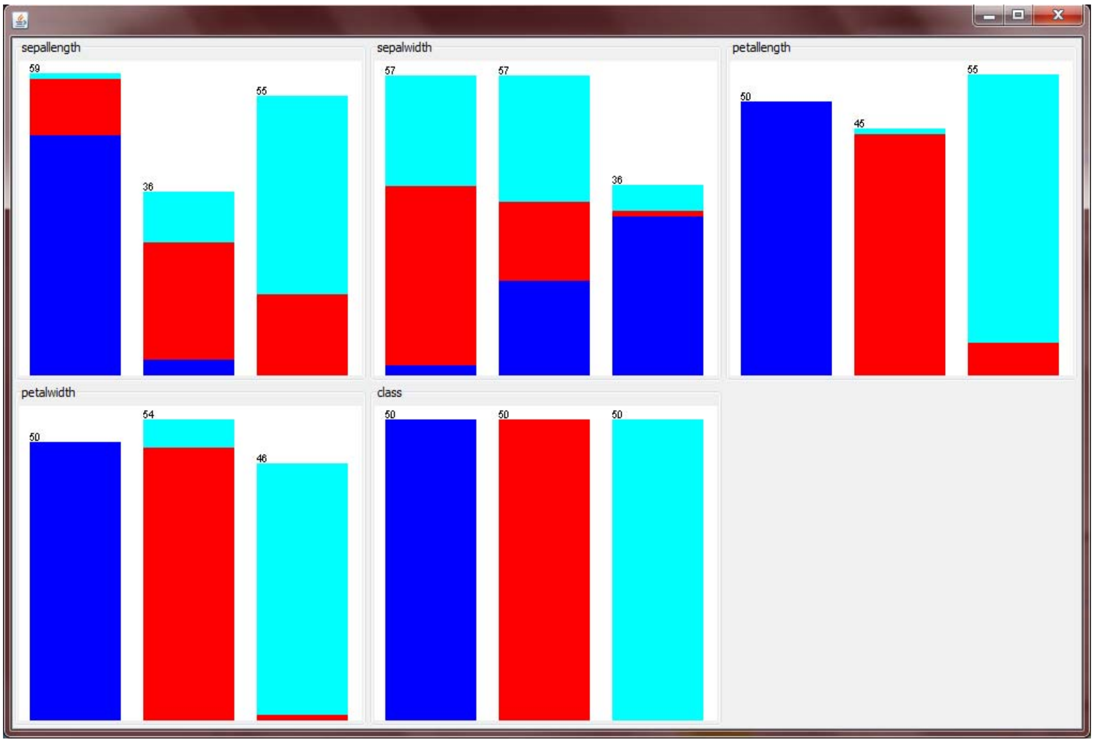


Fig 3: After Discretization values of iris data set in WEKA Tool.

5. Conclusion

In this paper, a study of classification algorithms such as Naïve-bayes, J48, Random Forest, RepTree, One R, Decision Table, and RBF Network has been experimented using iris data set. The classification algorithms compared, based on the performance measures namely accuracy and error rate using WEKA tool. Ten fold Cross validation testing used for the experiments. Results are shown in the table 6 and 7. From the results it is evident that OneR, and Decision Table are produce best classification accuracy compared to other studied classification algorithms. In future, research will be directed towards selection of different data sets, different behavioral patterns, and various algorithms.

References

[1] Aditi mahajan and Anita ganapati (2014), Performance evaluation of rule based classification algorithms, International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Vol. 3, Issue 10, pp. 5.

[2] Chitra nasa and Suman (2012), Evolution of different techniques forWEB Data, International Journal of Computer Application, Vol. 52, Issue 9, pp. 7.

[3] Data mining with WEKA tutorial classes (weka.waikato.ac.nz).

[4] Elsayad radwan and Adel M.A. Assiri (2013), Thyroid diagnosis based technique on rough sets with modified similarity relation, International Journal of Advanced Computer Science and Applications (IJACSA), Vol.4, Issue 10, pp. 7.

[5] Hemada and Vijayalakshmi (2013), Discretization technique using maximum frequent values and entropy criterion, International Journal of Advanced research in Computer Science and SoftwareEngineering (IJARCSSE), Vol. 3, Issue 11, pp. 7.

[6] Jiawei han and Micheline kamber, Data mining concepts and Techniques, second edition, pp. (76-96).

[7] Mangesh metkari and Madhavi pradhan (2015), Improve the classification accuracy of the heart disease data using discretization, International Journal of Innovative Research in Advanced Engineering (IJIRAE), Vol. 2, Issue 10, pp. 5.

[8] Nor liyana, Azuraliza and Zulaiha (2011), Performance study on data discretization technique using nutrition data set, International Symposium on Computing, Communication and Control (ISCCC), Vol. 1, Issue 9, pp. 5.

[9] Priyanka Sharma (2014), Comparative analysis of various decision tree classification algorithms using WEKA, International Journal on Recent and Innovation Trends in Computing and Communication, Vol.3, Issue 2, pp. 7.

[10] Sumit garg and Arvind Sharma (2013), Comparative analysis of data mining techniques of educational data set, International Journal of Computer Applications, Vol. 74, Issue 5, pp. 5.

[11] Tajun han, Sangbum lee and Sejong oh (2015), Improving discretization by post processing procedure, International Journal of Engineering and Technology (IJET), Vol. 7, Issue 2, pp. 5.

[12] Trilok chand Sharma and Manoj jain (2013), WEKA approach for comparative study of classification algorithm, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 4, pp. 7.

[13] Tutorial exercises for the WEKA Explorer, Data mining: Practical machine learning tools and techniques, chapter 17.

[14] UCI Machine learning Repository. (http://archive.ics.ui.edu/mi/)

[15] Vijayarani and Muthulakshmi (2013), Evaluating the efficiency of rule techniques for file classification, International Journal of Research in Engineering and Technology, Vol. 2, Issue 10, pp. 5.