2012 International Conference on Software and Computer Applications (ICSCA 2012)
IPCSIT vol. 41 (2012) © (2012) IACSIT Press, Singapore

Study on Classifiers using Genetic Algorithm and Class based Rules Generation

Maragatham G [1] and Lakshmi M [2]

[1] Research Scholar , Computer Science & Engg Dept, Sathyabama University , India , 600 119

[2] Professor & HOD, Computer Science & Engg Dept, Sathyabama University , India , 600 119

Abstract. With widespread use of databases, there exist huge amount of data in the repositories. Efficient decision making from these repositories is one of the main concern in data mining community. Therefore we need to depend on classification techniques for prediction. For summarizations, association rule mining is used. First, initial dataset is preprocessed for missing values. Then the classifiers are applied and their performance accuracy is studied. The data set contains 20 attributes. Choosing relevant attributes, improves classification results and better rule generation for decision making. Therefore we have used wrapper-subset evaluation with Genetic search for subset filtering and the performances of the classifiers such as Naïve Bayes classifier , decision table classifier and simple CART are studied. Since the reduced dataset with naïve bayes approach shows better improvement in classification, the same dataset is used for generating association rules with class based apriori. The experimental results shows the justification of the improved classified output with genetic approach and the generation of relevant class based association rules.

Keywords: Naïve Bayes, Genetic Algorithm (GA), classification, class based apriori.

1. Introduction

Data mining is considered as headway of Information technology area. The data mining is considered as a step by step process for knowledge discovery aspect. Commonly, Tasks in data mining includes descriptive and predictive. In this article we are discussing predictive data mining as well as well descriptive mining . In the former case, the performances of the classifier models are analyzed. In the later case, summarizations are obtained using association rule mining. Generally, a data mining process includes pre-processing step, mining step, post - processing step. In pre-processing stage, data cleaning technique is used for removing or reducing the presence of noise in the data. Also, missing values are treated with respective mean and modes from the training dataset. Since the dataset contains 20 attributes, classification accuracy and rule generation for decision making may be improved by filtering irrelevant attributes. Therefore to address this issue, attribute subset selection methods are used. In our article we have used Wrapper-subset evaluation with Genetic search measure as relevance test for filtering the irrelevant attributes for improving the scalability of the classifier. In the mining step, Naive Bayes classification, Decision table, CART are applied. Finally, the class based association rules are generated and the results are discussed. In this article ,we have considered dataset from UCI – repositories and considered classifiers Naïve Bayes, Decision table and CART. Initially all the classifiers are studied with hepatitis dataset. The study shows the accuracy of the Naive Bayes classifier outperforming the rest of the classifiers in terms of accuracy and error generated. Also , we have compared accuracy of the classifiers with reduced dataset. All the classifiers have shown improvement in classification accuracy and the error aspects. Even then, Naive bayes + GA approach outperforms other classifiers. Finally, the reduced dataset with naïve bayes approach is used for generating association rules with class based apriori algorithm. The details of the study is presented in section 3 and 4.

2. Related work

Mohd Fauzi bin Othman, Thomas Moh Shan Yau [1] have discussed performance study of different classification algorithms such as Bayes network, Radial Basis function, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors algorithm. They have used WEKA tool for study purposed and used

UCI data repository and concluded that the Bayes network classifier is better than the other classifiers in terms of accuracy. The concept of classification techniques in machine learning area is given in [2][3][5][8]. Classification based rule mining is discussed by Stefan Mutter [13]. Algorithms related to Classification, Clustering, Statistical learning, Association analysis, and Link mining are discussed by Xindongwu et al in [4]. The paper states the importance of the algorithms as well as research issues. The WEKA tool for analysis is given in [10] and the data set considered for the study is from [9]. For improving the classification standard filters are used. The most popular filter namely wrapper subset evaluation filter is discussed in [6]. Even-though filters are used for selecting relevant attributes, still for getting optimised classifier output , evolutionary concept – Genetic algorithm [12] [7] can be used.

3. Classification Algorithms

3.1. Classification :

The main task of Classification technique is to predict the value of a user-specified goal. Generally accuracy of any classifier is tested as, how well a classifier classifies unseen data. Initially all the classification algorithms tuned to the training dataset. Then, the classifier model is tested for accuracy with testing dataset. The evaluation criteria of a classifier [3] are measured in terms of speed, Robustness, Scalability, Interpretability etc.

3.2. Need for subset selection:

Feature selection is considered to be the core step to improve the classifier performance. For building robust learning model, Feature selection is required. That is to have optimal set of attributes for better results, subset selection is used. Wrapper subset evaluation with genetic search in [10][12] used. Genetic algorithm is a heuristic search technique for optimization is used as search technique with wrapper subset evaluation for improved model. The wrappers, divides the dataset into subsets and the wrapper evaluates each subset by running a model with the subset. Because of this nature the wrapper subset is computationally expensive.

3.3. Naive Bayes network:

It is a Statistical classifier which works on the principle of Bayes theorm using prior and posterior probabilities.[3]. Given a set of classes $C_1,C_2,...,C_m$. In order to classify a tuple X for a class C_i (i= 1,2...m). The conditional probability of X belonging to a particular class C_i is computed as follows:

$$P(C_i \mid X) > P(Cj \mid X) \quad \text{for all } 1 <= j <= m , j <> i;$$

The class C_i getting maximum value of P(Ci|X) is choosen for X. The analysis of Naive bayes approach is well stated in Table 2,3 & 4.

3.4. Decision table

Decision table[10][3] is one of the predictive classification technique in machine learning environment. In our study, the decision table gives 78.064 % accuracy in classification. While performance of classification is improved by using Wrappersub set evaluation – Filter – genetic search to be 84.51 %. The comparative study shows Naive bayes approach + GA outperforms decision table in terms of classification accuracies.

3.5. Simple CART :

Simple CART [10] , CART - Classification and Regression Trees . It is a technique used in data exploration and prediction , developed by Leo Breiman et al., CART has an inbuilt pruning facility for developing an optimal classifier model. The study shows, the simple CART outperforms decision table classifer in terms of all accuracy aspects. But,overall it is the Naive Bayes that outperforms all other classifiers.

3.6. CAR + Apriori.

Apriori algorithm + CAR [10][12] is a class based association rule mining algorithm for generating association rules. Generally the algorithm works with minimum support and minimum confidence constraints. The class based rules contains the frequent patterns as the rule antecedent and the consequence of the rule contains only the class variable.

4. Experimental Results and Discussion

4.1. Input Attributes

The analysis is done using WEKA 3.6.0 version [10]. The hepatitis [9] dataset contains 155 instances and 20 attributes. Initially missing values are treated with mean and modes from the training dataset. Then all the attributes are used by the classifiers for analysis. Next, by using WEKA's filter the relevant attributes are filtered. The subset evaluation function filters the attributes varyingly with respect to classifiers. For example Table 1 (partial) shows the details of filtered input for the Naive Bayes classifier after application of WEKA's Filter. The classification accuracy is given in Table 2 and Table 3. Also, the confusion matrices for the corresponding Classifiers are given in Table 4.

Table 1: Input Attributes

ATTRIBUTE		CLASS - A	CLASS - B
Gender	Male	1.0	17.0
	Female	33.0	108.0
	Total	34.0	125.0
Steroid	No	21.0	57.0
	Yes	13.0	67.0
	Total	34.0	124.0
Malaise	No	10.0	85.0
	Yes	24.0	39.0
	Total	34.0	124.0

4.2. Classification accuracy

The classification statistics shows the measure of the classifiers on different measures such as correctly classified instances, incorrectly classified instances, kappa statistics and different error measures. The correctly classified instance measure shows the ability of a classifier to correctly classify the data. The percentage of incorrect classification shows the measure of incorrectness of a classifier for the data set.

Among the classifiers, the Naïve Bayes approach with GA shows an accuracy of 89.0323% which outperforms the rest of the classifiers. The kappa statistics [11] is used to evaluate the reliability of any classifier in its classification task. The kappa value of +1 indicates complete agreement between the classification groups, -1 indicates the total disagreement among the groups classified , 0 indicates there exists no agreement in the classification group. Analyzing kappa statistics , NaiveBayes + GA has highest kappa score of 0.6614 rather than Decision tree and CART.The error rates gives the ability of a classifier for classifying incorrect instances.In this aspect Naïve Bayes + GA approach has got less error rate than reported by rest of the classifiers. Applying 10-fold classification on the dataset ,the Table 2 gives the classification details.

Table 2 : Classifier details for 10-fold classification

Statistics	Naïve Bayes	Naïve Bayes + GA	Decision table	Decision table + GA	Simple CART	Simple CART + GA
Correctly Classified Instances	83.371%	89.0323%	78.0645%	84.5161%	78.0645%	84.5161%
Incorrectly classified Instances	16.129%	10.9677%	21.9355%	15.4839%	21.9355%	15.4839%
Kappa Statistics	0.5242	0.6614	0.3305	0.4513	0.1292	0.4513
Mean absolute error	0.17	0.1649	0.2759	0.2637	0.2993	0.258

Root mean squared error	0.3721	0.3137	0.3805	0.3609	0.4224	0.3611
Relative absolute error	51.486%	49.949%	83.556%	79.8531%	90.6287%	78.1265%
Root relative square error	91.8835%	77.4788%	93.9934%	89.1376%	104.3052%	89.1632%

For measuring the classification accuracy - precision and recall are used. Precision is defined as the ratio of number of items correctly classified as true positives to the sum of both true positives and false positives. A precision value of 1.0 for a class A implies that every item labeled as Class A belongs to Class A but does not give detail about the number of items from Class A that were not correctly classified. The measures of true positives and negatives can be got from the confusion matrices Table 4 of the classifiers.

Recall is defined as the ratio of the sum of true positives to the sum of both true positives and false negatives. A Recall value of 1.0 for a class A means every item labeled as Class A belongs to Class A but does not give any information about how many other items were incorrectly labeled as class A. The Table 3 shows the highest precision and recall is supported by Naïve Bayes + GA approach . The assessment from Table 3 shows that the behavior of Naïve Bayes + GA has better results for TP and FP measures.

Table 3 : Detailed Accuracy of the Classifiers by Class - Dataset : Hepatitus

Classifiers	TP-Rate		FP-Rate		Precision		Recall	
	Class-A	Class - B	Class-A	Class - B	Class-A	Class - B	Class-A	Class - B
Naïve Bayes	0.656	0.886	0.114	0.344	0.6	0.908	0.656	0.886
Naïve Bayes + GA	0.719	0.935	0.065	0.281	0.742	0.927	0.719	0.935
Decision table	0.469	0.862	0.138	0.531	0.469	0.862	0.469	0.862
Decision table + GA	0.438	0.951	0.049	0.563	0.7	0.867	0.438	0.95
Simple CART	0.156	0.943	0.057	0.844	0.417	0.811	0.156	0.943
Simpe CART + GA	0.438	0.951	0.049	0.563	0.7	0.867	0.438	0.95

Note : (Class A belongs to the DIE & Class B Belongs to LIVE)

Table 4 : Confusion Matrices

a. Confusion matrix – Naïve Bayes

Class A	Class B
2	11
14	109

b. Confusion matrix – Naïve Bayes + GA

Class A	Class B
23	9
8	115

c. Confusion matrix – Decision table

Class A	Class B
15	17
17	106

d. Confusion matrix – Decision table + GA

Class A	Class B
14	18
6	117

e. Confusion matrix – Simple CART

Class A	Class B
5	27
7	116

f. Confusion matrix – Simple CART + GA

Class A	Class B
14	18
6	117

4.3. Class based Association rule generation:

Class Based rules generally takes the form of antecedent and consequent portions.[10]. The antecedent part contains the frequent patterns (Left hand side of the rule) and the consequent contains only class variable (Right hand side of the rule). These rules are generated with 60% as minimum support and 90% of minimum confidence value. Table 5 shows the output obtained with car + apriori.

Table 5 (Best Rules Found)

```
 1. SPIDERS=no VARICES=no PROTIME= (44.5-100 ) 100 ==> Class=LIVE 97conf:(0.97)
 2. SPIDERS=no ASCITES=no PROTIME= ( 44.5-100 098 ==> Class=LIVE 95 conf:(0.97)
 3. SPIDERS=no ASCITES=no VARICES=no PROTIME=(44.5-100 ) 98 ==> Class=LIVE 95    conf:(0.97)
 4. SPIDERS=no PROTIME=(44.5-100) 101 ==> Class=LIVE 97    conf:(0.96)
 5. ASCITES=no BILIRUBIN=(0.3-1.65) PROTIME=(44.5-100) 101 ==> Class=LIVE 96    conf:(0.95)
 6. SPIDERS=no VARICES=no 107 ==> Class=LIVE 101    conf:(0.94)
 7. SPIDERS=no ASCITES=no VARICES=no 105 ==> Class=LIVE 99    conf:(0.94)
 8. VARICES=no BILIRUBIN=(0.3-1.65) PROTIME=(44.5-100 )101 ==> Class=LIVE 95 conf:(0.94)
 9. LIVER_BIG=yes ASCITES=no PROTIME=(44.5-100) 100 ==> Class=LIVE 94 conf:(0.94)
10. BILIRUBIN=0.3-1.65 PROTIME=(44.5-100) 107 ==> Class=LIVE 100    conf:(0.93)
```

5. Conclusion

In this article, we have used the Data mining tool weka[10] to study the accuracy of classifiers and rule generation. The classifier performances are analyzed using GA approach. The study shows that the Naive Bayesian approach shows significant improvement in classification. Therefore, it is proved that Naive Bayesian approach outperforms , the decision tree classifier and CART classifier interms of both the classification accuracy and errors. For rule generation, the reduced dataset used in Naive bayes approach is used by class based apriori. These results are more informative and supportive for decision making.

References

[1] Mohd fauzi bin Othman, Thomas moh shan Yau, "Comparison of Different Classification Techniques using weka for Breast cancer", Vol 15, pp 520-523,2007.

[2] N. Friedman, D. Geiger , M. Goldszmidt, "Bayesian Network Classifiers", Machine learining,29,131-163 (1997), Kluwer Academic Publishers.

[3] Jiawei Han and Micheline Kamber, "Data Mining- Concepts and Techniques", Second edition, Elsevier Publicaitons.

[4] Xindongwu,VipinKumar,J.Ross Quinlan et al, "Top 10 algorithms in data mining ", Knowledge Information Systems (2008)14:1-37, DOI 10.1007/s10115-007-0114-2

[5] Lanley. P., Iba W and Thomas K, "An analysis of Bayesian Classification", Proceeding s of 10th National Conference of Artificial Intelligence, AAAI Press, Stanford 1992, pp 223-228.

[6] Kohavi R and John G, "Wrappers for feature subset selection", Artificial Intelligence journal, Special issue on relevance , 97(1-2),1997, pp.273-324.

[7] Alex.A. Freitas , "A survey of evolutionary algorithms for data mining and knowledge discovery ",Springer verlag ,Newyork, ISBN 3-540-43330-9,2003.

[8] Jiawei Han and Micheline Kamber , " Data Mining - concepts and Techniques", II Edition , Elesevier Publications.

[9] Data set: UCI http://archive.ics.uci.edu/ml/datasets

[10] Weka Data Mining tool – open source data mining tool : http://www.cs.waikato.ac.nz/~ml/weka.

[11] Kappa Measure : http://www.dmi.columbia.edu/homepages/chuangi/kappa

[12] Daniel T.Larose, " Data mining methods and models", Wiley – India Edition (2006)

[13] Stefan Mutter, "Classification using Association rules", A thesis of Diploma of computer science, Univeristy of Freiburg, Hamilton, NewZealoand, Aotearoa, 11th march 2004.