A Novel Ensemble Based Decision Tree Model for High Dimensional Biomedicine Data

Dande Rushalini[*1], Mr. K. Vijay Kumar[*2]

PG student, CSE Department, PVPSIT college/JNTUK University, Vijayawada, A.P, India
Asst. Professor, CSE Department, PVPSIT College/JNTUK University, Vijayawada, A.P, India

Abstract *–Knowledge discovery is an essential mechanism for the intelligent data analiysis to transform data in to meaniful information that will support for decision making. Data mining approaches support automatic exraction of data, and attempts to discover the hidden rules and patterns in data, and also detact relevant decision rules from the high dimensional dataset. Classification from imbalanced data is significantly affected the performance of the algorithm due to noise and high dimensionality. Sparsity and high dimensionality of the classifier algorithm becomes a major problem in many traditional decision tree models on medical datasets. A novel decision trees allows estimating on topmost features to assess the class prior probability and estimates the chance of misleading false positive patterns. In this research work, a new framework is proposed by integrating random forest decision tree for pattern analysis. Experimental results show that proposed model has better accuracy compare to existing approaches.*

Keywords — *PSO model, Disease detection, Random forest, classification ,UCI repository.*

I. Introduction

Heart diseases especially coronary heart disease is a very fatal and dangerous disease because if patient ignores its earlier symptoms, which seems to be a warning signs, it gives no time to patient for recovery and eventually may lead to death on spot. This is called as heart attack. It happens because the function of the arteries is to supply oxygen rich blood to heart but due to fatty and other substance the plaque is formed which turns normal coronary artery into narrowing of the coronary arteries. Today the buzz word is “ Health Care” all over the world. Early prediction of diseases can reduce the fatal rate of human. There are large and enormous data available in hospitals and medical related institutions.Information technology plays a vital role in Health Care. Diabetes is a chronic disease with the potential to cause a worldwide Health Care crisis

According to international Diabetes Federation 382 million people are living with diabetes world wide. By 2035, this will be doubled as 592 million. Early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors. Diabetes affects human organs such as kidney, eye, heart, nerves, foot etc.

Data mining is the discovery of knowledge in databases. Techniques of data mining help to process the data and turn them in to useful information. Prediction results from data mining are useful in various fields like Business Intelligence, Bioinformatics, Health care management, Finance etc. Medical field has wide amount as well as variety of data for processing and there exist many challenging tasks. This field requires accurate and timely mannered diagnosis. Based on the data used the accuracy and performance also vary.

Medical filed contains large amount of data that are needed to be processed. Data mining in medical field improves the quality of patient care and the prediction of health care patterns. Data mining tools helps us to discover unknown patterns, group the related items and decision making of health care oriented problems. Medical care is necessity, it gives patient and hope for a fruitful life. The collected data when published is used forsocial causes without harming the dignity of patients. Early detection of disease can increase the survivability of patients[1].

II. RELATED WORK

Data mining is a process to extract useful information from large database. It is a multidisciplinary field of computer science which involves computational process, machine learning, statistical techniques, classification, clustering and discovering patterns. Data mining techniques has proved for early prediction of disease with higher accuracy in order to save human life and reduce the treatment cost.

This paper explores various Data mining techniques such as Navie Bayes, MLP, Bayesian

Network, C4.5, Amalgam KNN, ANFIS, PLS-LDA, Homegenity-Based, ANN, Modified J48 etc. are analyzed to predict the diabetes disease. [2-4] combined KNN to improve the accuracy in prediction. In this K-means and KNN are combined to overcome the computational complexity of large number of dataset. And the training set is verified with fuzzy systems and neural networks to produce better result. [5] implemented genetic algorithm with data mining techniques to test the patients affected by diabetes based upon the fitness value and the accuracy chromosome value.

Support Vector Machine (SVM) is a regulatory algorithm introduced by Vapnik in 1995. The base of the algorithm is using the precision to generalize the errors. The algorithm makes "hyperplane" and divides the data into classes so that all samples belonging to one class will be categorized on one side and the rest on the other side. Linear SVM Classifier is defined for the SVM classifying task, and dividing them occurs provided that the chosen line involves the most marginalized sure.

Naïve Bayes Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. A naive Bayes classifier is a term dealing with a simple probabilistic classification based on applying Bayes theorem. In simple terms, a naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.

I. THE BELOW TABLE 1 SHOWS THE COMPARISON OF ALGORITHMS USED FOR DISEASE PREDICTION.

Disease Algorithm	Heart	Kidney	Liver	Diabetes	Cancer
Decision Tree	√	√	√	√	√
Naïve Bayes	√				√
Neural Networks	√	√			
Fuzzy			√		
SVM	√	√	√	√	
Multilayer Perceptron		√			√
Simple Logistic				√	

Artificial Neural Network is a data processing algorithm, originated from human brain. The system includes a large number of tiny processors to handle data processing. The processors act in the form of an interconnected network parallel to each other to solve a problem. Using programming knowledge, in this networks a data structure is designed that can act as neurons. This data structure is called the neuron.

III. PROPOSED MODEL

The standard filtering algorithms are not adaptive to currently the conditions when biomedical dataset is large. This might generate the false recommendations. In this model , a new recommended collaborating filter can be used in order to get active probe patterns dynamically based on the medical feature changing by using the dynamic likeness. The process for the proposed recommendations is as follows.

Step1.Determining the actual attribute features along with distributed weights R(i,c).This action is used to have the best attributes for kind of feature detection using R(i,c)=best feature selection method(data).This step is used to select the best weight for the attribute selection.

Step2.Using approved pearson's correlation to gauge the similarity pearson's correlation as follows, measures the linear correlation between 2 vectors of ratings.

$$sim(i,j)=\frac{\sum_{c\in I_{i,j}}(R_{i,c}-A_i)(R_{j,c}-A_j)}{\sqrt{\sum_{c\in I_{i,j}}(R_{i,c}-A_i)^2 * \sum_{c\in I_{i,j}}(R_{j,c}-A_j)^2}}$$

Where Ri,c is the attribute weight of the given dataset.

Step3. Neighbor Selection

After computing the similarity between the attributes .Next we need to find k neighborhood selection .

If(sim(i,j>thres)

Then

Continue;

Else

Remove attribute;

Random Forests

That needs to be classified is put down each of the tree gives The random forests are an esemble of unpurned classification or regression trees and each tree is constructed by a different bootstrap sample from the original data using a tree classification algorithm.After the forest is formed ,a new object that needs to be classified is put down each of the tree in the forest for classification. Each tree gives a vote that indicates the tree's decision about the class of the object. The forest chooses the class with the most words for the object. The random forest algorithm(for both classification and regression) is as follows:

1. From the training of *n* samples draw *n* tree bootstrap samples.

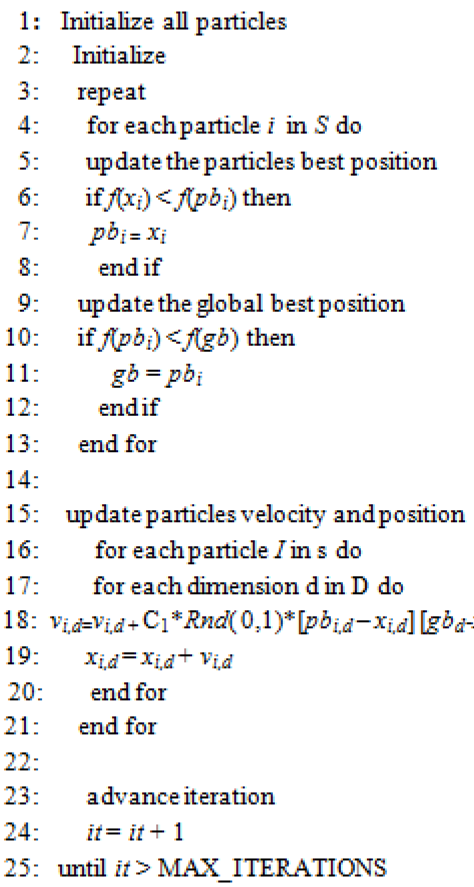
2. For each of the bootstrap samples, grow classification or regression tree with the following

modification; at each node, rather than choosing the best split among all predictors, randomly sample *m* try of the predictors and choose the best split among those variables. The tree is grown to the maximum size and not pruned back.

Bagging can be thought of as the special case of random forests obtained when *m* try=*p*, the number of predictors.

3. Predict new data by aggregating the predictions of the *n* tree trees(i.e., majority word for classification, average for regression).

Pseudo-code of the PSO algorithm



```
Initialize all particles
 Initialize
  repeat
   for each particle $i$ in $S$ do
   update the particles best position
   if $f(x_i) < f(pb_i)$ then
    $pb_i = x_i$
    end if
  update the global best position
  if $f(pb_i) < f(gb)$ then
      $gb = pb_i$
    end if
  end for

 update particles velocity and position
     for each particle $I$ in s do
     for each dimension d in D do
$v_{i,d}=v_{i,d}+C_1*Rnd(0,1)*[pb_{i,d}-x_{i,d}]\ [gb_d-x_{i,d}]$
     $x_{i,d}=x_{i,d}+v_{i,d}$
     end for
  end for

   advance iteration
   $it = it + 1$
until $it >$ MAX_ITERATIONS
```

i particles index, used as a particle identifier;

d dimension being considered, each particle has a postion and a velocity for each dimension;

it iteration number, the algorithm is iterative;

$x_{i,d}$ position of particle i in dimension d;

$v_{i,d}$ velocity of particle i in dimension d;

C_1 acceleration constant for the cognitive component;

Rnd stochastic component of the algorithm, a random value between 0 and 1;

$pb_{i,d}$ the location in dimension d with the best fitness of all the visited locations in that dimension of particle i;

C_2 acceleration constant for the social component;

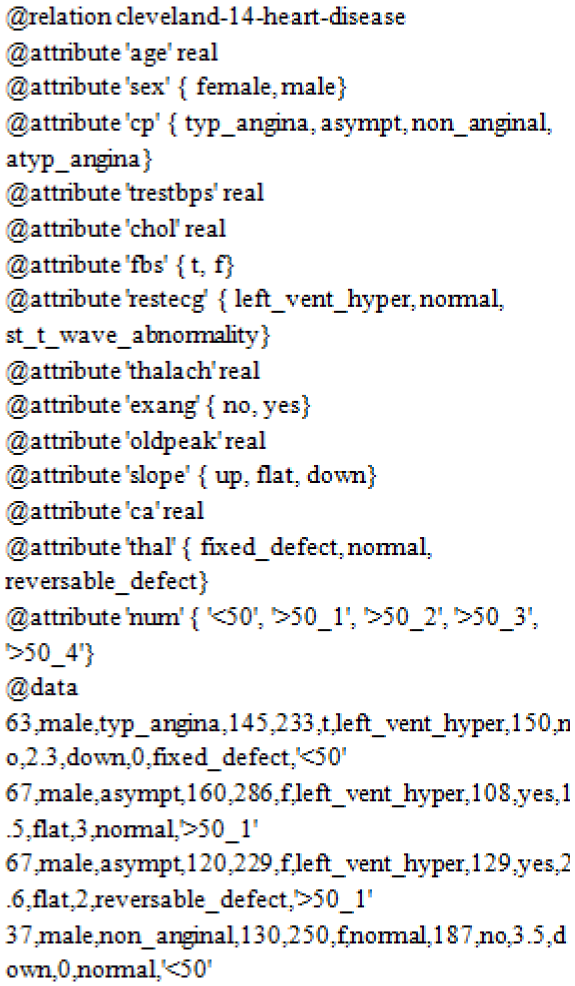
IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

DATASETS

In our experiment, we investigated four different classification models along with pre-processing techniques.

Sample dataset:

Heart-c Dataset:



```
@relation cleveland-14-heart-disease
@attribute 'age' real
@attribute 'sex' { female, male}
@attribute 'cp' { typ_angina, asympt, non_anginal,
atyp_angina}
@attribute 'trestbps' real
@attribute 'chol' real
@attribute 'fbs' { t, f}
@attribute 'restecg' { left_vent_hyper, normal,
st_t_wave_abnormality}
@attribute 'thalach' real
@attribute 'exang' { no, yes}
@attribute 'oldpeak' real
@attribute 'slope' { up, flat, down}
@attribute 'ca' real
@attribute 'thal' { fixed_defect, normal,
reversable_defect}
@attribute 'num' { '<50', '>50_1', '>50_2', '>50_3',
'>50_4'}
@data
63,male,typ_angina,145,233,t,left_vent_hyper,150,n
o,2.3,down,0,fixed_defect,'<50'
67,male,asympt,160,286,f,left_vent_hyper,108,yes,1
.5,flat,3,normal,'>50_1'
67,male,asympt,120,229,f,left_vent_hyper,129,yes,2
.6,flat,2,reversable_defect,'>50_1'
37,male,non_anginal,130,250,f,normal,187,no,3.5,d
own,0,normal,'<50'
```

41,female,atyp_angina,130,204,f,left_vent_hyper,17
2,no,1.4,up,0,normal,'<50'
56,male,atyp_angina,120,236,f,normal,178,no,0.8,up
,0,normal,'<50'
62,female,asympt,140,268,f,left_vent_hyper,160,no,
3.6,down,2,normal,'>50_1'

```
57,female,asympt,120,354,f,normal,163,yes,0.6,up,0,normal,'<50'
63,male,asympt,130,254,f,left_vent_hyper,147,no,1.4,flat,1,reversable_defect,'>50_1'
53,male,asympt,140,203,t,left_vent_hyper,155,yes,3.1,down,0,reversable_defect,'>50_1'
57,male,asympt,140,192,f,normal,148,no,0.4,flat,0,fixed_defect,'<50'
56,female,atyp_angina,140,294,f,left_vent_hyper,153,no,1.3,flat,0,normal,'<50'
56,male,non_anginal,130,256,t,left_vent_hyper,142,yes,0.6,flat,1,fixed_defect,'>50_1'
44,male,atyp_angina,120,263,f,normal,173,no,0,up,0,reversable_defect,'<50'
52,male,non_anginal,172,199,t,normal,162,no,0.5,up,0,reversable_defect,'<50'
57,male,non_anginal,150,168,f,normal,174,no,1.6,up,0,normal,'<50'
48,male,atyp_angina,110,229,f,normal,168,no,1,down,0,reversable_defect,'>50_1'
54,male,asympt,140,239,f,normal,160,no,1.2,up,0,normal,'<50'
```

Proposed Approach

```
==========
cp = typ_angina
|  trestbps < 139
| |  chol < 198.5 ==> <50
|  |  chol >= 198.5 ==> <50
|  trestbps >= 139 ==> <50
cp = asympt
|  ca < 0.5
| | thal = fixed_defect ==> <50
| | thal = normal
| | |  age < 51.5 ==> <50
|  |  |  age >= 51.5
| | | | exang = no
| | | | |  chol < 300.5 ==> <50
|  |  |  |  |  chol >= 300.5
| | | | | |  age < 61.5 ==> <50
|  |  |  |  |  |  age >= 61.5 ==> <50
| | | | exang = yes
| | | | |  trestbps < 122 ==> <50
|  |  |  |  |  trestbps >= 122 ==> <50
| | thal = reversable_defect
| | |  oldpeak < 0.65
| | | |  chol < 237.5
| | | | |  age < 42 ==> <50
|  |  |  |  |  age >= 42 ==> <50
|  |  |  |  chol >= 237.5 ==> <50
|  |  |  oldpeak >= 0.65 ==> >50_1
|  |  |  chol >= 158.5
| | | |  chol < 328 ==> <50
|  |  |  |  |  chol >= 328 ==> <50
|  |  oldpeak >= 2.7 ==> <50
| thal = reversable_defect
| |  oldpeak < 1.9
| | | slope = up ==> <50
| | | slope = flat ==> <50
| | | slope = down ==> <50
|  |  oldpeak >= 1.9 ==> >50_1
cp = atyp_angina
|  age < 56.5
| | thal = fixed_defect ==> <50
| | thal = normal ==> <50
| | thal = reversable_defect ==> <50
|  age >= 56.5 ==> <5
```

Table 2: Accuracy Performance of the proposed and traditional models

Algorithm	TruePositive	Accuracy
RandomForest	0.783	84.51
NeuralNetworks	0.852	89.43
NN+PSO	0.931	92.54
Proposed	0.962	97.43

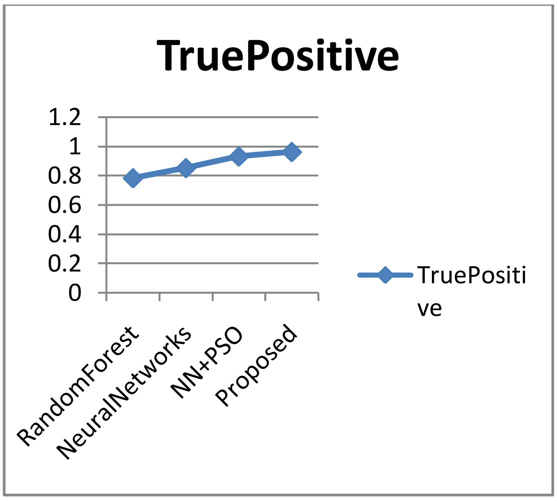


Figure 1: Comparative graph of proposed and existing model

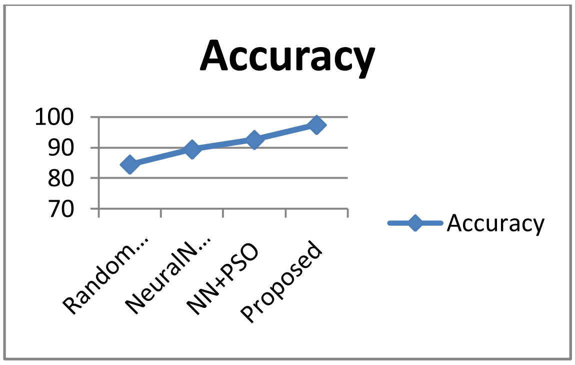


Figure 2: Comparative graph of proposed and existing model

V. CONCLUSION

Classification from imbalanced data is significantly affected the performance of the algorithm due to noise and high dimensionality. Sparsity and high dimensionality of the classifier algorithm becomes a major problem in many traditional decision tree models on medical datasets. A novel decision trees allows estimating on topmost features to assess the class prior probability and estimates the chance of misleading false positive patterns. In this research work, a new framework is proposed by integrating random forest decision tree for pattern analysis. Experimental results show that proposed model has better accuracy compare to existing approaches.In future this work can be extended to optimize the realtime disease prediction in web based applications.

REFERENCES

[1] Quinlan J R,"Simplifying Decision Tree," Internet Journal of Man-Machine Studies,1987,27,pp.221-234

[2] Yang Xue-bing,Zhang Jun, "Decision Tree Algorithm and its core technology", techno Logy and development, 2013.

[3] Qu Kai-she ,Wen Cheng-li, Wang Jun-hong, "An improved algorithm of ID3 algorithm," Computer Engineering and Applications, 2003,(25),pp.104-107

[4] Mao Cong-liYi Bo, "The most simple decision tree generation algorithm based on decision-making degree of coordination ,“ Computer Engineering and Design,2008,29(5),pp.1250-1252

[5] Huang Ai-hui,"Improvement and application of decision tree C4.5 algorithm ,“ Science Technology and Engineering,2009, (1),pp.34-37

[6] J. Gehrke, R.Ramakrishnan, and V. Ganti, "Rainforest, a framework for fast decision tree construction of large datasets", in Springer Netherlands-Data mining and knowledge discovery vol.4. Issue(2-3) July 2000.

[7] M. Kantardzic “Data Mining. Concepts, Models, Methods and Algoritms”. John Wiley and Sons Inc, 2003.

[8] Xu.M.Wang, J.and Chen.T. “Improved decision tree algorithm: ID3+” Intelligent Computing in Signal Processing and Pattern Recognition, Vol.345, pp.141-149, 2006

[9] Quinlan, J. R. “C4.5: Programs for Machine Learning” Morgan Kaufmann, San Mateo, CA 1993.

[10] Lewis, R.J. “An Introduction to Classification and Regression Tree (CART) Analysis” Annual Meeting of the Society for Academic Emergency Medicine, Francisco 2000.

[11] Ruoming Jin, Ge Yang and Gagan Agrawal, “Shared memory parallelization of Data mining algorithms: Techniques, Programming interface and Performance”, IEEE Transactions on Knowledge & data engineering, 2005.

[12] Song Xudong, Cheng Xiaolan “Decision tree Algorithm based on Sampling” IFIP International conference on Netwok and Parallel Computing-Workshops 2007.