IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 8, Issue 5 (Jan. - Feb. 2013), PP 75-78
www.iosrjournals.org

A Survey on Classification and Rule Extraction Techniques for Datamining

Tulips Angel Thankachan[1], Dr. Kumudha Raimond[2]
[1](PG Scholar/ Department of computer Science, Karunya University, Coimbatore)
[2](Professor/ Department of computer Science, Karunya University, Coimbatore)

Abstract : *Classification is a data mining function that assigns similar data to categories or classes. The main goal is to accurately predict the class for each data. Different classification algorithms such as C4.5, k-nearest neighbor (KNN) classifier, Naive Bayes, SVM (Support Vector Machine), Apriori, and AdaBoost have been used for data mining applications. This paper provides a survey of different classification algorithms for data mining applications.*

Keywords: *Classification, Classifier, Data Mining, Rule Extraction, class imbalance problem*

I. INTRODUCTION

One of the common tasks performed in machine learning and knowledge discovery is classification. It includes assigning a decision class label to a set of unclassified objects. These objects are described by a set of attributes. For inducing various forms of classification knowledge, learning algorithms are applied over the unclassified objects. And using this knowledge the objects can be classified and the classification system is called classifier. Classifier is used to classify the datasets into various categories or classes based on its features. The typical measure used for evaluating classifiers is the classification accuracy.

A survey is conducted to compare various methods for classifying datasets and also for extracting more optimized rules for those datasets. Various approaches are used for classification of datasets and rule extraction. As there are numerous techniques for classification and rule extraction, the categorization of these techniques based on their application is necessary. So, the objective is to classify the various classification and rule extraction techniques and to find out their merits and demerits.

II. CLASSIFICATION OF MULTICLASS LEARNING PROBLEM

In machine learning the multiclass learning problem is the classification of various instances or datasets into more than two classes. Kemel Polat et al. [1] proposed a novel hybrid classifier to classify multi-class learning problems. The classifier is based on C4.5 decision tree classifier and one-against-all approach. This classifier has been used on the multi-class learning problems such as dermatology, image segmentation, and lymphography datasets. The main aim is to improve the classification accuracy. The performance evaluation methods such as sensitivity, specificity and 10-fold cross validation are used. Based on these methods 96.17%, 95.17%, 87.95% classification accuracies are obtained for the above datasets respectively and the results show that the proposed method has produced very promising results in the classification of multi-class learning problems. It can be used for pattern recognition applications.

III. CLASSIFIERS FOR CLASSIFICATION AND RULE EXTRACTION

Jerzy Stenfnowski [4] proposed two approaches to examine the application of the rule induction algorithm called MODLEM, the bagging approach and the n^2-classifier. The bagging approach is that it combines homogeneous classifiers generated from learning problems. The n^2-classifier is a specialized approach used to solve multi-class learning problems by using a set of binary classifiers. These two approaches are used to increase the classification accuracy and to evaluate the performance of multiple classifiers. The result shows that the MODLEM algorithm is efficiently used with multiple classifiers.

Miguel Rodriguez [7] proposed a new method of data distribution in computer networks called Efficient Distributed Genetic Algorithm for classification Rule extraction in data mining (EDGAR). This has been done by the spatial partitioning of the search space into several semi-isolated nodes. Based on this proposed classifier, the first applicable rule from the ordered set of rules returns the assigned class. The main advantage of this proposed classifier is that it responds quickly for large number of rules.

Jacek Jelonek [8] proposed a method called n^2-classifier, to solve the multi-class learning problems. Each pair of classes is trained based on $(n^2-n)/2$ binary classifier. All the experiments are carried out on typical benchmark datasets. Using stratified version of 10-fold cross validation, the classification accuracy is estimated. It is realized that neural network performs well compared to n^2-classifier.

For reducing the dimensionality in case-based reasoning (CBR) classifiers, Maria Salamo [14] investigated feature selection based on rough sets. For the uncertainty management, the rough set theory seems to be an effective tool for data mining. There are two central contributions in this paper: three strategies are developed for feature selection and for estimating the relevance of attributes based on rough set theory, several measures are proposed. Feature selection and instance selection are the main problems faced by CBR. The proposed system is applicable across a wide range of learning algorithms.

S. Dehuri et al. [16] proposed a new algorithm called Elitist Multi-Objective Genetic Algorithm (EM`OGA) for mining classification rules. It is used on large databases. This proposed algorithm emphasized on comprehensibility, interestingness of the rules and predictive accuracy. The simple GA and EMOGA ran for 500 generations over the nursery, zoo and adult datasets. The whole dataset is divided into training and testing set. Twofold cross validation has been applied over the training and testing sets. The experimental result confirms that the rule discovery algorithm has a clear edge over simple genetic algorithm.

Matthew N. Anyanwu et al. [17] presents an evaluation of serial implementations of the decision tree algorithms to categorize the commonly used algorithms. Based on the volume of data, decision tree classification algorithm can be implemented in a serial or parallel manner. Statlog dataset are used to evaluate the performance of the serial decision tree algorithms. The serial implementations of decision tree algorithms are memory resident, fast and easy to implement compared to parallel implementation of algorithms. The results show that the serial implementation of decision tree algorithms such as C4.5 and SPRINT algorithms have better improvement in classification accuracy compared to other decision tree algorithms.

For mining classification rules from large databases Basheer M. Al-Maqaleh et al. [10], proposed a GA based approach. A classification rule is of the form: "If P Then D" and it is a high-level knowledge representation. These forms of representation of rules are highly comprehensible for the user. The proposed algorithm used to mine the classification based on the evaluating measures such as accuracy, comprehensibility and coverage (completeness). Experimental results show that the proposed GA is suitable for classification and rule mining and it shows higher classification accuracy.

To improve the classification accuracy, Pushpalatha Pujari [11] proposed the feature selection technique and ensemble model. In feature selection technique, the subsets of relevant features are selected from datasets to build robust learning models. Ensemble model is used to improve the accuracy of classification by combining the prediction of multiple classifiers. The performances are measured using Receiver Operating Characteristics (ROC) and Gain chart. Based on the experiments performed over various datasets, ensemble model performs well for classification.

For overcoming the limitations of traditional KNN algorithm, N. Suguna et al. [19] proposed a new algorithm. In this proposed algorithm GA is combined with KNN algorithm called Genetic KNN (GKNN). This algorithm has been performed over five datasets collected from UCI data repository and its performance is compared with SVM, CART (Classification and Regression Tree) and traditional KNN. From the comparison it proved that this algorithm reduces the complexity of traditional KNN and also it improves the classification accuracy.

IV. COMBINATION OF CLASSIFIERS

Jerzy Stefanowski et al. [3] proposed an experimental study of using the rule induction algorithm MODLEM in the multiple classifier scheme called combiner. The main aim of this proposed method is to improve the classification accuracy. Experiments are done over various benchmark datasets and found that the combiner classifier is having higher classification accuracy than the single classifiers.

In Jerzy Stefanowski et al. [6] , another combiner method is proposed. In this method two rough sets based filtering approaches combined with rule based classifiers. It is mainly suited for handling imbalance datasets. This paper shows a higher improvement in sensitivity and gain.

Based on the various characteristics of input data, S. Y. Sohn et al. [13] proposed a method to compare the performance of classifier methods using logistic regression. The combination method includes modified random subspace method, bagging, parametric fusion, classifier selection. Taguchi design has been used for typically unknown combination function among input variables. Monte carlo simulation is used to improve the classification accuracy of classifier combination methods based on various data characteristics.

Bikash Kanti Sarkar et al. [20] present a hybrid classifier called DTGA (Decision Tree and Genetic Algorithm) which is a combination of C4.5 and GA as GA is more suitable for getting more optimized solutions. The experiments are done on UCI repository datasets. In DTGA, the dataset is first passed to C4.5 to generate rules. After discretizing the rules, GA is applied to refine the rules. This proposed model increases the classification accuracy and able to classify imbalance datasets.

V. CLASSIFICATION OF LARGE/SMALL-DISJUNCT RULES

Deborah R. Carvalho et al. [2] proposed a hybrid decision tree/ GA method. In this hybrid approach, two specifically designed GA algorithms are used for discovering rules of examples belonging to small disjuncts and conventional decision tree algorithm are used for producing rules of examples belonging to large disjuncts. The advantage of this hybrid method is that they are flexible and robust. This hybrid method is having two versions, C4.5/GA-Small and C4.5/GA-Large-SN. The performance of both the versions are compared with C4.5 and "double C4.5". And the better results are shown by the C4.5/GA-Large-SN and it has been considered as the best solution for small disjunct problems.

Deborah R Carvalho et al. [9] addresses to discover rules to predict the class based on the value of the attributes. These rules decrease the classification accuracy and they are error prone. For overcoming all the limitations, a hybrid decision-tree/ GA approach is proposed. And the performance result of this hybrid model has been compared with three versions of C4.5 over eight domain data sets. In all the comparisons the hybrid model achieves better predictive accuracy.

For discovering small-disjunct rules in the form "If P Then D", a classification algorithm based on Evolutionary Algorithm (EA) has been proposed by Basheer M Al-Maqaleh et al. [12]. The experiments are carried out over several UCI dataset repositories. From large datasets, the small disjunct rules are generated using this successful application of GA. It is having appropriate crossover and mutation operators, flexible chromosome encoding, and suitable fitness function. The results show that the proposed algorithm is much more efficient for better rule extraction.

VI. CLASSIFICATION OF IMBALANCE PROBLEMS

Nathalie Jakowicz [5] proposed to find out the type of discrepancy that is most destructive for a standard classifier that expects balanced class distributions. In this paper the comparison of various methods are done. The schemes for handling class imbalance problems are: re-sampling, down-sizing and learning by recognition. Both the re-sampling and down-sizing methods are very efficient especially as the sensitivity gets larger. The recognition based approach has higher performance when applied on the majority class. However, down-sizing and re-sampling approaches are more effective than recognition based approach.

Comparisons of data mining approaches to deal with imbalanced data sets are presented by Jerzy W. Grzymala-Busse [15]. The foremost approach is based on saving the unique rule set, induced by the Learnable Evolution Model 2 (LEM2) algorithm, and altering the rule potency for all rules for the minor class during classification. In the next approach, rule set was divided: the rule set for the superior class was induced by LEM2, whereas the rule set for the minor class was induced by EXPLORE. Results of these approaches shows that both increases the sensitivity compared to the original LEM2.

VII. COMPARISON

Table 1: Comparison of different classification and rule extraction techniques

Classification	Merits	Demerits
Classification of multi-class learning problem	•Higher classification accuracy, sensitivity and specificity	• Not able to classify imbalance datasets
Different methods for classification and rule extraction	• It does not increase the computation time. • Bagging approach increases classification accuracy. • n^2-classifier increases accuracy of multiple classes. • EDGAR shows considerable speedup. • Provides higher predictive accuracy. • Coupling strategy improves classification accuracy. • CBR retrieves the most relevant information in large datasets.	• Losing simple and easy interpretable structure of knowledge. • It does not perform well in some cases. • Not suitable for multi class classification problems. • CBR system is sensitive to noisy and unreliable data. • Decrease in classification accuracy.
Combination of classifiers	• Can able to classify all dataset problems. • Led to greater increase in sensitivity and gain. • Performs well and provides higher classification accuracy.	• Combiner strategy did not improve classification accuracy compared with single classifiers. • Proposed approach insufficient for detecting all unsafe cases. • Classifier combination methods may lose the information.

Classification of Large/Small-Disjunct rules	• Hybrid model provides higher accuracy. • Good solution for small-disjunct problems.	• Hybrid model is slower than pure C4.5 model. • Increase in computational time. • For processing larger datasets it takes more processing time.
Classification of imbalance problems	• Sensitivity increases with complexity of domain. • Can use large datasets. • Increase in gain compared to original LEM2. • Less expensive.	• It is not possible to use on different types of imbalance problems. • Keeping weak rules also.

VIII. CONCLUSION

The comparison of various classification and rule extraction methods are done in this paper. From this survey it is clear that the classification and rule extraction techniques can be applied for any type of dataset such as imbalance, multi-class etc. Each method is having its own benefits and drawbacks. The main advantage among all the classification is that, it increases the classification accuracy. In many of the research works such as [1], [3], [4], [6], [8], [11], [13], [14] and [15], optimization technique was not used. In Hybrid GA-based classifier [20], GA can be considered as the best optimization technique for improving the accuracy of classification. Specific datasets are used for performance evaluation in the works of [1], [19] and [11]. But in Hybrid GA-based classifier [20], the performance is evaluated using many general data sets. There is an improvement in accuracy in [2], [3] and [7]. Compared to that, hybrid GA-based classifier model, it is having good improvement in classification accuracy. In some technique, they can classify imbalance datasets [6], [7] and in some can solve multiclass classification problem [1], [3],[4],[8],[11]. But the hybrid GA-based classifier [20] can classify the imbalance dataset and also can handle multiclass classification problem. In most of the works data sampling is not used. But [19], [6] and [2] used random sampling. Two types of data sampling techniques are used in [20]. From these comparisons of various methods of classification and rule extraction techniques, it is clear that the hybrid model with optimization algorithm improves the classification accuracy.

REFERENCES

[1] Kemal Polat and SalihGunes , "A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems" , *Expert Syst. Appl.*, vol. 36, no. 5, pp. 1587-1592, Jul. 2007.

[2] Deborah R. Carvalho and Alex A. Freitas, "A Hybrid Tree/Genetic Algorithm Method for DataMining", *Applied soft computing*, vol. 35, pp. 650-673, 2005.

[3] Jerzy Stefanowski and SlanwomirNowaczyk, "An Experimental Study of Using Rule Induction Algorithm in Combiner Multiple Classifier", *IISN 0973-1873* vol. 2, no. x pp. xxx-xxx 2006.

[4] Jerzy Stefanowski, "The bagging and n² - classifiers based on rules induced by MODLEM", *Pattern Recognition ,* 2002.

[5] Nathalie Japkowicz, " The Class Imbalance Problem: Significance and Strategies", *computers in biology and medicine*, 2006.

[6] Jerzy Stefanowski and SzymonWilk, "Combining rough sets and rule based classifiers for handling imbalanced data", 2001.

[7] Miguel Rodriguez, Diego M. Escalante and Antonio Peregrin, "Efficient Distributed Genetic Algorithm for Rule extraction", *Applied soft computing*, vol. 36, pp. 733-743 jan 2010.

[8] JacekJelonek and Jerzy Stefanowski, "Experiments on solving multiclass learning problems by n²- classifier", 2004.

[9] Deborah R.Carvalho and Alex A. Freitas, "A Genetic Algorithm for Discovering Small- Disjunct Rules in Data Mining", *Advances in artificial intelligence*, 2007.

[10] Basheer M. Al-Maqaleh and Hamid Shahbazkia, "A Genetic algorithm for Discovering Classification Rules in Data Mining", *International Journal of Computer Applications*, vol. 41, no.18, pp. 40-44, march 2012.

[11] PushpalataPujari and JyotiBala Gupta, "Improving Classification Accuracy by Using Feature Selection and Ensemble Model", *International Journal of Soft Computing and Engineering*, vol. 2, pp. 380-386, may 2012.

[12] Basheer M. Al-Maqaleh, Mohammed A. Al-Dohbai and Hamid Shahbazkia, "An Evolutionary Algorithm for Automated Discovery of Small-Disjunct Rules", *International Journal of Computer Applications*, vol. 41, no.8, pp. 33-37, march 2012.

[13] S.Y Sohn, H.W. Shin, "Experimental study for the comparison of classifier combination methods", *Pattern Recognition*, vol.40,pp. 33-40, june 2007.

[14] Maria Salamo, Maite Lopez-Sanchez, "Rough set based approaches to feature selection for Case-Based Reasoning classifiers", *Pattern Recognition Letters*, vol.32, pp. 280-292, sep 2011.

[15] Jerzy W. Grzymala-Busse, Jerzy Stefanowski and SzymonWilk, "A Comparison of Two Approaches to Data Mining from Imbalanced Data", *Springer*, pp.757-763, 2004.

[16] S. Dehuri, S. Patnaik, A. Ghosh, R. Mall, Application of elitist multi-objective genetic algorithm for classification rule generation, Applied soft computing, vol. 8, pp. 477–487, march 2008.

[17] Matthew N. Anyanwu, Sajjan G. Shiva, Comparative Analysis of Serial Decision Tree Classification Algorithms, International Journal of Computer Science and Security, (IJCSS) Vol. 3, no.3, pp. 230-240, 2004.

[18] T.G. Dietterich, Ensemble methods in machine learning, in: Proceedings of the*1st International Workshop on Multiple Classifier Systems*, LNCS vol. 1857,Springer Verlag, 2000, pp. 1–15.

[19] N.Suguna and Dr.K.Thanushkodi, "An Improved k- Nearest Neighbour Classification Using Genetic Algorithm", *International Journal of Computer Science Issues*, vol. 7, no.2, pp. 18-21, july 2010.

[20] Bikash Kanti Sarkar, Shib Sankar Sana, Kripasindhu Chaudhuri, "A genetic algorithm based rule extraction system", *Applied soft computing*, vol.12, pp. 238- 254, September 2011.