Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5

Asha Gowda Karegowda
Siddaganga Institute of Technology. Tumkur, India.

Punya V
Siddaganga Institute of Technology , Tumkur, India.

M.A.Jayaram
Siddaganga Institute of Technology , Tumkur, India

A.S .Manjunath
Siddaganga Institute of Technology, Tumkur, India.

ABSTRACT

Medical Data mining is the process of extracting hidden patterns from medical data. This paper presents the development of a hybrid model for classifying Pima Indian diabetic database (PIDD). The model consists of two stages. In the first stage, the K-means clustering is used to identify and eliminate incorrectly classified instances. The continuous data is converted to categorical form by approximate width of the desired intervals, based on the opinion of medical expert. In the second stage a fine tuned classification is done using Decision tree C4.5 by taking the correctly clustered instance of first stage. Experimental results signify the cascaded K-means clustering and Decision tree C4.5 has enhanced classification accuracy of C4.5. Further rules generated using cascaded C4.5 tree with categorical data are less in numbers and easy to interpret compared to rules generated with C4.5 alone with continuous data. The proposed cascaded model with categorical data obtained the classification accuracy of 93.33 % when compared to accuracy of 73.62 % using C4.5 alone for PIMA Indian diabetic dataset.

General Terms

Medical data mining, clustering, rule based classification using decision tree C4.5

Keywords

K-means clustering, Categorical data, rule based classification, Decision tree C4.5, Pima Indian Diabetics.

1. INTRODUCTION

The data mining functionalities are used to specify the kind of patterns to be found in the data-mining task. The data mining functionalities mainly include association rule mining, classification, prediction & clustering. Association analysis is used for discovering interesting relations between variables in large databases, which in given in the form of rules to user. Classification predicts the class labels. Prediction is used to access the value of an attribute that a given sample is likely to have. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Classification is supervised learning algorithms in contrasts with clustering, which are unsupervised learning algorithm [1]. Classification is a supervised model, which maps or classifies a data item into one of several predefined classes. Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The most common classification data mining techniques are Case-Based Reasoning, Decision tree, Back propagation neural network, Radial basis neural network, Bayesian classification, Rough set Approach, Fuzzy Set Approaches, and K-nearest Neighbor classifiers. In this paper a cascaded K-means clustering and Decision tree has been used to categorize diabetic's patients. Literature survey of classification of diabetic data set is briefed in section 2. For the sake of completeness Decision tree and K-mean clustering have been briefly explained in section 3 and 4. Preprocessing of diabetic data set and working of cascaded K-means clustering and Decision tree classifier is explained in section 5, followed by results and conclusion in section 6 and 7 respectively.

2. RELATED WORK ON DIABETIC DATA SET CLASSIFICATION

2.1 Diabetics

Diabetes mellitus is a disease in which the body is unable to produce or unable to properly use and store glucose (a form of sugar). Glucose backs up in the bloodstream causing one's blood glucose or "sugar" to rise too high. There are two major types of diabetes. In type 1 diabetes, the body completely stops producing any insulin, a hormone that enables the body to use glucose found in foods for energy.

People with type 1 diabetes must take daily insulin injections to survive. This form of diabetes usually develops in children or young adults, but can occur at any age.

Type 2 (also called adult-onset or non insulin-dependent) diabetes results when the body doesn't produce enough insulin and/or is unable to use insulin properly (insulin resistance). This form of diabetes usually occurs in people who are over 40, overweight, and have a family history of diabetes, although today it is increasingly occurring in younger people, particularly adolescents[2], [3].

World Health Organization (WHO) report had shown a marked increase in the number of diabetics and this trend is expected to grow in the next couple of decades. In the International Diabetes Federation Conference 2003 held in Paris, India was labeled, as "Diabetes Capital of the World," as of about 190 million diabetics worldwide, more than 33 million are Indians. The worldwide figure is expected to rise to 330 million, 52 million of them Indians by 2025, largely due to population growth, ageing, urbanization, unhealthy eating habits and a sedentary lifestyle. Poorly managed diabetes can lead to a host of long-term complications among these are heart attacks, strokes, blindness, kidney failure, blood vessel disease.

2.2 Literature review of classification of diabetic dataset

A lot of research work has been done on various medical data sets including Pima Indian diabetes dataset. Classification accuracy achieved for Pima Indian diabetes dataset using 22 different classifiers is given in [4] and using 43 different classifiers is given in [5]. The performance of proposed cascaded model using k-means and Decision tree is is compared with [4] and [5]. The results of [5] and [4] are shown in Table 1 and Table 2 respectively. The accuracy of most of these classifiers is in the range of 66. 6% to 77.7%. Hybrid K-means and Decision tree [6] achieved the classification accuracy of 92.38% using 10 fold cross validations for continuous data. Further cascaded learning system based on Generalized Discriminate analysis (GDA) and Least Square Support Vector Machine (LS_SVM), showed accuracy of 82.05% for diagnosis of Pima dataset [7]. Further authors have achieved classification accuracy of 72.88 % using ANN, 78.21% using DT_ANN where decision tree C4.5 is used to identify relevant features and given as input to ANN [8], 79.50% using Cascaded GA_CFS_ANN, relevant feature identified by Genetic algorithm with Correlation based feature selection is given as input to ANN [9], 77.71% using GA optimized ANN, 84.10% using GA optimized ANN with relevant features identified by decision tree and 84.71% with GA optimized ANN with relevant features identified by GA_CFS[10]. Authors have achieved an accuracy of of 96.68% for diabetic dataset using cascaded k-means and K-nearest neighbor [11].

3. DECISION TREE C4.5

Decision trees represent a supervised approach to classification. A decision tree is a simple tree structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. The basic decision tree induction algorithm ID3 [12] was enhanced by C4.5[13, 14]. The WEKA classifier package has its own version of C4.5 known as J4.8. Information gain and gain ratio measures are used by ID3 and C4.5 as splitting criterion respectively. The summary of decision tree algorithm is given

i. Choose an attribute that best differentiates the output attribute values.
ii. Create a separate tree branch for each value of the chosen attribute.
iii. Divide the instances into subgroups so as to reflect the attribute values of the chosen node.
iv. For each subgroup, terminate the attribute selection process if:
(a). The members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value.
(b). The subgroup contains a single node or no further distinguishing attributes can be determined. As in (a), label the branch with the output value seen by the majority of remaining instances.
v. For each subgroup created in (iii) that has not been labeled as terminal, repeat the above process.

4. K-MEANS CLUSTERING

K-means [15] is one of the simplest unsupervised learning algorithms and follows partitioning method for clustering. K-means algorithm takes the input parameter, k as number of clusters and partitions a dataset of n objects into k clusters, so that the resulting objects of one cluster are dissimilar to that of other cluster and similar to objects of the same cluster. In k-means algorithms begins with randomly selected k objects, representing the k initial cluster center or mean. Next each object is assigned to one the cluster based on the closeness of the object with cluster center. To assign the object to the closest center, a proximity measure namely Euclidean distance is used that quantifies the notion of closest. After all the objects are distributed to k clusters, the new k cluster centers are found by taking the mean of objects of k clusters respectively. The process is repeated till there is no change in k cluster centers. K-means algorithm aims at minimizing an objective function namely sum of squared error (SSE). SSE is defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2 \qquad (1)$$

Where E is sum of the square error of objects with cluster means for k cluster. p is the object belong to a cluster Ci and mi is the mean of cluster Ci .The time complexity of K-means is O (t*k*n) where t is the number of iterations, k is number of clusters and n is the total number of records in dataset.

K-means partitioning algorithm: (Input is k is the number of clusters, D is input data set. Output is k clusters.)

i. Randomly choose k objects from D as the initial cluster centers.
ii. Repeat
iii. Assign each object from D to one of k clusters to which the object is most similar based on the mean value of the objects in the cluster.
iv. Update the cluster means by taking the mean value of the objects for each of k cluster.
v. Until no change in cluster means/ min error E is reached.

5. CASCADED K-MEANS AND DECISION TREE C4.5

5.1 Data preprocessing

The PIMA diabetic database consist of two categories in the data set (i.e. Tested positive , Tested Negative) each having 8 features :Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function and Age (years]. The data is availed from UCI Machine Learning Repository available at http://www.ics.uci.edu/~mlearn/MLRepository.html the data processing techniques, when applied prior to mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining. Data preprocessing is a significant step in the knowledge discovery process, since quality decisions must be based on quality data. A total of 768 cases are available in PIDD. 5 patients had a glucose of 0, 11 patients had a body mass index of 0, 28 others had a diastolic blood pressure of 0, 192 others had skin fold thickness readings of 0, and 140 others had serum insulin levels of 0. After deleting these cases there were 392 cases with no missing values (130 tested positive cases and 262 tested negative) [16].

5.2 Working of Proposed method

In the first stage of proposed model, simple K-means clustering (with k = 2) of Weka tool, is applied to 392 diabetic patients samples as obtained in section 5.1. The wrongly classified samples are eliminated to get final 299 samples. Further as a part of preprocessing the continuous data is converted to categorical form by approximate width of the desired intervals, based on the opinion of medical experts as shown in table 3. Finally in the second stage, the correctly classified samples from first stage and the categorical data is given as input to Decision tree C4.5 (weka J4.8). The data is partitioned using (a) 60-40 ratio partitioning method (training-test) and (b) 10 fold cross validation. For the sake of completeness few of the performance metrics have been discussed. True positive (TP) corresponds to the number of positive examples correctly predicted by the classifier. False negative (FN) corresponds to the number of positive examples wrongly predicted as negative by the classifier. False positive (FP) corresponds to the number of negative examples wrongly predicted as positive by the classifier. True negative (TN) corresponds to the number of negative examples correctly predicted by the classifier. The true positive rate (TP rate) or sensitivity is the fraction of positive examples predicted correctly by the model. TP Rate = TP/ (TP + FN). The false positive rate (FP rate) or Specificity the fraction of negative examples predicted as a positive class. FP Rate = FP/ (TN + FP).Precision is the fraction of records that actually turns out to be positive in the group the classifier has declared as a positive class. Precision = TP / (TP + FP). Recall is the fraction of positive examples correctly predicted by the classifier. Recall = TP / (TP + FN). F-measure is used to examine the tradeoff between recall and precision. Measure = 2*TP/ (2*TP+FP + FN).

6. RESULTS

In this paper an eager learner: Decision tree classifier (C4.5) has been used for classification of standard medical database namely PIMA diabetic. Incorrect labeled instance are eliminated using K-means clustering followed by converting the continuous data to categorical data by consulting medical experts. The resultant dataset is used to train and test the diabetic data set using two method (a) dividing training data and test data using 60-40 ratio (b) 10 fold cross validation method. The performance of Decision tree with unprocessed data compared to proposed cascaded k-means with Decision tree using processed data by means of TP rate, FP rate, Precision, Recall and F-measure is also computed for tested positive and tested negative class is shown in table 4. For unprocessed diabetic data, the classification accuracy of Decision tree using 10 fold cross validation and 60-40 training–testing partitioning of data was found to be almost same. However for the proposed cascaded model with categorical data, the performance of Decision tree with 60-40 training-testing partitioning of data outperforms the decision tree with 10 fold cross validation by an order of 1.38 %. Kappa value , mean absolute value and classification accuracy for unprocessed and processed data using C4.5 using 60-40 ratio (partitioning of training and test data) and 10 fold cross validations is shown in figure 1 and 2 respectively. Experimental results show the improvement in accuracy Diabetic data set using proposed cascaded method: k-means with Decision tree (with categorical data) by an order of 19.50% of classification compared to Decision tree C4.5 alone with unprocessed data.

Further the rules generated by decision tree with categorical data are less in number and are easy to interpret compared to rules generated by decision tree with continuous data. The cascaded model with categorical data resulted in only 9 rules (figure 4) compared to 20 rules (figure 3) by decision tree alone with unprocessed numerical data. Kappa statistics K value for proposed cascaded model (k-means + classifier with feature selection) is equal to 0.85, (which is almost near to 1) clearly illustrates the match between the proposed cascaded classifier and and the real world output. The proposed model achieved the sensitivity and specificity for tested positive class as 0.89 and 0.48 compared to 0.695 and 0.24 with Decision tree alone using unprocessed data. Further for the tested positive class, the tradeoff between recall and precision for the proposed model is found to be 0.89 compared to 0.64 using Decision tree alone. The rules generated by the proposed cascaded model are given below.

1. If Plasma=low then class=> Tested Negative
2. If Plasma =medium & Age=low & Pedigree =low then
   Class => Tested Negative
3. If Plasma =medium & Age=low & Pedigree =medium
   & Diastolic BP=medium then Class=> Tested Negative
4. If Plasma=medium & Age=low & Pedigree =medium
   & Diastolic BP =low then Class=> Tested Negative
5. If Plasma=medium & Age=low & Pedigree =medium
   &Diastolic BP =high then Class=> Tested Positive
6. If Plasma =medium & age=high then
   Class => Tested Positive
7. If Plasma =medium & Age=low & Pedigree =high
   then Class=> Tested Positive
8. If Plasma =medium & Age=medium then
   Class=> Tested Positive
9. If Plasma =high then Class=> Tested Positive

7. CONCLUSIONS

The performance of classification algorithm depends on the quality of data. The K-means clustering is used to identify and eliminate incorrectly classified instances. Further the continuous data is converted to categorical data by consulting medical expert's advice. The correctly classified instance by k-means is used as input to Decision tree after conversion of continuous data to categorical data. The proposed cascaded shows improved classification of 93.33% for PIMA diabetic dataset using 60-40 % training–testing partitioning method with preprocessed data. Further results showed that the performance of cascaded model with categorical data generated comparatively less number of rules which are easy to interpret compared to rules generated by Decision tree with unprocessed data. The classification accuracies obtained by the proposed cascaded K_means clustering and Decision tree classifier is one of the best results compared with the results of Decision tree reported in the literature.

8. REFERENCES

[1] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, San Francisco, Morgan Kauffmann Publishers, (2001)

[2] Editorial, Diagnosis and Classification of Diabetes Mellitus, American Diabetes Association, Diabetes Care, vol 27, Supplement 1, (Jan 2004).

[3] The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus: Follow up report on

the Diagnosis of Diabetes Mellitus. Diabetic Care 26, pp.3160- 3167, (2003).

[4] Michie, D., Spiegelhalter, D. J., & Taylor, C. C., Machine learning, neural and statistical classification. Ellis Horwood ,(1994).

[5] Humar, K., & Novruz, A. Design of a hybrid system for the diabetes and heart diseases. Expert Systems with Applications, 35, 82–89 ,(2008).

[6] B.M Patil, R.C Joshi, Durga Tosniwal, Hybrid Prediction model for Type-2 Diabetic Patients, Expert System with Applications, 37, 8102-8108 (2010).

[7] Polat, K., Gunes, S., & Aslan, A., A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Systems with Applications, 34(1), 214–221(2008) .

[8] Asha Gowda Karegowda, MA.Jayaram, Integrating Decision Tree and ANN for Categorization of Diabetics Data , International Conference on Computer Aided Engineering, December 13-15, IIT Madras, Chennai, India (2007).

[9] Asha Gowda Karegowda and M.A. Jayaram, Cascading GA & CFS for Feature Subset Selection in Medical Data Mining , International Conference on IEEE International Advance Computing Conference (IACC'09), Thapar University, Patiala, Punjab India (Mar 2009).

[10] Asha Gowda Karegowda, A.S. Manjunath, M.A. Jayaram Application Of Genetic Algorithm Optimized Neural Network Connection Weights For Medical Diagnosis Of Pima Indians Diabetes, International Journal on Soft Computing (IJSC), Vol.2, No.2.(May 2011).

[11] Asha Gowda Karegowda , M.A. Jayaram, A.S. Manjunath ,Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients , International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-3, (Feb 2012).

[12] J.R. Quinlan, Induction of Decision Trees, Machine Learning 1: pp.81-106, Kluwer Academic Publishers, Boston, (1986).

[13] J.R. Quinlan, San Mateo, C4.5 Programs for Machine Learning: Morgan Kaufmann, (1993).

[14] J.R. Quinlan, Bagging, Boosting and C4.5, In Proc. 13th National Conf. Artificial Intelligence (AAAI'96), pp. 725-730. Portland, (Aug, 1996).

[15] MacQueen, J.B., Some Methods for classification and analysis of multivariate observations. In Proceedings of 5th Berkeley symposium on mathematical statistics and probability (pp. 281–297). Berkeley: University of California Press (1967)..

[16] JJoseph L.Breault, Data Mining Diabetic Databases: Are rough Sets a Useful Addition?, http://www.galaxy.gmu.edu/interface/I01/I2001Proceedings/Jbreault

Table 1. The Classification accuracy proposed model compared with different machine learning methods on Pima Indian diabetes dataset (Humar &Novruz,2008)

Method	Accuracy %	Reference
Proposed Model k-means+DT continuous data)	93.33 %	This paper
K-means + KNN,k=5	96.68	Asha et.al[11]
Log disc	77.7	Stat log
IncNet	77.6	Norbert Jankowski
DIPOL92	77.6	Stat log
Linear Discr. Anal.	77.5–77.2	Stat log; Ster & Dobnikar
SMART	76.8	Stat log
GTO DT (5 · CV)	76.8	Bennet and Blue
kNN, k = 23, Manh, raw, W	76.7 ± 4.0	WD-GM, feature weighting 3CV
kNN, k = 1:25, Manh, raw	76.6 ± 3.4	WD-GM, most cases k = 23
ASI	76.6	Ster & Dobnikar
Fisher discr. analysis	76.5	Ster & Dobnikar
MLP + BP	76.4	Ster & Dobnikar
MLP + BP	75.8 ± 6.2	Zarndt
LVQ	75.8	Ster & Dobnikar
LFC	75.8	Ster & Dobnikar
RBF	75.7	Stat log
NB	75.5–73.8	Ster & Dobnikar; Stat log
kNN, k = 22, Manh	75.5	Karol Grudzin´ ski
MML	75.5 ± 6.3	Zarndt
SNB	75.4	Ster & Dobnikar
BP	75.2	Statlog
SSV DT	75.0 ± 3.6	WD-GM, SSV BS, node 5CV MC
kNN, k = 18, Euclid, raw	74.8 ± 4.8	WD-GM
CART DT	74.7 ± 5.4	Zarndt
CART DT	74.5	Stalog
DB-CART	74.4	Shang & Breiman
ASR	74.3	Ster & Dobnikar
SSV DT	73.7 ± 4.7	WD-GM, SSV BS, node 10CV strat
C4.5 DT	73.0	Stalog
C4.5 DT	72.7 ± 6.6	Zarndt
Bayes	72.2 ± 6.9	Zarndt
C4.5 (5 · CV)	72.0	Bennet and Blue
CART	72.8	Ster & Dobnikar
Kohonen	72.7	Statlog
kNN	71.9	Ster & Dobnikar
ID3	71.7 ± 6.6	Zarndt
IB3	71.7 ± 5.0	Zarndt
IB1	70.4 ± 6.2	Zarndt
kNN, k = 1, Euclides, raw	69.4 ± 4.4	WD-GM
kNN	67.6	Statlog
C4.5 rules	67.0 ± 2.9	Zarndt
OCN2	65.1 ± 1.1	Zarndt
QDA	59.5	Ster, Dobnikar

Table2. The Classification accuracy of different machine learning methods on PIMA Indian diabetes dataset (Michie et al.,1994)

Machine learning method	Accuracy%
Discrim	77.5
Quaddisc	73.8
Logdisc	77.7
SMART	76.8
ALLOC80	69.9
k-NN	67.6
CASTLE	74.2
CART	74.5
IndCART	72.9
NewID	71.1
AC2	72.4
Baytree	72.9
NaiveBay	73.8
CN2	71.1
C4.5	73
Itrule	75.5
Cal5	75
Kohonen	72.2
DIPOL92	77.6
Backprop	75.2
RBF	75.7
LVQ	72.8

Table 3. Categorical Partitioning of PIMA diabetic data set attributes based on Medical Experts opinion.

Attributes name	Partitioned data

Number of times of pregnancy
--- (# Preg)
	{low, medium, high}
	{ <3, 3-5 ,>6}

Plasma glucose concentration every 2 hours in an oral glucose tolerance test --- (Plasma)	{low, medium, high }
	{< 95 , 95-150 , >150}

diastolic blood pressure (mm Hg)
---- (Diast BP)
	{low, normal, high }
	{ <70, 70-100, >100}

triceps skin fold thickness (mm)
---- (skin)
	{low, medium, high }
	{ <21, 21-40, > 40}

2-Hour serum insulin (mu U/ml)
---- (insulin)
	{normal, medium, high}
	{<140, 140-200, > 200}

body mass index (weight in kg/(height in (mm)2) ---- BMI	{normal, obese, overweight }
	{<23 , 23-29, > 29}

diabetes pedigree function
--- Pedigree
	{low, medium, high }
	{ <0.4, 0.4- 0.8, > 0.8}

Age in years ---- Age	{young, middle aged, senior } { 20-34, 35-46, >46}
Class	{ Tested Positive, Tested Negative}

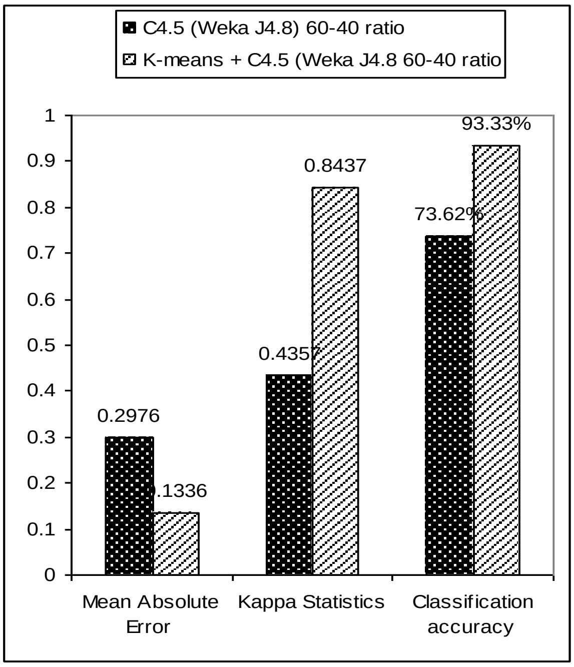

Figure 1. The Classification accuracy, kappa and mean absolute value of c4.5 alone vs proposed method for PIMA Indian diabetes dataset using 60-40 partitioning of data into training and test data

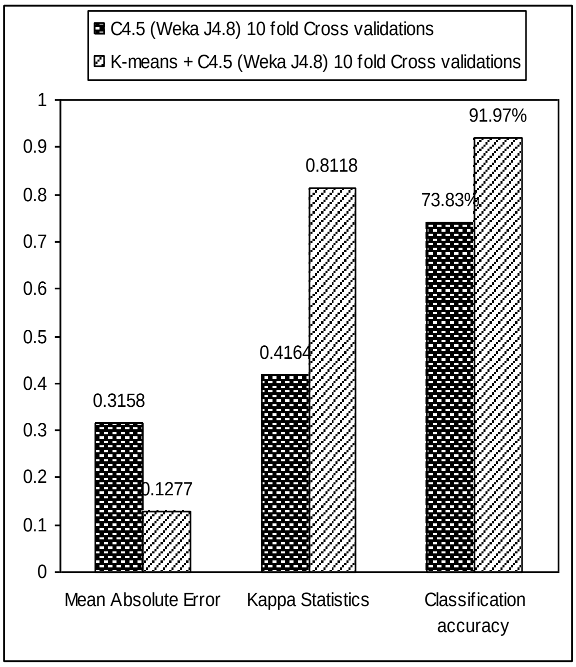

Figure 2. The Classification accuracy, kappa and mean absolute value of c4.5 alone vs proposed method for PIMA Indian diabetes dataset using 10 fold cross validation method.

Table 5.Comparing various performance measures of C4.5 alone vs proposed model for Pima Indian diabetes

Classification Algorithm	Partitioning method for Data	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Decision tree C4..5 (Weka J4.8)	60-40 ratio	Tested _Positive	0.695	0.243	0.598	0.695	0.643
		Tested Negative	0.757	0.305	0.827	0.757	0.791
	10 fold Cross validations	Tested Positive	0.597	0.186	0.632	0.597	0.614
		Tested Negative	0.814	0.403	0.79	0.814	0.802
Proposed method : K-means + C4..5 (Weka J4.8)	60-40 ratio	Tested Positive	0.892	0.048	0.892	0.892	0.892
		Tested Negative	0.952	0.108	0.952	0.952	0.952
	10 fold Cross validation	Tested Positive	0.833	0.039	0.909	0.833	0.87
		Tested Negative	0.961	0.167	0.924	0.961	0.942

```
Plasma <= 127
|  BMI <= 26.4: Tested Negative (132.0/3.0)
|  BMI > 26.4
|  |  Age <= 28: Tested Negative (180.0/22.0)
|  |  Age > 28
|  |  |  Plasma <= 99: Tested Negative (55.0/10.0)
|  |  |  Plasma > 99
|  |  |  |  pedigree <= 0.561: Tested Negative (84.0/34.0)
|  |  |  |  pedigree > 0.561
|  |  |  |  |  preggree <= 6
|  |  |  |  |  |  Age<= 30: Tested Positive (4.0)
|  |  |  |  |  |  Age > 30
|  |  |  |  |  |  |  Age <= 34: Tested Negative (7.0/1.0)
|  |  |  |  |  |  |  Age > 34
|  |  |  |  |  |  |  |  BMI <= 33.1: Tested Positive (6.0)
|  |  |  |  |  |  |  |  BMI > 33.1: Tested Negative (4.0/1.0)
|  |  |  |  |  #preg > 6: Tested Positive (13.0)
Plasma > 127
|  BMI <= 29.9
|  |  Plasma <= 145: Tested Negative (41.0/6.0)
|  |  Plasma > 145
|  |  |  Age <= 25: Tested Negative (4.0)
|  |  |  Age > 25
|  |  |  |  Age <= 61
|  |  |  |  |  BMI <= 27.1: Tested Positive (12.0/1.0)
|  |  |  |  |  BMI > 27.1
|  |  |  |  |  |  Diast_BP <= 82
|  |  |  |  |  |  |  pedigree <= 0.396: Tested Positive (8.0/1.0)
|  |  |  |  |  |  |  pedigree > 0.396: Tested Negative (3.0)
|  |  |  |  |  |  Diast_BP > 82: Tested Negative (4.0)
|  |  |  |  Age > 61: Tested Negative (4.0)
|  BMI > 29.9
|  |  Plasma <= 157
|  |  |  Diast_BP <= 61: Tested Positive (15.0/1.0)
|  |  |  Diast_BP  > 61
|  |  |  |  Age <= 30: Tested Negative (40.0/13.0)
|  |  |  |  Age > 30: Tested Positive (60.0/17.0)
|  |  Plasma > 157: Tested Positive (92.0/12.0)
```

Figure 3. Decision tree C4.5 (WEKA J4.8) for unprocessed PIMA on Pima Indian diabetes (continuous

```
Plasma = low: Tested Negative (115.0/2.0)
Plasma = medium
|  Age = low
|  |  pedigree = low: Tested Negative (89.0/10.0)
|  |  pedigree = high: Tested Positive (1.0)
|  |  pedigree = medium
|  |  |  Diast_BP = medium: Tested Negative (6.0/1.0)
|  |  |  Diast_BP = low: Tested Negative (0.0)
|  |  |  Diast_BP = high: Tested Positive (2.0)
|  Age = high: Tested Positive (11.0/1.0)
|  Age = medium: Tested Positive (17.0/3.0)
Plasma = high: Tested Positive (58.0/2.0)
```

Figure 4. Decision tree C4.5 (WEKA J48) for preprocessed Pima Indian diabetes (categorical data).