ISSN (Print): 0974-6846
ISSN (Online): 0974-5645
Indian Journal of Science and Technology, *Vol 12(34), DOI: 10.17485/ijst/2019/v12i34/145506, September 2019*

Knowledge Discovery and Sense Making for Early Diagnosis of Diabetes: A Hybrid Model combining LFD and Forest PA

Kiran Gurung[1], Abeer Alsadoon[1]*, Chandana Withana[1], Angelika Maag[1] and Amr Elchouemi[2]

[1]Study Group Australia, Department of IT, Sydney Campus, Australia; chandanapw@yahoo.co.uk, aalsadoon@studygroup.com, cwithana@studygroup.com, monem.rahma@yahoo.com
[2]Department of IT, Colorado State University, Gobal campus, United States; amr.elchouemi@hpe.com

Abstract

Objective: Even though there are many advanced and sophisticated data mining techniques that are being used to enhance the quality of health services, many of them still fail to produce accurate predictions when applied to real-time health datasets. Therefore, our aim is to devise the most accurate prediction model for an early diagnosis of diseases based on previously recorded patient data without performing any extensive laboratory tests. **Methods/Statistical Analysis and Finding:** Datasets are discretized by converting numerical attributes into categorical attributes. To such datasets, a decision forest algorithm is applied to produce a diverse group of classifiers. The algorithm eliminates the appearance of identical attributes in subsequent trees. Evaluation of a proposed hybrid model showed that the new technique had successfully improved the accuracy of classification and prediction by 9~10%. The accuracy is increased by reducing the information loss using a low-frequency discretization technique and by enhancing classification capabilities by generating a diverse group of classifiers. **Application/Improvement:** The proposed hybrid model combines two advanced techniques Low-frequency Discretization to reduce information loss during attribute discretization and to increase diversity and Forest PA to increase the accuracy of classifiers

Keywords: Data Mining, Data Visualization, Early Diagnosis of Diabetes, Forest PA, Knowledge Discovery, LFD, Sophisticated Data Mining Techniques

1. Introduction

Data mining is a part of artificial intelligence that is used to discover hidden sequences, patterns and information through examining relationships between features of data in large datasets, that otherwise could be very hard to extract[1]. Today, data mining is being used in various sectors including healthcare, medicine, for forecasting, transportation, government and others[2]. It has become the ultimate weapon to deliver advanced intelligence and solutions to every sector of business and government. Classification is one of the commonly used data mining techniques in the medical and health sectors[1].

Classification of data can be achieved using decision trees, a method which has been found useful in various business sectors including medical areas mainly as a prediction model which has drastically increased the use of prediction applications[1]. Our focus is on devising and studying the results of prediction models by combining the advantages of LFD with a decision forest algorithm with penalizing attributes. Forest PA basically discourages the same attributes to appear repeatedly in subsequent trees[3,4] producing more diverse and robust trees and increasing the accuracy of the prediction model.

The remainder of the paper is organized as follows: In Section 2, we identify and discuss the features of some of the existing classification and prediction models in a literature review. In Section 3, we outline the proposed solution. Section 4 introduces implementation and results discussion. This is followed by concluding comments and a discussion of future research.

*Author for correspondence

In[5] introduced a new technique of numerical attribute discretization by categorizing the value of numerical attributes based on low-frequency and attribute interdependency.

Low-frequency discretization techniques reduce information loss which accounts for its effectiveness and efficiency when compared to other existing discretization techniques. Furthermore, it removes the need for user input of the number of intervals and frequencies in each interval.

In[6] worked on improving an existing algorithm DBSCAN (Density Based Spatial Clustering of Application of Noise) to IDBSCAN (Improved Density Based Spatial Clustering of Application of Noise). The main purpose of this solution was to identify spatial data from other kinds of data using a density clustering technique. Moreover, the clustering technique implemented in this solution did not require any predefined number of clusters, making it more like data driven clustering. In[7] devised an improved version of the k-means clustering technique which significantly reduces processing time by reducing the number of cycles required to examine the dataset for the need to adjust the centroids of clusters. It reduced the execution time for dealing with complex and voluminous data generators such as health monitoring system and sensors. In[8] investigated and worked on developing a model which can identify recessive conjecturable rules for a single cluster. The improved algorithm identifies as many conditions as possible to define a cluster. However, although it can detect recessive conjecturable rules, it failed to provide details of how the dependency between two possible conjecturable rules affects accuracy.

In[9] successfully increased generalization accuracy by introducing improvements in splitting criteria for continuous attributes while creating a decision tree. This improved algorithm reduced the generalization error rate by 1.08% over its predecessor model. However, it produced small depth classifiers which directly affect its classification accuracy. In[10] introduced a model for pandemic influenza prediction based on a classification and association mining algorithm (CBA). The researchers were the first to introduce CBA which proved superior in comparison with other existing clustering techniques in pandemic influenza detection. However, the model was tested only against the pandemic influenza dataset. In[11] devised a decision tree model to examine the risk of laryngeal pathology in Korean adults. This model implemented many advanced and sophisticated algorithms and techniques such as CART, Rao-Scott chi-square test, ANOVA and 10-fold cross validation to improve the quality of results in terms of accuracy. It successfully classified risk factors associated with laryngeal pathology among Korean adults. In[12] designed a model to study factors related to hypertension in a sample of Iran population. Again, this model only used a single classifier rather than implementing an ensemble of classifiers which could have improved the performance of the model. Hence, the solution is not of importance to our work.

In[13] introduced a technique for classifying uncertain data based on a nearest neighbour class. Usually, uncertain data are discarded from the dataset as it hinders the performance of data mining algorithms and affects the quality of the overall results. In[1] devised a model to study the risk factors related to all diabetes type 2 diseases. The authors devised a predictive model for an early diagnosis of diabetes type 2 without initially going through any kind of extensive laboratory test. The model was based on a decision tree algorithm in conjunction with several techniques to boost the performance of the predictive model. In[14] devised a k Rare-class Nearest Neighbour Classification algorithm (KRNN) which can classify rare classes accurately. Moreover, the solution was featured with dynamic nearest neighbourhood formulation and posterior class probability estimation which can be used to create bias towards the rare class[13,14]. However, the solution failed to deal with the presence of multiple rare classes. Therefore, this algorithm does not add any values to our proposed model. In[4] developed a new decision forest algorithm using a concept of punishing attributes to avoid the repetition of a participation of the same attributes in subsequent tree generation processes. This algorithm generates a group of highly accurate classifiers considering the impact of all categorical attributes rather than single class attributes. Moreover, to attain strong diversity in the set of generated classifiers, the algorithm uses the concept of penalizing attributes used in the last tree to generate the subsequent trees[3,4]. Therefore, the concept forest PA can be added to our proposed model to make it more effective and accurate in terms of prediction results and knowledge discovery.

In[1] devised a prediction model for screening and diagnosing Type-2 Diabetes Mellitus disease without requiring laboratory tests. The prediction was made based on risk factors related to T2DM. Early diagnosis gives doctors and patients a chance to take special health measures promptly. This work had used C4.5 algorithm in a J48 version of implementation to create a single decision tree to successfully study the features associated with T2DM. They also implemented a weak classifier boosting algorithm, AdaboostM1 to enhance the effectiveness and efficiency of decision tree classification and prediction results. In addition, the use of a 10-fold cross validation technique generated a high-quality classifier by performing repetitive training and testing of the dataset. The problem with this solution lies in two elements. The data pre-processing phase and the model creation phase. The rectangle with red dash border shows the area of limitations in Figure 1. The real dataset usually contains both numerical and categorical attributes. The numerical attribute has a very high value range, as a result of which the performance of the data mining algorithm decreases significantly in terms of learning and training time, thereby, affecting its capability of knowledge discovery. Moreover, there are many techniques which do not accept data which contains numerical attributes as it increases computational complexity. Therefore, as for this T2DM model, the decision tree will not be as effective as it could be because of the presence of numerical attributes such as age and BMI. Similarly, the use of a single classifier or a decision tree is a drawback of this model, as it limits the capability of extracting rules or patterns to reduce numbers, thus limiting the capability of knowledge discovery and prediction. Hence, the model offers the possibility of improving its performance in terms of knowledge discovery and prediction by implementing a group of classifiers rather than one classifier as it will generate large number rules associated with data. See Figure 1.

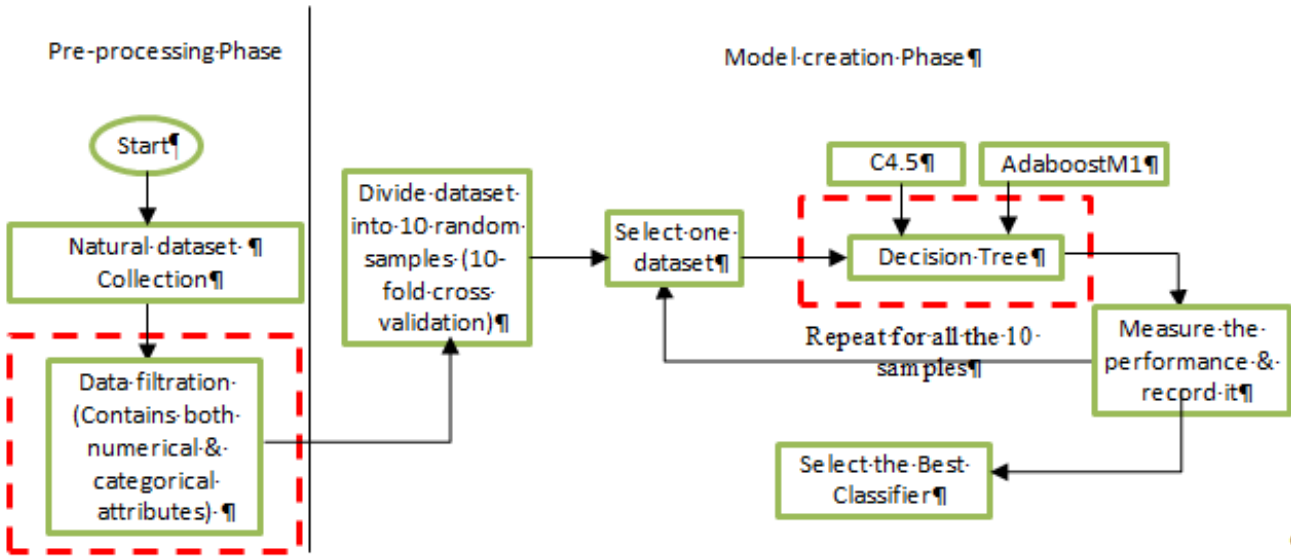

Figure 1: Block diagram of state of art

2. Proposed Solution

To improve the accuracy of the current best model, the following techniques were introduced and implemented. The proposed model in Figure 2 has the following two steps:

- Discretization of numerical attributes.
- Decision Forest with Penalizing attributes.

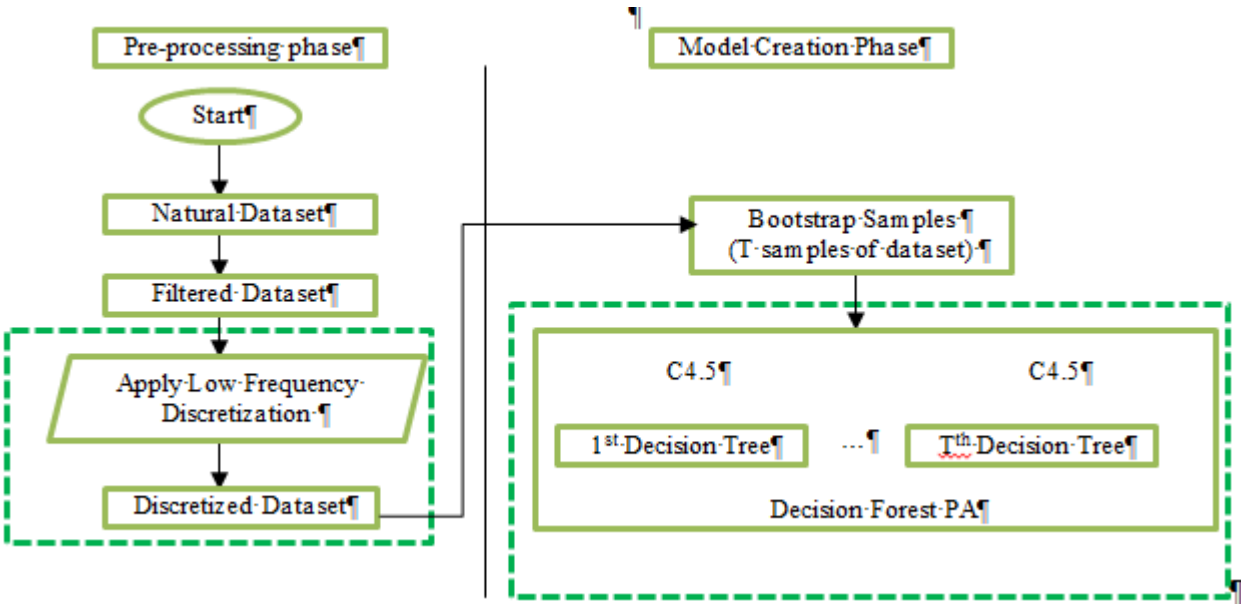

Figure 2. Block diagram of proposed system.

2.1 Discretization of Numerical Attributes

It has been proven that the C4.5 algorithm works much better with discrete or categorical attributes than numerical attributes. The presence of continuous attributes cost the decision tree in terms of capability of classification and prediction accuracy. Therefore, conversion of numerical attributes into categorical attributes is important from the classification accuracy point of view. The process of converting numerical attributes to categorical is called discretization of numerical attributes. The discretization task is performed before the actual tree making process starts. As for our solution, we will be discretising numerical attributes of a dataset using a Low-frequency Discretization (LFD) method. Basically, it spots the split point around the sparse region where the frequency of the attribute's value is less in order to reduce the information loss[5]. Therefore, it keeps important information intact. LFD discretises numerical attributes in an order based on average correlation value of numerical attributes to the categorical attributes in a dataset.

The average correlation value of a numerical attribute α_j is calculated using the Equation (1)[5].

$$\alpha_j = \frac{\sum_{k;\forall A_k \in A_c} \eta_{jk}}{c}$$

Where, A_c represents the categorical attributes and A_k is one of the categories, similarly c represents the number of categorical attributes. η_{jk} represents the correlation ratio.

Unlike other existing discretization algorithms, LFD is the only algorithm which considers the influence of all the categorical attributes while discretizing a numerical attribute. Moreover, once the numerical attribute is discretized, its influence is also measured to discretize other remaining attributes and so on.

2.2 Decision Forest by Penalizing Attributes (Forest PA)

A decision forest is an example of ensemble decision trees. A decision forest can extract more rules and patterns than a single decision tree. Therefore, it makes it superior to one single decision tree in terms of producing a greater capability of classification and prediction accuracy. Therefore, I believe the implementation of decision forests rather than one single decision tree will increase the prediction capability of the model. As for our solution, we will work with Forest PA as it helps to generate more diverse and accurate trees individually[4]. Hence, it produces a group of the best classifiers which will greatly improve the accuracy rate of classifying new data and so increase the prediction accuracy of the model.

The main feature of Forest PA is that it has penalising attributes. It uses two basic concepts assigning and releasing the weight of attributes. The attribute that appears at the higher level are assigned lower weight (i.e. are highly penalized) and those which appear at the lower level are assigned higher weight (less penalized) so that no attributes tested at higher level appear again in subsequent tree generation. Therefore, it helps to generate a diverse decision tree.

The weight assignment is done using the following Equation (2) given by[4]:

$$WR^{\lambda} = \begin{cases} \left[0.0000, e^{-\frac{1}{\lambda}}\right], & if\ \lambda = 1 \\ \left[e^{-\frac{1}{\lambda-1}} + \rho, e^{-\frac{1}{\lambda}}\right], & if\ \lambda > 1 \end{cases}$$

Where, λ is level at which an attribute is tested in the tree and ρ is a constant value 0.0001.

Equation (2) produces high weight value for high value of λ and low weight value when the value of λ is small, thereby preventing those attributes from appearing in an immediate decision tree which were tested at the higher level node of current decision trees. It ensures that all other important attributes are given chance to be tested at near the root node to produce diverse decision forests at the end.

However, if a particular attribute is penalized highly, then the probability of participation of such attribute will become less in subsequent trees and after some point, the good attributes might not appear in trees, thus decreasing the quality of trees as such. Therefore, to avoid such situation, Forest PA also introduces a weight release mechanism for an attribute whose value is calculated using Equation (3) given by[4]:

$$\sigma_i = \frac{1.0 - \omega_i}{(\eta + 1) - \lambda}$$

Where, ω_i weight of the attribute, η height or depth of the tree in which the attribute has participated and λ level at which attribute is tested.

Therefore, concepts of penalizing attributes and releasing its penalty by certain amount help the attribute to appear again in the process of creating a decision tree, ensuring that all important attributes are participating equally after certain intervals depending on their importance. Technically and theoretically, the hybrid concept of combining LFD and Forest PA seems to be more powerful to produce more accurate models of prediction and classification as each of them has already proven their efficiency and performance in the field of data mining.

3. Experimental Results and Discussion

The data were collected from the online repository system maintained by Kaggle group which was provided free of cost. The collected dataset contains records of a diabetes screening test. Each record has nine attributes which are age, Body Mass Index (BMI), blood pressure, glucose, insulin, diabetes pedigree function, skin thickness, pregnancies and outcome. Among nine attributes, the outcome is a categorical attribute which represents patients with specific record details and tested positive or negative for diabetes. The remaining eight attributes are numerical attributes and were used as test nodes during the creation of trees or classifiers for the model.

The model's recall, precision and accuracy were calculated using the following equations based on[1]:

$$\text{Recall} = TP / (TP + FN) \quad (4)$$

$$\text{Recall} = TP / (TP + FN) \quad (5)$$

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (6)$$

Initially, a dataset was pre-processed to remove records with missing values so that there would not be any complications and alteration of extracted logic rules later in the process. After that, the dataset was discretized to discretize the natural dataset containing both categorical and numerical attributes. We considered all influences of the categorical attribute on the numerical attribute while converting these to categorical attributes. At first, the average correlation value of each numerical attribute related to all the categorical attributes was calculated using Equation (1). Based on the average co-relation value, the order of discretizing numerical attributes was set from high to low. After that, the vote value for all possible cut points was measured to identify the best point for splits to discretize it. After all the attributes were discretized, decision forest PA was applied to the discretized dataset.

The following Tables 2–5 show the confusion matrix for the current model as well as for the proposed model. In Table 2, the confusion matrix of the current best model reveals that it could identify 50.52% of healthy persons

from others, whereas only 17% of patients were identified which is less than that for healthy specimen. Similarly, the confusion matrix of the proposed model in Table 3 shows that could separate 57.81% of healthy individuals out of 768 records, while only 20.18% of patients were identified from others which is still less than that of healthy ones. However, the capability of identifying patients correctly was increased for both healthy and not healthy (diabetic) individuals. The Table 4 below describes and compares the precision and false rate of both models. A total of 768 clean records were obtained after the pre-processing of the dataset which was used for generating the ultimate classification and prediction model. Table 5 compares the detailed evaluation of the current model and the proposed model based on parameters including precision, ROC area, accuracy, FP rate etc. Table 5 shows that the accuracy of patient classification and prediction is increased by 10−15% in the proposed model. Precision value has been improved which means the proposed model is able to identify patients more accurately. Similarly, both models had lower FP rates at 0.413 and 0.239 respectively which signifies that the ability to identify healthy people is more precise than identifying the patients. However, the TP rate that is the ability to identify the patients correctly has increased by some considerable margin in the proposed model.

The result also shows that area under the ROC curve has been increased by 8−10% approximately. The graphical representation of both ROC curves is shown in the Figure 3.

Table 1. Pseudocode of the proposed model

INPUT: Diabetic dataset
OUTPUT: Decision forest

Step 1. Discretise the dataset using LFD

a. Calculate the co-relation value of all the numerical attribute in relation to categorical attribute
b. *Average correlation value of the j^{th} attribute,*

$$\alpha_j = \frac{\sum_{k_j \forall A_k \in A_c} \eta_{jk}}{c}$$

c. Rank the numerical attribute based on its co-relation value
d. Calculate the average vote for every possible cut points of the numerical attribute
e. Select the cut point with the highest vote value which ensures the cut point lies near the region with the lowest frequency value of a numerical attribute.
f. Discretise all numerical attributes in order based on rank value assigned to them previously. categorical attributes only.
g. Apply the Forest PA on the discretised dataset obtained from the Step 1 to generate classifiers.

Step 2. Obtain the Forest PA

a. Specify the number of trees or classifiers to generate
b. Start creating the classifier or decision tree
c. Apply the bootstrap technique on the original training dataset
d. Assign the default weight value for all the attribute $\omega = 1.0$
e. Create the decision tree
f. Recalculate the weight value for the entire attribute participating in the current decision tree, and the respective increment value η.
g. Repeat the process from step i to iv until the required number of trees are generated.
h. Increase the weight value of those attributes that are not participating in the subsequent trees. As $\omega = \omega + \eta$

Table 2. Confusion matrix for current best model

Class	Diabetic	Healthy
Diabetic	130	138
Healthy	112	388

Table 3. Confusion matrix for proposed model

Class	Diabetic	Healthy
Diabetic	155	103
Healthy	66	444

Table 4. Classification result of current best and proposed model

Classification Result	Correctly classified (%)	Incorrectly classified (%)
Current Best Model	518 (67.44%)	250 (32.55%)
Proposed Model	599 (77.99%)	169 (22.01%)

Table 5. Evaluation matrix of current best and proposed model (Note: ROC = Receiver Operating Characteristics, FP = False Positive, TP = True Positive

Evaluation Measure	Current Best			Proposed Model		
Class	Diabetic (pos)	Healthy (neg)	Average	Diabetic (pos)	Healthy (neg)	Average
ROC Area	0.687	0.687	0.687	0.744	0.744	0.744
F-Measure	0.510	0.756	0.670	0.541	0.779	0.696
Recall	0.485	0.776	0.674	0.504	0.808	0.702
Precision	0.537	0.738	0.668	0.584	0.752	0.694
TP	0.485	0.776	0.674	0.504	0.808	0.702
FP	0.224	0.515	0.413	0.082	0.396	0.239
Accuracy	0.675			0.779		

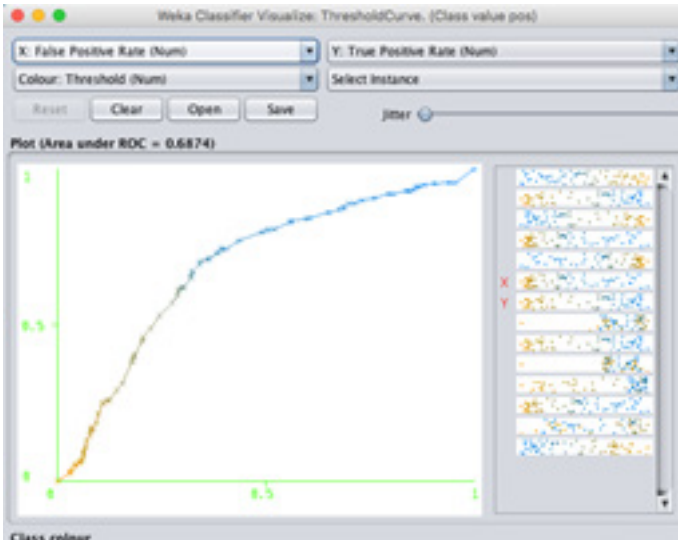

Figure 3-a. ROC curve for current model

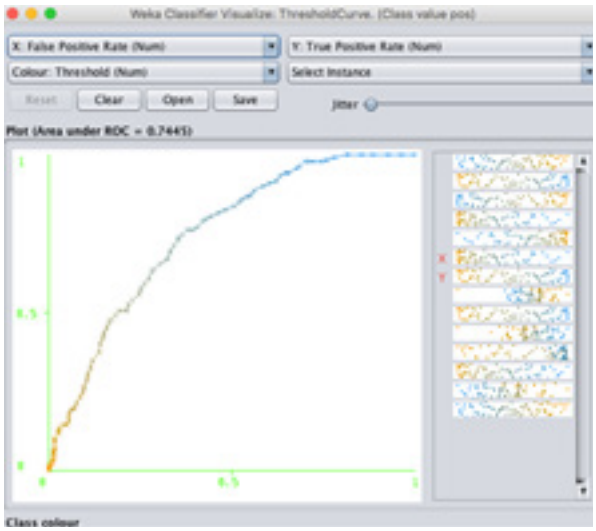

Figure 3-b. ROC curve for proposed model

where *x* axis represents False Positive (FP) and y axis represents True Positive (TP) values.

Figure 3. Classifier model of current best system

4. Conclusion

This study has tested the concept of data discretization and ensemble classifier which proved to be one of the important aspects of the classification and prediction model. Data discretization discretizes continuous attributes, whereas the decision forest played an important role in generating several classifiers making it more effective in generating logic rules for classification as opposed to one classifier with fewer rules. The proposed model uses low-frequency attribute value region as a cut point considering the influence of all the categorical attributes in the record to minimize information loss in conjunction with Forest PA to generate diverse and accurate classifiers from the discretized dataset. The proposed model only works if there is at least one categorical attribute. Therefore, further research needs to be carried out in future to make improve the independence of categorical attributes. Using a hybrid concept by combining LFD and Forest PA for our proposed system, we were able to improve the performance of classification and prediction of the current best model significantly. Overall accuracy has increased by 9~10% along with many other important factors such as F-measures and ROC curve performance while dealing with real-time datasets. The improvement in the performance and accuracy of the model will greatly assist medical experts to make informed decisions when operating in real-time environments with the real-time datasets.

5. References

1. Habibi S, Ahmadi M, Alizadeh S. Type 2 Diabetes Mellitus screening and risk factors using decision tree: Results of data mining. Global Journal of Health Science. 2015; 7(5):304–10. PMCid: PMC4803907. https://doi.org/10.5539/gjhs.v7n5p304.
2. Advantages and disadvantages of data mining. 2017. https://www.zentut.com/data-mining/advantages-and-disadvantages-of-data-mining/.
3. Adnan MN, Islam MZ. Forest CERN: A new decision forest building technique. Proceedings of the Twentieth Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD); 2016. p. 304–15. https://doi.org/10.1007/978-3-319-31753-3_25.
4. Adnan M, Islam M. Forest PA: Constructing a decision forest by penalizing attributes used in previous trees. Expert Systems with Applications. 2017; 89:389–403. https://doi.org/10.1016/j.eswa.2017.08.002.
5. Geaur Rahman M, Zahidul Islam M. Discretization of continuous attributes through low frequency numerical values and attribute interdependency. Expert Systems with Applications. 2016; 45:410–23. https://doi.org/10.1016/j.eswa.2015.10.005.
6. Sharma A, Gupta R, Tiwari A. Improved density based spatial clustering of applications of noise clustering algorithm for knowledge discovery in spatial data. Mathematical Problems in Engineering. 2016; 1564516:1–9. https://doi.org/10.1155/2016/1564516.
7. Haraty R, Dimishkieh M, Masud M. An enhanced k-means clustering algorithm for pattern discovery in healthcare data. International Journal of Distributed Sensor Networks. 2015; 11(6):615–740. https://doi.org/10.1155/2015/615740.
8. Huang T, Hsu W, Chen Y. Conjecturable knowledge discovery: A fuzzy clustering approach. Fuzzy Sets and Systems. 2015; 221:1–23. https://doi.org/10.1016/j.fss.2012.12.006.
9. Tzirakis P, Tjortjis C. T3C: Improving a decision tree classification algorithm's interval splits on continuous attributes. Advances in Data Analysis and Classification. 2016; 11(2):353–70. https://doi.org/10.1007/s11634-016-0246-x.
10. Kargarfard F, Sami A, Ebrahimie E. Knowledge discovery and sequence-based prediction of pandemic influenza using an integrated classification and association rule mining (CBA) algorithm. Journal of Biomedical Informatics. 2015; 57:181–8. PMid: 26232668. https://doi.org/10.1016/j.jbi.2015.07.018.
11. Byeon H. The risk factors of laryngeal pathology in Korean adults using a decision tree model. Journal of Voice. 2015; 29(1):59–64. PMid: 25008378. https://doi.org/10.1016/j.jvoice.2014.04.004.
12. Tayefi M, Esmaeili H, Saberi Karimian M, Amirabadi Zadeh A, Ebrahimi M, Safarian M. The application of a decision tree to establish the parameters associated with hypertension. Computer Methods and Programs in Biomedicine. 2017; 139:83–91. PMid: 28187897. https://doi.org/10.1016/j.cmpb.2016.10.020.
13. Angiulli F, Fassetti F. Nearest neighbor-based classification of uncertain data. ACM Transactions on Knowledge Discovery from Data. 2015; 7(1):1–35. https://doi.org/10.1145/2435209.2435210.
14. Zhang X, Li Y, Kotagiri R, Wu L, Tari Z, Cheriet M. KRNN: k Rare-class Nearest Neighbour classification. Pattern Recognition. 2014; 62:33–44. https://doi.org/10.1016/j.patcog.2016.08.023.