International Journal of Computer Applications (0975 - 8887)
Volume 139 - No.11, April 2016

Cascaded Modeling for PIMA Indian Diabetes Data

M.S. Barale
Deapartment of Statistics,
Shivaji University, Kolhapur,
INDIA 416004
baralemahesh12@gmail.com

D.T. Shirke
Deapartment of Statistics,
Shivaji University, Kolhapur,
INDIA 416004
dts_stats@unishivaji.ac.in

ABSTRACT

This paper develops the cascaded models for classification of PIMA Indian diabetes database. The k-nearest neighbour method is used to impute the missing data and the processed data is used for further classification. This is done in two steps, in first step k-means clustering algorithm is used for extracting hidden patterns in data set then in second step the classification is done by using suitable classifier. k-means algorithm combined with artificial neural network classifier and k-means algorithm combined with logistic regression classifier achieve classification accuracy above 98%.

Keywords

Missing data, Clustering, Classification

1. INTRODUCTION

Data mining refers to extracting knowledge from large amount of data. It includes tools like clustering, classification, prediction and association analysis. It is important to have quality data for the best possible performance of any of the data mining algorithm. The quality of data in the sense that presence of missing instances and outliers, results misleading outcome. So the treatment on missing cases and outlier is required. There are many methods for imputing the missing cases such as regression, k-nearest neighbour method for missing data imputation. PIMA Indian diabetic (PID) data is considered, where problem under consideration is to develop a model for predicting a class label. A brief review of the literature related to this problem is reported in section 2. Pre-processing of PID data is described in section 3 with k-nearest neighbour method of data imputation. Data mining methods which are used for classification are discussed in section 4. For the sake of completeness procedure for predicting a class label described in section 5. Results and conclusions are reported in sections 6 and 7 respectively.

2. RELATED WORK ON PID DATA AND CLASSIFICATION

2.1 Diabetics

Diabetes mellitus is a chronic disease that occurs when the pancreas is no longer able to make insulin, or when the body cannot make good use of the insulin it produces. Insulin is a hormone made by the pancreas, that acts like a key to let glucose from the food one eats pass from the blood stream into the cells in the body to produce energy. All carbohydrates foods are broken down into glucose in the blood. Insulin helps glucose get into the cells. There are two major forms of diabetes. Type 1 debates is characterized by a lack of insulin production and Type 2 results from the body's ineffective use of insulin. World health organization (WHO) reported total deaths from diabetes are projected to rise by more than 50% in the next 10 years. Type 2 diabetes is much more common than Type 1 diabetes. In 2012 diabetes was the direct cause of 1.5 million deaths and 80% of diabetes deaths occur in low and middle income countries as per WHO (2015) report.

2.2 Literature review of classification of Diabetic dataset

The PID database availed from UCI Machine Learning Repository available at *http://www.ics.uci.edu/ $\sim$ mlearn/MLRepository.html* consists of two categories namely tested positive and tested negative. It has 8 features as : number of times pregnant, plasma glucose concentration at 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).
A lot of research work has been done on the PID dataset, with a problem of developing a model for predicting the class label. The classification accuracy of various 65 classifiers were discussed by Karegowda et. al (2012) and also developed model named as cascaded k-means and decision tree, which has classification accuracy 93.33% for categorized and no missing data. Also performance of cascaded model using k-means and k nearest neighbour for continuous and no missing data is compared with the other classifiers discussed earlier in Karegowda et. al (2012) and which has the accuracy 96.68%. The hybrid prediction model proposed by Patil et. al (2010) achieved 92.38% accuracy. The cascaded Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM) developed by Polat et. al (2006) achieved 82.05% accuracy for diabetic dataset. Karegowda et. al proposed models ANN and DT_ANN with accuracy 72.88% and 78.21% respectively (2007), further they propose cascaded GA_CFS_ANN model with accuracy79.50% (2009). The accuracy of the other classifiers discussed in Table [2] and it is ranging from 77.2% to 77.7%.

Table 1. Percentage of missing

Sr. no.	Attribute name	Missing values	Missing percentage
1	2 hours serum insulin (muU/ml)	374	49
2	Triceps skin fold thickness (mm)	227	7 30
3	Diastolic blood Pressure	35	5
4	Body mass index (kg/ $(mm)^2$)	10	1
5	Plasma glucose concentration (mm Hg)	5	1

3. DATA PRE-PROCESSING

3.1 Missing Values Imputation

One relevant problem in data quality is the presence of missing data. Missing data may have different sources such as death of patients, equipment malfunctions, refusal of respondents to answer certain questions, and so on. The impact of missing data on various data mining tools are well explained by Brown and Kros (2003). Data pre-processing is an important step in the data mining process. Here in overall 51% cases have missing values of one or more than one attributes. Percentage missing values of various attributes are given in Table 1. Karegowda et. al (2012) have discarded all the cases having one or more missing attribute values from 768 cases. They have left with 392 cases for modelling.
There are 7 cases has four attribute values missing, 26 cases have three attribute values missing and 201 cases have two attribute value missing. After deleting these cases there are 142 cases with one attribute value missing and 392 cases with no missing. These 142 missing values imputed using method of data imputation k-nearest neighbour (k=3). Therefore 534 observations are useful for further analysis (178 *tested positive* cases and 369 *tested negative*). The following method is used for data imputation.
In k-nearest neighbour method, the missing values of an instance are imputed considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a distance function. The algorithm is as follows:

(1) Divide the data set D into two parts. Let D_m be the set containing the instances in which at least one of the features is missing. The remaining instances will complete feature information form a set called D_c.
(2) For each vector x in D_m: divide the instance vector into observed and missing parts as x=$[x_0; x_n]$, where x_0 is observed part while x_n is the missing attribute(s) part.
(3) Calculate the distance between the x_0 and all the instance vectors from the set D_c. Use only those features in the instance vectors from the complete set D_c, which are observed in the vector x.
(4) Use the k closest instances vectors (k-nearest neighbours) and perform a majority voting estimate of the missing values for categorical attributes. For continuous attributes replace the missing value using the mean value of the attribute in the k nearest neighbour. The median could be used instead of the mean.

3.2 Outlier Detection and Treatment

Presence of outlier has considerable effect on the accuracy of prediction model. Outliers can be detected in the PID dataset using box plot. The attribute serum insulin has the large number of outliers therefore the corresponding data 36 cases are eliminated from data set leaving with 498 cases for the modelling. In this data set 341 cases have label *tested negative*, while 157 cases have label *tested positive*.

4. DATA MINING METHODS

k-means algorithm is used to extracting hidden patterns in the data. k-means algorithm is an unsupervised learning technique, which is used for clustering the datasets into k; a predetermined number of clusters such that objects of one cluster are dissimilar to that of other cluster and similar to objects of the same cluster. One may refer to Han and Kamber (2012) for the details. The extracted data used to building classifier.
Artificial neural network(ANN), logistic regression(LR), support vector machine(SVM) and decision tree(DT) algorithms are supervised learning techniques, which is applied to a data with class labels. These classifiers predict class labels based on corresponding independent attributes. The LR model is well described by Alan Agresti (2007). For a good discussion of neural networks and support vector machine used as statistical tools, see Han and Kamber (2012).

5. WORKING PROCEDURE

In first stage of modelling, simple k-means clustering algorithm with(k=2) is applied to 498 sample cases obtained from data preprocessing for extracting hidden patterns. The misclassified sample cases are removed from data to get final 349 samples (237 *tested positive* cases and 112 *tested negative*). In second stage correctly classified samples from first stage given as input to the LR, ANN, SVM and DT. The dataset is randomly partitioned in two sets with 70% training and 30% testing and 10 fold cross validation also done. To check the adequacy of classifiers performance measures need to be taken into account. True positives (TP) refers to the positive cases that were correctly labelled by the classifier, while true negatives (TN) are the negative cases that were correctly labelled by the classifier. False positives (FP) are the negative cases that were incorrectly labelled, while false negatives (FN) are the positive cases that were incorrectly labelled. Some evaluation measures of classifiers are as follows,

(1) Accuracy: The accuracy of the classifier on a given test data set is the percentage of test cases that are correctly classified by classifier. *Accuracy=(TN)/(TP+TN+FP+FN)*.
(2) Specificity: Specificity is the true negative rate ie. Proportion of negative cases that are correctly identified.
Specificity= (TN)/(TN+FP).
(3) Sensitivity: Sensitivity is the true positive rate ie. Proportion of positive cases that are correctly identified.
Sensitivity = (TP)/(TP+FN).

6. RESULTS

In this article the classifiers LR, ANN, SVM and DT are used for classification of the imputed PIMA Indian Diabetes database. After k- means clustering the mislabelled instances are removed from database and remaining instances are used as input to the classifiers ANN, LR and DT. Further data divided in to training and testing dataset using 70-30 ratio. The performance of the proposed classifier is compared with cascaded k-means with DT and cascaded k-means with KNN, k=5 by using evaluation measures sensitivity, specificity and accuracy of the classifiers. As

compared to these two models the proposed models are superior. The sensitivity, specificity and accuracy are given in Table 3. The comparison with relative classifiers which has accuracy greater than 77.2 are shown in Table 2. and Rule generated by using k-means+DT is given in Figure 1.

7. CONCLUSIONS

The classification performance depends on the quality of the data. A data pre- processing is done properly then accuracy of the classifier may get increased. Data is to be used for classification is selected by clustering algorithm where cases which are correctly grouped are only considered for classification. The developed models cascaded k-means combined with LR and k-means combined with ANN give the accuracy above 98% which is improvement as compared to cascaded models in (2012). The model k-means and SVM achieved the accuracy 95.31%. The processed(imputed) data is also analyzed using ANN, SVM and LR. The classification accuracy of these classifiers are given in Table[2]. As compared with accuracy the proposed model cascaded k-means and LR is best among the methods considered here.
The performance of reported classifiers can be improved by using ensemble methods which can be taken as future work.

Acknowledgement

Both the authors are thankful to University Grants Commission, New Delhi for providing financial assistance to carry out the research work under Special Assistance Programme (F.520/8/DRS-I/2016(SAP-I)).

Table 2. The classification accuracy comparison of proposed model with other models

Method	Accuracy in %	Reference
k-means+LR *	99.33	This paper
k-means+ANN(MFP) *	98.57	This paper
k-means+DT *	97.99	This paper
k-means+SVM *	97.13	This paper
data K-means+ANN (RBF)*	95.70	This paper
k-means+KNN,k=5	96.68	Karegowda et al.
k-means+DT continuous data	93.33	Karegowda et. al.
Hybrid Prediction Model(HPM)	92.38	B.M. Patil et. al
GDA and LS-SVM	82.05	Kemal Polat et. al
SVM *	79.32	This paper
DT *	76.10	This paper
ANN *	75.10	This paper
ANN *	75.10	This paper
IncNet	77.6	Norbert Jankowski
DIPOL92	77.6	Stat log
Linear Discr. Anal	77.5-77.2	Stat log; Ster & Dobnikar

where * denotes use of imputed dataset for classification.

Table 3. The classification accuracy comparison of proposed model with other models

Method	Partitioning method	Specificity	Sensitivity	Accuracy
k-means+LR *	70-30 ratio	1	0.991	99.33
k-means+ANN(MFP)	10 fold	0.992	0.974	98.57
k-means+DT *	10 fold	0.983	0.973	97.99
k-means+SVM *	10 fold	0.983	0.947	97.13
k-means+ANN (RBF)*	10 fold	0.97	0.929	95.70
SVM *	10 fold	0.745	0.807	79.32

where * denotes use of imputed dataset for classification.

8. REFERENCES

[1] Alan Agresti Department of Statistics University of Florida Gainesville, Florida, An Introduction to Categorical Data Analysis 2^{nd} Edition, (2007).

[2] A. G. Karegowda, M. A. 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).

[3] A. G. 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).

[4] A. G. Karegowda, Punya V., M.A. Jayaram and 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).

[5] A. G. Karegowda, Punya V., M.A. Jayaram and A.S. Manjunath, Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5, International Journal of Computer Applications ISSN: 0975 – 8887, Volume 45, (May 2012).

[6] B. M. Patil , R.C. Joshi, Durga Toshniwal, Hybrid prediction model for Type-2 diabetic patients, Expert Systems with Applications, Volume 37 ISS: 8102–8108, (2010).

[7] Gustavo E. A. P. A. Batista and Maria Carolina Monard, University of Sao Paulo, A Study of k- Nearest Neighbour as an Imputation Method.

[8] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, San Francisco, Morgan Kauffmann Publishers, 3^{rd} edition, (2012).

[9] Kayaer, K., & Yildirim, T., Medical diagnosis on pima Indian diabetes using general regression neural networks, artificial neural networks and neural information processing (pp. 181–184), Istanbul, Turkey, (2003).

[10] Kemal Polat, Salih Gunes and Ahmet Arslan, A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine, Expert Systems with Applications, Volume 34 ISS: 482–487, (Jan 2008).

[11] Marvin L. Brown and John F. Kros, Data Mining and the Impact of Missing Data, Industrial Management & Data Systems, Volume 103, ISS: 611–621, (2003).

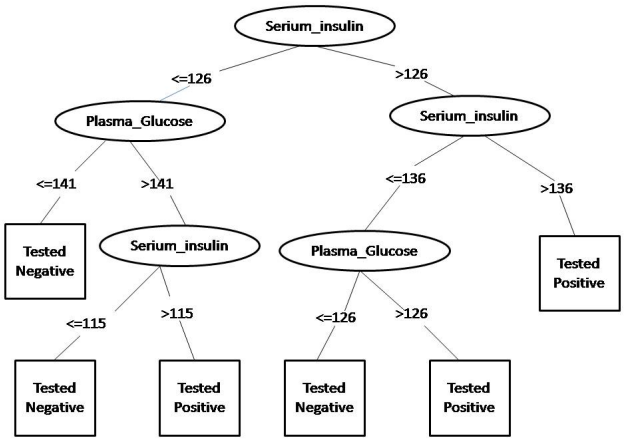

Fig. 1 Decision Tree (Weka J48)

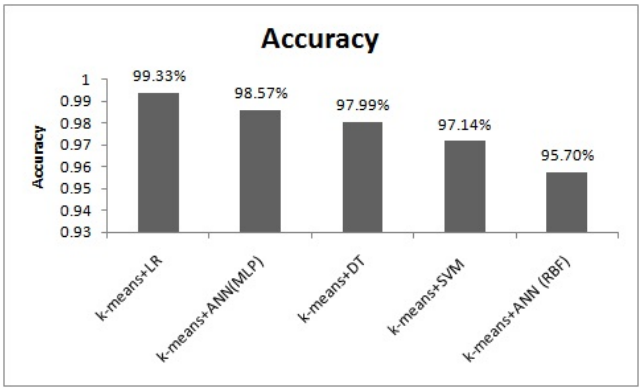

Fig. 2 The Classification accuracy using proposed models for PIMA Indian data

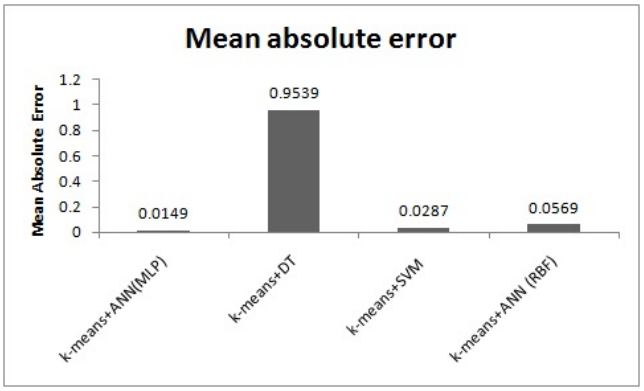

Fig. 4 Mean absolute error values of proposed models for PIMA Indian data 10 fold cross validation.

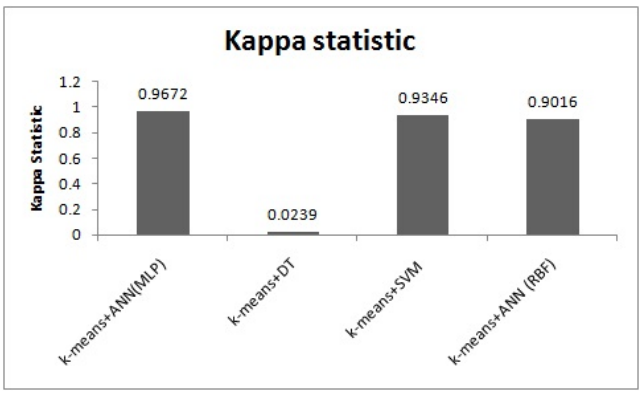

Fig. 3 Values of kappa statistic of proposed models for PIMA Indian data using 10 fold cross validation.