CLASSIFICATION OF SPECT HEART DATASET USING EFFICIENT MODIFIED RANDOM FOREST ALGORITHM

Mrs K.Ammulu 1, Dr T.Venugopal 2

1Associate Professor, Department of Computer Science, Dravidian University, Kuppam, Andhrapradesh
kambhamammulu@gmail.com
2Professor & Head, Dept of CSE & IT, JNTUH, Sultanpur, Medak, Telangana
t_vgopal@rediffmail.com

Abstract: Data mining is one of the most powerful technique and which is embedded in many applications. The scholar is working in data mining domain proposed many solutions to solve classification problems. Random forest algorithm is a one of the ensemble learning method used for data classification and regression of dataset. Random forest algorithm facing classification problem and it is not classifying dataset accurately. In this paper, we propose modified random forest algorithm to merge random forest algorithm and particle swarm optimization algorithm. To analyze the performance of proposed method, SPECT heart dataset with 17 variables and 10299 instances of data is taken as an input from UCI repository, confusion matrix is used to measure accuracy, recall, F1, and precision. Compare with random forest algorithm proposed method exhibit better results.

Keywords: Data Mining, Classification, Modified Random Forest Algorithm, SPECT Heart Dataset, Prediction

I. INTRODUCTION

Data mining is concerned with locating hid relationships present in commercial enterprise data presents groups to make predictions for eventual use. It is the procedure of records-driven extraction of not so evident however rather helpful data from large databases. Data mining has risen as a key commercial enterprise intelligence technology.

The preference of data mining is to extract implicit, previously unexplored and doubtlessly beneficial (or actionable) styles from statistics. Data mining encompass many up to date tactics along with classification (neural networks, k-nearest neighbor, naive Bayes classifier, and decision trees), clustering (density-based clustering, k-means, hierarchical clustering), association (constraint-based association, multilevel association, multidimensional, one-dimensional). Years of education display that data mining is a technique, and it’s helpful application calls for post processing (presentation, understandability, summary), facts preprocessing (cleaning, noise/outlier removal, dimensionality reduction) true knowledge of problem domains and domain facility.

All traditional algorithms are exaggerated to some amount by the class imbalance crisis. Also, the accurate choice of the metric (or consolidation of metrics) to assess – and ultimately improve, is essential for the accomplishment of a data mining effort in such areas, since most of the time improving one metric degrades others.

A first unique contribution obtainable in this paper is the modification of the random forest algorithm based on the particle swarm optimization technique which assesses the performance of traditional classification algorithms under imbalanced class distributions. The main objective of our research was to study the data mining techniques and the different methodologies in data mining. Then the random forest is improvised using the optimization techniques (i.e., particle swarm optimization technique). Decision trees for regression have not been studied yet in the context of using a best class for classifying, despite the fact that many interesting real-world applications require a more accurate classification technique. Within this study, we aimed to develop tree based technique for the classification by integrating it with the optimization approaches.

Our goal was to follow the main developments within the line of algorithms for the classification trees from the dataset, that is, include particle swarm optimization inside the tree learning algorithms that is random forest with the particle swarm optimization algorithm, introduced to select a best class from the dataset to obtain a better classifying technique.

Remaining sections of the paper outlined as, Section II covers related background study, Section III represents modified random forest algorithm, Section IV described performance evaluation results, and section V covered conclusion of the paper.

II. RELATED WORK

In [1], authors Surekha et al proposed a new method of feature subset selection method to analyze thyroid disease. They have studied and compared the performance of RST, EC, PSQ, and Genetic Algorithms. They have used thyroid dataset for evaluation and classification is done by using Naive Bayesian and K-Nearest Neighbor. The exploratory impacts demonstrated that EC approach for FSS outflanks RST based technique in each the expressions of delivering least subset of components and resulting in better arrangement precision.

In [2], researchers Taheri et al introduced optimization techniques based on Naive Bayes Classifier. Various kinds of textual data are available in web. In their method they are
trying to improve understanding capability of textual classifier and design a tool for text data classification. They described the assumptions made inside of the derivation of Naive Bayes, noting primary homes and providing approaches for its extension and development. Next, they checked the best of Naive Bayes parameter estimates and their impact on classification. Their experimental results identified the reason behind the improvements that can be determined in multiclass category with Naive Bayes using Error-Correcting.

In [3], scholars Francisco et al proposed classification method using both K Means and Random decision forests algorithms. In area of self sustaining and sensible automobiles, the purpose of pedestrian classification is to reduce quantity of accidents. The object class accuracy relies upon at the sort of classifier and the extracted item capabilities used for classification. Support Vector Machines (SVM), is taken into consideration the most effective classifier for this challenge. But, it relies upon on quite a number of factors that require researchers to perform numerous modifications to achieve an amazing result with ok overall performance. This take a look at affords a promising opportunity with fewer parameters, which works on huge datasets, and decreased runtime. It additionally has the benefit of allowing the facts evaluation between each step of the algorithm.

In [4], authors Haris et al investigated a method for design making using Data Mining. Data mining tools may be very beneficial to expect the destiny developments and overall performance, allowing decision makers to make forecasting on the data collected. In optimization consider many issues to identify minimum and maximum optimal solutions. The management inside the introduction of decision making procedure derived from datasets should use unique techniques to boost up the transformation of records in one-of-a-kind levels. But the integration of data mining and optimization is appeared as a brand new combination method to enhance the quality of the decision making procedure.

In [5], researchers Tanupriya et al proposed Intelligent Data Mining classification algorithm to analyze lung and oral cancer dataset. The main purpose of their research is to design GUI as the way to input the affected patient’s document. Thus count on whether or not or no longer the patient is suffering by the lung and oral cancer by way of the use of the rule based classifier. The forecast is done from the technique of mining the current informative statistics of the affected person and records repository or the reference cost for each attributes. The reference values are taken from expert doctors, and net repository. On their research a contrast of Multilayer perceptron and logistic set of rules have been used to analyze patient’s data to detect cancer.

In [6], authors Bang et al proposed ensemble Naive Bayes algorithm to analyze asthma patient’s dataset. In this work, they used supervised learning strategy and a new classification algorithm called Naive Bayes ensemble, which can classify unlabeled topics from a new dataset into the clusters/subtypes previously recognized from an in advance dataset so that you can still make full use of the brand new dataset for similarly evaluation. They also endorse a technique to robotically choose single classifiers to be protected inside the ensemble, which can be based on each type performance and the diversity of the single classifiers. The Naive Bayes ensemble outperforms 7 single classifiers and other ensemble techniques when they used to evaluated their overall performance.

In [7], Sunita et al conducted a survey on application of particle swarm optimization in Data Clustering. They introduce the LeGO (gaining knowledge of for worldwide Optimization) technique for optimization where in Machine learning is used to predict the final results of a computation. They planned to apply the proposed method to support vector device to examine the connection between the places of starting to the final results.

In [8], researchers Jiawei et al proposed map reduced technique based on random forest algorithm. It is a new aggressive technique is developed for teach agents to play games. This method makes use of particle swarm optimizers (PSO) to teach neural networks to predict the desirability of states within the leaf nodes of a game tree. The new technique is applied to the Tic-Tac-Toe game and compared with the overall performance of an evolutionary approach. In the absence of any manager to provide target outputs to the NNs, the game tree is largely used to quantify the overall performance or fitness of an NN. A population of NNs, in which person NNs compete towards a sample of the opposite people for survivals.

### III. MODIFIED RANDOM FOREST ALGORITHM

The main approach in the decision tree learning is decision tree pruning. The process of removing the tree that present poor voting at the final classification output by reducing the tree size is defined as the tree pruning. The main advantage of this is reducing the tree size, the complication of the build tree, and the reduction of over-fitting. In most of the cases, random forest algorithm will not use this approach. Alternatively, random forest classifier use ntree as parameter that responds to the number of decision trees which is created during the ensemble bagged forest classifier. The random forest algorithm describes the datasets training for each decision tree.

### ALGORITHM: RANDOM FORESTS TRAINING

Step 1: Set: Number of classes = N, Number of features = M  
Step 2: Let: m determine the number of features at a node of decision tree, (m ≤ M)  
Step 3: for each decision tree do  
Step 4: Select randomly: a subset (with replacement) of training data that represents the N classes and use the rest of data to measure the error of the tree  
Step 5: for each node of this tree do  
Step 6: Select randomly: m features to determine the decision at this node and calculate the best split accordingly. /* No tree pruning used */  
Step 7: end for  
Step 8: end for

Random forest algorithm takes more time to build multiple classifiers even it is highly parallel algorithm. This can’t be processed in a usual core (need multiple core) and also not without any graphical processing unit. Initially, the dataset needs to be trained in order to build the classifier at the testing phase. The large number of trees will take more time for the real-time prediction of the classifier. However the more number of trees will provide more accuracy in final. Once it is trained, the remaining dataset is treated as a testing dataset where the prediction of class is not available at shorter duration. The list of parameters used in Random forest algorithm is listed in Table I.

| Table I. Random Forest Parameters Optimized |
ALGORITHM: MODIFIED RANDOM FORESTS TRAINING

Step 1: Initialize Population. The population size is set to pop size, the max iteration time is set to max gen, the position of the binary particle is \( X_k = \{ Z_k,1, Z_k,2, \ldots \} \), \( k = 1, 2, \ldots \) pop size, the velocity is \( V \), the learning factors are \( c_1, c_2 \), and the weight is \( w \).

Step 2: Combine the PSO with RF classification and calculate the fitness function \( F = \max(1/t) \), \( t = 1 \).

Step 3: WHILE (till the ending criterion)

Step 4: FOR \( p = 1 \) to \( n \) particles

Choose attributes
Divide training data and testing data using K-fold Cross-validation
Train the data available for it
Classify the remaining data available for test
Store the estimated rate in an array

Step 5: For NEXT \( p \)

Step 6: Update particle’s velocity and position \( V_k + 1 \) velocity and position \( X_k + 1 \) position of particles. \( P_k \) be the optimal position of an individual particle, \( P_{gk} \) be the optimal position of all particles, and \( rand \) be a random number uniformly distributed in the range \( (0,1) \).

Step 7: If \( gen+max gen \), the algorithm will terminate; otherwise, return to Step 6.

Step 8: NEXT generation ends until it reaches the stopping criterion

Modified Random forest algorithm is one of the best authentic learning algorithms available for enormous data sets where it constructs more accurate classifier. There is adequate approach to obtain the missing data and it also maintains the accuracy even there is huge amount of data is missing. This provides an internal impartial estimation of error occurred due to generalization as the forest built prior. The prototypes are processed by the data given for the interaction between the classification and variables.

IV. PERFORMANCE EVALUATION

List of notations used in the proposed method of classification:

- \( D \) = Dataset
- \( N \) = Number of instances
- \( TP \) = True Positive
- \( FP \) = False Positive
- \( TPR \) = True Positive Rate
- \( FPR \) = False Positive Rate
- \( P \) = positive instances
- \( N^* \) = negative instances
- \( A \) = Accuracy
- \( P \) = Precision

\( R \) = Recall
\( F1 \) = Score
\( r \) = range of misclassified instances

The confusion matrix illustrates the more accuracy of the solution to a classification problem. The classification result can be obtained from the true positive, false true, true negative and false negative rate. The term true and false refer to whether the predicted corresponds to the trusted external judgments. \( P \) as positive instances and \( N^* \) as negative instances then the true positive rate and false positive rate is TPR and FPR respectively. Precision is defined as the range of true positives over the number of true positives plus the number of false positives. Recall is defined as the number of the true positives over the number of true positives plus the number of false negatives. The F1 score is a measure of the test’s accuracy where both the precision and recall are used to compute the rate. Accuracy is explains as the number of all correct predictions split by the total number of the dataset.

\[
P = \frac{TP}{TP+FP}
\]

\[
R = \frac{TP}{TP+FN}
\]

\[
F1 = 2 \times \frac{TP}{2TP+FN+FP}
\]

\[
A = \frac{TP+TN}{P+N}
\]

\[
TPR = \frac{TP}{TP+FN}
\]

\[
FPR = \frac{FP}{FP+TN}
\]

The following table represents the confusion matrix values for the human activity recognition using smart phones with 10299 instances with 17 variables. The SPECT Heart dataset contains attribute such as overall diagnosis, F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11, F12, F13, F14, F15, F16, F17, F18, F19, F20, F21, F22. The table shows the true and false positive count. The dataset type is multivariate. The following table II represents the confusion matrix values for both the test and the train data. The accuracy of random forest algorithm and the modified random forest algorithm for the SPECT Heart dataset is measured by varying number of trees, results are shown in Table III. Precision, recall, F1 score, Accuracy of RF and MRF is evaluated and corresponding graph is shown in figure 1. Accuracy with number of trees is compared and corresponding graph is depicted in figure 2.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Range</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ntree</td>
<td>1 – 100</td>
<td>Discrete</td>
</tr>
<tr>
<td>Sampsize</td>
<td>100 – 10000</td>
<td>Discrete</td>
</tr>
<tr>
<td>Mtry</td>
<td>1 - max. no. of predictors in dataset</td>
<td>Discrete</td>
</tr>
<tr>
<td>Number of features for finding best split node (m)</td>
<td>Min:2 – Max:100</td>
<td>Discrete</td>
</tr>
</tbody>
</table>

Table II. Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>Predicted Positive</th>
<th>Predicted Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Positive</td>
<td>TP : 179</td>
<td>FN : 04</td>
</tr>
<tr>
<td>Observed Negative</td>
<td>FP : 09</td>
<td>TN : 75</td>
</tr>
</tbody>
</table>

\[
P = \frac{TP}{TP+FP} = 0.9521
\]

\[
R = \frac{TP}{TP+FN} = 0.9781
\]

\[
F1 = 2 \times \frac{TP}{2TP+FN+FP} = 0.9648
\]

\[
A = \frac{TP+TN}{N+N^*} = 0.9513
\]

Table III. Accuracy of RF and MRF for the Wine dataset

<table>
<thead>
<tr>
<th>Number of Trees</th>
<th>Accuracy of RF</th>
<th>Accuracy of MRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.375661</td>
<td>0.557875</td>
</tr>
<tr>
<td>20</td>
<td>0.408931</td>
<td>0.606395</td>
</tr>
<tr>
<td>30</td>
<td>0.442201</td>
<td>0.654915</td>
</tr>
</tbody>
</table>
data is supplied as a input to classification algorithm. Random forest algorithm is a traditional algorithm and facing accuracy problem. We proposed modified random forest algorithm to resolve the problems faced by the random forest algorithm. The experimental work is implemented by using the java language in Net beans IDE. Proposed modified random forest algorithm shows better results when compared with existing algorithm.

REFERENCES


V. CONCLUSION

Data mining is a tool to extract interesting patterns. In our paper we focus on data mining classification technique. SPECT heart dataset with 17 variables and 10299 instances of data of

$\begin{array}{|c|c|c|}
\hline
\text{Value} & \text{Precision} & \text{Recall} \\
\hline
40 & 0.475471 & 0.703435 \\
50 & 0.508741 & 0.751955 \\
60 & 0.542011 & 0.800475 \\
70 & 0.575281 & 0.848995 \\
80 & 0.608551 & 0.897515 \\
90 & 0.641821 & 0.946035 \\
100 & 0.675091 & 0.994555 \\
\hline
\end{array}$

Figure 1. Performance comparison of RF and MRF algorithms

Figure 2. Accuracy Vs. Number of trees in RF and MRF algorithms