International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 6 (2015) pp. 14467-14477

http://www.ripublication.com

An Experimental Analysis of Classification Mining Algorithm For Coronary Artery Disease

P. Ganesan[1], S. Sivakumar[1*], S. Sundar[3]
[1]Department of Mathematics, Anna University
Chennai, Tamil Nadu, India 600 025.
sivaiit79@gmail.com
[2]Department of Mathematics, IIT Chennai.
slnt@iitm.ac.in

Abstract

Data mining means to search the data from large amount of database. Classification is one of the well-known supervised learning techniques in data mining. We introduce three independent mining algorithm Navie Bayes (NB), Support Vector Machine (SVM) and Decision Tree(DT) classifier during the classification to improve precision, recall, f-measure and accuracy rates. These three algorithms, NB, SVM and DT classifier are useful and efficient, has been tested in medical dataset for heart disease and solving classification problem in data mining. In this paper, we compare the three different algorithms and the results indicate that decision tree algorithm has achieved a high accuracy rate of 91.3% and error rate 8.7% out of other algorithms.

Key Words: Data mining, Classification, Decision Tree, SVM

Introduction

Data mining techniques and applications are used in a wide range of fields, including banking, social science, business industries, bioinformatics, weather, forecasting, health care and big data [9,11]. The classification problem has been extensively investigated by the research community. A number of different approaches to build accurate classification have been proposed (e.g Bayes classifier [11,12], Decision tree [14], and Support Vector Machine [13]). In classification, we give a set of example records or the input data, called the test data set, with each record consisting of various attributes.

An attribute can be either a numerical attribute or a categorical attribute. If values of an attribute belong to an ordered domain, the attribute is called a numerical attribute (e.g. Age, Weight, Sports, Sleep, and Drink). A categorical attribute, an another one, has values from an unordered domain (e.g Sex and BP). Classification is

the process of splitting a dataset into mutually exclusive groups, called a class, based on suitable attributes.

In this world, various kinds of heart diseases are one of the most common reasons for death, out of these diseases coronary artery diseases(CAD) is a major kind where 25% of people die all of sudden without any prior symptoms. CAD affects the heart causes severe heart attack in patients. Once the symptoms of the diseases are recognized, previous of adequate first aid would reduce the severity of disease side effects. At present angiography is used to diagnose the disease which is quite expensive. Hence many researchers have began to use data mining for diagnosing CAD.

This paper is organized accordingly: The related works and description of the technical aspects of the used data mining methods in section 2. The introduction of the dataset for heart disease in section 3. The experimental and comparative results in section 4. and finally conclude the paper and future research directions.

Related Works and Methods

This research work is mainly based on three algorithms SRINT, SLIQ and ID3, The algorithms are briefly introduced which are as follows. SPRINT planned classification algorithm called as SPRINT that eliminates all memory limitations that restrict decision-tree algorithm, which proves that planned algorithm is fast and scalable [5].

Supervised Learning In Quest (SLIQ) is a decision tree classifier that can handle both numerical and categorical attributes, It builds compact and accurate trees. It uses a novel pre sorting technique in the tree-growth phase to reduce the cost of evaluating numeric attributes. This sorting procedure is integrated with a breadth first tree growing strategy to enable classification of disk-resident datasets. SLIQ uses a fast sub setting algorithm to determining splits for categorical attributes[4].

Iterative dichotomizer 3 (ID3) which is used to build a decision tree was first developed by Quinlan (1979). It is a top-down approach starting with selecting the best attribute to test at the root of the tree[8]. The selection of the best attribute in ID3 is based on an information theory approach or entropy. Entropy is used to measure how informative a node is. This algorithm uses the criterion of information gain to determine the goodness of a split. The attribute with the greatest information gain is taken as the splitting attribute, and the data set is split for all distinct values of the attribute.

Support Vector Machine

Support Vector Machine (SVM) develops hyper plane in a large dimensional space which can be used for several important data mining and statistical analysis related to classification and other problems. Logically, the best separation is achieved by the hyper plane containing the largest functional margin distance from the nearest training data points of any class. Statistically, with large margin, low generalization error can be achieved for the classifier [13].

Naïve Bayes

Naïve bayes is a data mining technique that shows success in classification in diagnosing heart disease patients [11]. Naïve bayes is based on probability theory to find the most likely possible classifications [12]. This algorithm uses the Bayes formula, which calculates the probability of a data record *X* having the class label c_i:

$$P(c_i / X) = \frac{P(X / c_i)\ P(c_i)}{P(X)}$$

The probability of data record *P(X)*, can be safely eliminated as it does not depend on the label. The class label c_i, with the largest conditional probability value, determines the category of the data record [10].

Decision Tree algorithm

Classification is the most important and familiar technique in data mining. The decision tree is a structure that includes root node, branch and leaf (or) terminal node. Each internal node denotes a condition on attribute, each branch denotes the ending of test and each leaf node holds the class label and counts the value. Decision tree uses a "divide and conquer" technique to split the data into subsets based on the condition. The result of decision tree is in the form of rule-based or tree-based.

Tree Building

An initial decision tree is grown-up in this phase by continually partitioning the training data from a given condition. The training set is split into two or more partitions using an attribute based on the condition. This procedure is repeated recursively until all the data set in each partition belongs to one class [4,5]. A simple building of a decision tree is shown in Figure - 1.

Decision Tree Algorithm

Make Tree (Training Data T)
    Partition (T);
Build Tree (Data set S)
        if (all records in *S* are in same class)
    return;
        for each attribute *A*
        Use best split found to partition S_1 into S_2;
        Partition (S_1);
        Partition (S_2);

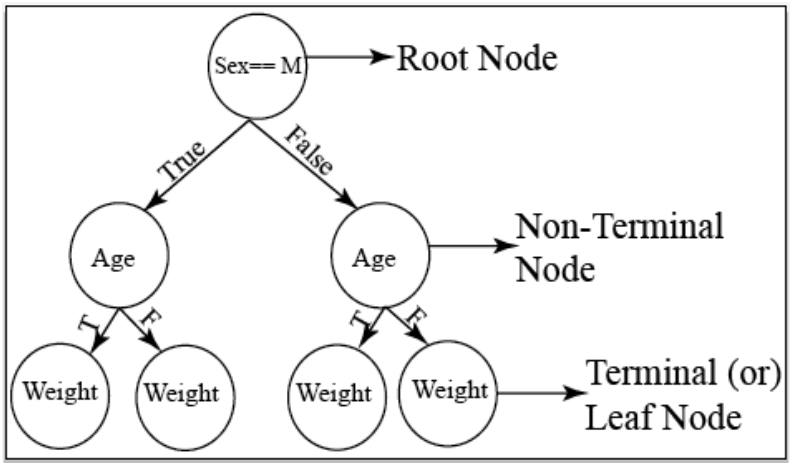

Figure 1: Example of a decision tree

Splitting Points

A splitting point is used to evaluate the "goodness" of the different splits for an attribute. we use the *gini* index, initially proposed in [4,5,7], based on our knowledge with SLIQ and SPRINT. If a data set S contains n classes, gini(S) is defined as $gini(S) = 1 - \sum p_j^2$ where p_j is the relative frequency of class *j* in *S*.

If a split divides S into two subsets S1 and S2, the index divided data $gini_{split}(S)$ is given by $gini_{split}(S) = \frac{n_1}{n} gini(S_1) + \frac{n_2}{n} gini(S_2).$ The benefit of this index is that it requires computation only at the distribution stage of the class values in each of the partitions. To discover the best split point for a node, we search each of the node's attribute lists and calculate split based on that attribute. The attribute containing the split point with the lowest value for the *gini* index is then used to split the node. We used two types of attributes (i) Numerical attribute, a binary split of the form $A \leq v$, where v is a real number, is used for numeric attributes (e.g. Age, Weight, Sports, Drinks). (ii) Categorical attributes, If S(A), is the set of possible values (e.g. BP, Sex).

Heart Disease Dataset

The performance of these three algorithms namely NB, SVM, DT was tested in a medical database for Heart Disease dataset from UCI machine learning repository (available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease) [6]. The data set has 76 features of the attributes. Table- 1 describes the data for heart disease. The medical dataset contains data from reviews conducted among patients, each of which has 14 features. All features can be considered as on indicators of coronary artery disease for a patient. The dataset holds records of the following attributes.

Table 1: UCI Dataset Of Heart Disease

Attributes Name	**Attribute Type**	**Description**
Sex	Discrete	Value 1= Male, Value 0 = Female
Age	Continuous	30-86
Cp = Chest pain type	Discrete	value 1: typical angina, value 2: atypical angina, value 3: non-anginal pain, value 4: asymptomatic.
Chol	Continuous	serum cholestoral in mg/dl
Fbs	Discrete	fasting blood sugar > 120 mg/dl) 1 = true; 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality, 2 =showing probable or define left ventricular hypertrophy by Estes'criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes, 0 = no
oldpeak	Discrete	ST depression induced by exercise relative to rest
Slope Discrete	Discrete	The slope of the peak exercise segment : 1 = up sloping, 2 = flat, 3= down sloping
Ca	Discrete	number of major vessels (0-3) colored by flourosopy
Thal	Discrete	3 = normal, 6 = fixed defect, 7 = reversable defect
Diagnosis	Discrete	Diagnosis classes: 0 = healthy, 1= possible heart disease
Trestbps:	Continuous	resting blood pressure Categorical [High BP, Low BP, Normal BP], this is class values

Experimental and Comparison

In this section, we describe the test database and experimental analysis and the current evaluation results for three algorithms namely NB, SVM, DT classifier.

In this experimental analysis, Navie Bayes, Support vector Machine and Decision tree algorithms performance were compared based on their application in the medical datasets. Rapid miner tool is used in our experiment analysis. The applications of rapid miner software is being used for banking, research, education and weather datasets. It helps in coordinated activities in machine learning, data mining, text mining and web mining. It supports all the mining process to get valid and clear visualization with accurate results, 10 fold cross-validation was applied to the input datasets in the experiment.

Experimental Step Up

A brief description about the classification process by all three algorithms Navie Bayes, Support vector Machine and Decision tree are given below:

Data mining methods also need an evaluation procedure. This procedure is used to verify the models generated by three algorithms namely NB,SVM and DT. Classification techniques can be evaluated using the data labels in a supervised learning methods. The different matrices are used to evaluate the classification algorithm, such as confusion matrices for Precision, Recall, F-measure and Accuracy.

Table 1: Confusion Matrix

	L	H	N
L	tpL	eLH	eLN
H	eHL	tpH	eHN
N	eNL	eNH	tpN

The confusion matrix is shown in Table 1. In the confusion matrix, the diagonal elements are correctly classified data and the rest of elements are incorrectly classified data. Precision is defined as the ratio between the true positive value and both the true positive and false positive values.

Table 2: Confusion Matrix Using Naive Bayesian

	L	H	N	Total
L	2189	89	47	2325
H	428	427	33	888
N	279	72	87	438

The confusion matrices for Naive Bayesion algorithm is shown in Table-2. This classification uses the class values of L-low BP, H- high BP and N-Normal BP. The result from the confusion matrix is discussed for each class as given below

There are 2325 items found are classified into class value for L-low BP, 2189 of these items are exactly classified into class L, 89 of these items are incorrectly classified into class H, finally 47 of these items are incorrectly classified in to class N. There are 888 items found are classified into class value for H-high BP, 428 of these items are exactly classified into class L, 427 of these items are incorrectly classified into class H, finally 33 of these items are incorrectly classified into class N. There are 438 items found are classified into class value for N-Normal BP, 279 of these items are exactly classified into class L, 72 of these items are incorrectly classified into class H, finally 87 of these items are incorrectly classified into class N.

Table 3: Confusion Matrix Using Decision Tree

	L	H	N	Total
L	2202	72	51	2325
H	1	820	67	888
N	90	38	310	438

The confusion matrices for Decision tree algorithm is shown in Table-3. This classification uses the class values of L-low BP, H- high BP and N-Normal BP. The result from the confusion matrix is discussed for each class as given below

There are 2325 items found are classified into class value for L-low BP, 2202 of these items are exactly classified into class L, 72 of these items are incorrectly classified into class H, finally 51 of these items are incorrectly classified into class N. There are 888 items found are classified into class value for H-high BP, 1 of these items are exactly classified into class L, 820 of these items are incorrectly classified into class H, finally 67 of these items are incorrectly classified into class N. There are 438 items found are classified into class value for N-Normal BP, 90 of these items are exactly classified into class L, 38 of these items are incorrectly classified into class H, finally 310 of these items are incorrectly classified into class N.

Table 4: Confusion Matrix Using Support Vector Machine

	L	H	N	Total
L	2015	187	123	2325
H	427	399	62	888
N	237	105	96	438

The confusion matrices for Decision tree is shown in Table-4. This classification uses the class values of L-low BP, H- high BP and N-Normal BP. The result from the confusion matrix is discussed for each class as given below

There are 2325 items found are classified into class value for L-low BP, 2015 of these items are exactly classified into class L, 187of these items are incorrectly classified into class H, finally 123 of these items are incorrectly classified into class N. There are 888 items found are classified into class value for H-high BP, 427 of these items are exactly classified into class L, 399 of these items are incorrectly classified into class H, finally 62 of these items are incorrectly classified into class N. There are 438 items found are classified into class value for N-Normal BP, 237 of these items are exactly classified into class L, 105 of these items are incorrectly classified into class H, finally 96 of these items are incorrectly classified into class N.

Precision

It is used to represent the fraction of retrieved data from connect datasets, that are relevant to the search. *precision* will be used to represent how many instance have been correctly classified in the confusion matrix table (correct classified data is true positive and incorrect classified data is error positive).

$$\Pr ecision = \frac{tpL}{tpL + eHL + eNL}$$

Where tpL is represented as true positive for the class L and eHL and eNL are represented as false positive.

Recall

It is used to represent the fraction of retrieved data from connect datasets, that are relevant to the query that are successful. It is used to find out the ratio between the true positive and both true positive and false positive values.

$$\mathrm{Re} call = \frac{tpL}{tpL + eLH + eLN}$$

Where tpL is represented as true positive for the class L and eLH and eLN are represented as error positive.

F-measure This is evaluated by the harmonic mean between precision and recall.

$$F - Measure = 2 * \frac{\Pr ecision * \mathrm{Re} call}{\Pr ecision + \mathrm{Re} call}$$

Accuracy This is calculated as the proportion of true positive, true negatives and true results from all the given data.

$$Accuracy = \frac{tpL + tpH + tpN}{tpL + eLH + eLN + tpH + eHL + eHN + eNL + eNH + tpN}$$

$$Error\ Rate = 1 - Accuracy.$$

The decision tree classification algorithms runs on the test datasets and process each records. It classifies the records into Low BP, High BP and Normal BP. The results are verified for the above proposed classifier. These results are shown in table 5,6 and 7. The total records are used in the experiment as shown in tables 2, 3 and 4.

Table 5: Confusion Matrix Using Naive Bayesian

	Confusion Matrices			Results			Accuracy %
	L	H	N	Precision	Recall	F-Measure	
L	2189	89	47	75.60	94.20	84.70	74.00
H	428	427	33	72.60	48.08	87.30	
N	279	72	87	52.00	19.80	28.90	

Table 6: Confusion Matrix Using Decision Tree

	Confusion Matrices			Results			Accuracy %
	L	H	N	Precision	Recall	F-Measure	
L	2202	72	51	96.03	94.71	95.33	91.30
H	1	820	67	88.17	92.34	90.19	
N	90	38	310	72.43	70.77	71.64	

Table 7: Confusion Matrix Using Support Vector Machine

	Confusion Matrices			Results			Accuracy %
	L	H	N	Precision	Recall	F-Measure	
L	2015	187	123	75.21	86.66	80.54	68.70
H	427	399	62	57.74	44.93	50.48	
N	237	105	96	34.16	21.91	26.73	

Table 8: Average of The Precision, Recall and F-Measure

Algorithm Names	Results		
	Precision	Recall	F-Measure
Naive Bayesion	66.73	54.02	66.96
Decision Tree	85.54	85.94	85.72
Support Vector Machine	55.70	51.16	52.55

The results are evaluated by the precision, recall, and f-measure and compared the results with three algorithms namely NB, SVM and DT. It shows the better accuracy for the classification as shown in the figure 2. The final result, the decision tree algorithm seems to be good performance for the supervised learning methods.

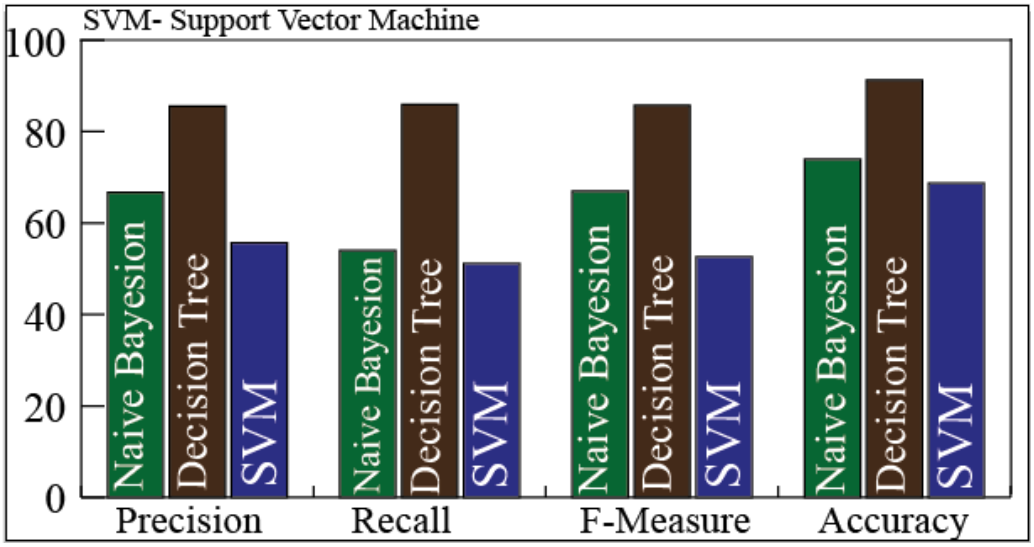

Figure 2: The comparison of classification precision, recall F-measure and accuracy

Conclusion

The decision tree classification is probably the most popular method in data mining and easily to understand the rule. This algorithm to apply in medical database for the

classification of blood pressure evaluation analysis and it is based on rule. In this case study, multiple split points using numerical and categorical attributes. This algorithm is based on SLIQ, it is used a pre-sorting method in the tree growth phase to reduce the cost of evaluating numeric attributes and SRINT that eliminates all memory limitations it restrict decision tree algorithm, which proves that planned algorithm is fast and scalable. Evaluation and analysis results are achieved classified from data sets with different attributes to explain the usefulness of the planned approach to increase the accuracy of the classifier. The experiment and comparison results showed that, Decision tree algorithm has been achieved high accuracy 91.3% and error rate 8.7% out of other algorithms.

In addition, the features used in this study, can be measured with affordable costs and side effect. Hence, applying the proposed approach can identify the CAD state with higher probability and low costs. In future, we aim to consider predicting state of each artery independently. Moreover, it is obvious that true diagnosis of diseased people is more important than true identification of healthy ones. Therefore, another goal to meet is using cost sensitive algorithm to consider this factor. Finally, larger datasets, more feature and also broader data mining approaches, could be used to achieve better and more interesting results.

References

[1] Joachims T., 1998, "Text categorization with support vector machines: Learning with many relevant feature," ECML'98. pp.137-142.

[2] Quinlan J., 1993, "C4.5:Programs for Machine Learning, ". The Morgan Kaufmann.

[3] AI-hegami A. "Pruning Based Interestingness of Mined Classification Patterns. International Arab Journal of Information Technology. Vol. 6, No. 4, pp. 336-343, 2009.

[4] Manish M., Rakesh A., and Jorma R., 1996, "SLIQ: A Fast Scalable Classifier for Data Mining," Int. Conference on Extending Database Technology(EDBT'96), Avignon, France.

[5] John S., Rakesh A., and Manish M., 1996, "SPRINT: A Scalable Parallel Classifier for Data mining, ". proceedings of the 22nd VLDB Conference Mumbai (Bombay), India.

[6] UCI Machine Learning Repository (2013). Available from: http://archive.ics.uci.edu/ ml/datasets.html

[7] Breiman L., Friedman J.H., Olshen R.A., and Stone C.J., 1984, "Classification and Regression Trees, " Wadsworth, Belmont.

[8] Quinlan J R.,1979, " Discovering rules by induction from large collections of examples, " Expert Systems in the Micro Electronic Age, Edinburgh University Press, 168–201.

[9] Quinlan J R., 1996, "Improved use of continuous attributes in C4.5, " Journal of Artificial Intelligence Research 4: 77-90.

[10] Roohallah A., Jafar H., Mohammad J H., Hoda M., Reihane B., Asma G., Behdad B.,and Zahra A S., 2013, " A data mining approach for diagnosis of coronary artery disease, Computer Methods and Programs in Biomedicine, pp.53-61.

[11] Sitar-Taut, V.A., et al.,2009, "Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics.

[12] Wu, X., et al.,2007, "Top 10 algorithms in data mining analysis," Knowl. Inf. Syst.

[13] Bennett K.P., and Blue J.A.,1998, "A support vector machine approach to decision tree," In proceedings IJCNN'98, pp. 2396–2401.

[14] Han j., and Kamber M., 2006, "Data mining : Concepts and techniques," Morgan Kaufmann Pulishers, San Francisco, CA.