International Journal of Scientific Research & Engineering Trends
Volume 1, Issue 6, Nov.-2015, ISSN (Online): 2395-566X

Review: Efficient and Scalable Multiple Class Classification

Esha Shrivastava*, Yogdhar Pandey
Department of Computer Science & Engineering
Sagar Institute of Research & Technology
*Email: shrivastavaesha@gmail.com

***Abstract* – Classification plays an important role in data mining and the need for building classifiers across multiple databases is driven by applications from various domains. Examples include market basket transaction data from different branches of a whole sale store, network intrusion detection, and molecular genetic data analysis. To perform data mining from multiple databases, the traditional way was to integrate all the databases, and then apply the adequate algorithm. However, the huge dataset after integrating will be difficult to deal with. Therefore we need a fundamentally different approach for multi-database mining. The main idea of this approach is making bridges across the multiple databases with some useful links, in order to build the data mining model.**

***Keywords* – Data Mining, Classification, Clustering, Association Rule, Bee Colony Approach.**

I. INTRODUCTION

Classification [1] is an important subject in data mining and machine learning, which has been studied extensively and has a wide range of applications. Classification based on association rules, also called associative classification, is a technique that uses association rules to build classifier. Generally it contains two steps: first it finds all the class association rules (CARs) whose right-hand side is a class label, and then selects strong rules from the CARs to build a classifier. In this way, associative classification can generate rules with higher confidence and better understandability comparing with traditional approaches. Thus associative classification has been studied widely in both academic world and industrial world, and several effective algorithms [2, 3] have been proposed successively. However, all the above algorithms only focus on processing data organized in a single relational table. In practical application, data is often stored dispersedly in multiple tables in a relational database. Simply converting multi-relational data into a single flat table may lead to the high time and space cost, moreover, some essential semantic information carried by the multi-relational data may be lost. Thus the existing associative classification algorithms cannot be applied in a relational database directly. We propose a novel algorithm, CMAR, for associative classification which can be applied in multi-relational data environment. The main idea of CMAR is to mine relevant features of each class label in each table respectively, and generate strong classification rules. By relevant features, we mean two kinds of frequent close item sets: single table item sets in the target table and cross table item sets in non-target tables. Experiment results show that the above two kinds of item sets have contained sufficient relevant features of class labels. Then we breadth-firstly generate strong classification rules from these item sets with a pruning strategy used in this step. After that, a classifier can be easily built to predict unseen objects' class labels.

II. MOTIVATION

Association rule mining algorithms are often compared using time complexity. That is an important issue of the mining process, but the quality of the resulting rule set is ignored. On the other hand there are approaches to investigate the discriminating power of association rules and use them according to this to solve a classification problem. This research area is called classification using association rules. It has to deal with a large number of rules. Therefore, rule selection and rule weighting are essential for these approaches in classification. An important aspect of classification using association rules is that it can provide quality measures for the output of the underlying mining process. The properties of the resulting classifier can be the base for comparisons between different association rule mining algorithms. A certain mining algorithm is preferable when the mined rule set forms a more accurate, compact and stable classifier in an efficient way. The introduction of this quality measures particularly the accuracy of the classifier kills two birds with one stone. First, in this thesis we are interested in the comparison of the quality of different mining algorithms. Therefore, we use classification using association rules. Secondly, classification using association rules can be improved itself by using a mining algorithm that prefers highly accurate rules. In this thesis we compare different rule mining strategies and different approaches to classification using association rules. We present the impact of different mining algorithms on the resulting classifiers as well as the impact of different classification schemes.

III. OBJECTIVE

There are some limitation and problem of classification algorithm. Now recently researcher choose association classification for classification algorithm and used genetic algorithm for the optimization of classification rate of association classification. In our case the results are improved because the genetic algorithm is a heuristic function. The heuristic function gives an optimal result. Now we adopted evolutionary optimization of classification of association rule with the help of bee colony optimization.

IV. LITERATURE SURVEY

Xiao-Lin Li [4] Probabilistic relational models (PRMs) extend the Bayesian network representation to incorporate a much richer relational structure. Existing probabilistic relational model (PRM) learning approaches based on search and scoring usually perform a heuristic search for the highest scoring structure. In this paper, they propose the maximum likelihood tree based immune binary particle swarm optimization (MLT-IBPSO) method to learn structures of PRMs from relational data. First, a maximum likelihood tree (MLT) is generated from the data sample, and a population is created according to the MLT. Then, immune theory is combined with particle swarm optimization (PSO) for searching the structures. As a result, the probabilistic structure is learned based on the proposed method. Experiments show that the MLT-IBPSO method can learn structures from relational data effectively.

Bahareh Bina[5] An important task in multi-relational data mining is link-based classification which takes advantage of attributes of links and linked entities, to predict the class label. The relational Naive Bayes classifier exploits independence assumptions to achieve scalability. They introduce a weaker independence assumption to the effect that information from different data tables is independent given the class label. The independence assumption entails a closed-form formula for combining probabilistic predictions based on decision trees learned on different database tables. Logistic regression learns different weights for information from different tables and prunes irrelevant tables. In experiments, learning was very fast with competitive accuracy.

Geetha Manjunath [6] Practical usage of machine learning is gaining strategic importance in enterprises looking for business intelligence. However, most enterprise data is distributed in multiple relational databases with expert-designed schema. Using traditional single-table machine learning techniques over such data not only incur a computational penalty for converting to a flat form (mega-join), even the human-specified semantic information present in the relations is lost. In this paper, we present a practical, two-phase hierarchical meta-classification algorithm for relational databases with a semantic divide and conquer approach. They propose a recursive, prediction aggregation technique over heterogeneous classifiers applied on individual database tables. The proposed algorithm was evaluated on three diverse datasets, namely TPCH, PKDD and UCI benchmarks and showed considerable reduction in classification time without any loss of prediction accuracy.

Tahar [7] In this paper author present an improved decision tree classification algorithm DTHR for multiple relational databases. DTHR can perform accurate classification with data stored in multiple databases. The Support Vector Regression model is used in predicting the usefulness of links through the databases. To perform accurate classification, DTHR adopts a low inter-data base communication strategy which actions with high benefit-cost ratio. The experiments performed on both real and synthetic databases showed that DTHR compared with previous approaches, achieves high accuracy in nearly the same running times. Although our experimental results have been encouraging, there remain many possibilities for future work. Currently DTHR uses SVR to build the prediction model based on some properties of links. It is interesting to find more properties to improve usefulness of links. Although DTHR performs a quite broad search to build the decision tree, it is interesting to study other classification approaches like SVM, Neural Networks and naive Bayes classification on multiple relational databases to achieve better accuracy and speed up DTHR.

Marko Debeljaka [8] Nearly three-quarters of the genetically modified maize (the insect resistant type MON 810, also called Btmaize) produced in the EU are cultivated in Spain, where the share of Bt maize cultivation in some regions(Catalonia) is very high (above 70%). In order to ensure coexistence with the production of conventional maize and satisfy the 0.9% EU threshold for adventitious presence of authorized genetically modified (GM) material in conventional (non-GM) maize crops, a set of preventive coexistence measures must be applied. These measures usually include the setup of large and fixed isolation distances, pollen barriers, flowering coincidence, crop rotation and other measures, which are very hard to fulfill in a multi-field setting. Basic empirical and modeling studies that explore the feasibility of coexistence between GM and non-GM crops focus on pair-based interactions between fields while multi-field studies build upon them, attempting to consider the complexity of gene flow under crop management practices.In this study, we use the methodology of relational data mining (which can take into account several coexistence measures at the same time) to predict gene flow from GM to non-GM maize fields under multi-field crop management practices at a local scale. The approach extends the pair-based assessments of out-crossing rate by considering all neighboring fields within the entire study area, along withthe farming practices applied to them. The estimation of the out-crossing rates is performed by using a

PostgreSQL relational database that is analyzed with the algorithm TILDE for building relational classification trees. In building the trees, TILDE explores the relations describing spatial aspects, maize flowering and crop management practices for the 400 ha maize oriented production area Pla de Foixà in Catalonia, Spain, in the period 2004–2006. Our approach proposes a new methodology to predict the level of adventitious presence on a multifield setting, where the influence of more than one GM field is considered at the same time. The Structure of the obtained models can be used in the design of coexistence measures of the second generation, which should not be used individually but treated as synergetic coexistence measures, offering different alternatives to achieve a particular coexistence threshold (e.g., 0.9%, 0.45%, or 0.1%). The possibility to consider multiple measures simultaneously makes farmers more flexible in their management decisions as compared to the rigid use of isolation distance only, which is currently the most commonly recommended coexistence measure.

Dewan Md. [9] In this paper, they introduce two independent hybrid mining algorithms to improve the classification accuracy rates of decision tree (DT) and naïve Bayes (NB) classifiers for the classification of multi-class problems. Both DT and NB classifiers are useful, efficient and commonly used for solving classification problems in data mining. Since the presence of noisy contradictory instances in the training set may cause the generated decision tree suffers from over fitting and its accuracy may decrease, in our first proposed hybrid DT algorithm, we employ a naïve Bayes (NB) classifier to remove the noisy troublesome instances from the training set before the DT induction. Moreover, it is extremely computationally expensive for a NB classifier to compute class conditional independence for a dataset with high dimensional attributes. Thus, in the second proposed hybrid NB classifier, we employ a DT induction to select a comparatively more important subset of attributes for the production of naïve assumption of class conditional independence. In future work, other classification algorithms, such as naïve Bayes tree (NBTree), genetic algorithms, rough set approaches and fuzzy logic, will be used to deal with real-time multi-class classification tasks under dynamic feature sets.

Recent research having lower predictive accuracy which lead to tends, combine existing [1] log-linear model with probabilistic techniques. While a search for informative aggregate features is computationally expensive, when it succeeds, the new aggregate features can increase the predictive accuracy. There are several possibilities for a combined hybrid approach. (i) Once good aggregate features are found, they can be treated like other features and used in a decision tree. (ii) A simple decision forest [2] is fast to learn and can establish a strong baseline for evaluating the information gain due to a candidate aggregate feature. (iii) The regression weights can be used to quickly prune uninformative join tables with or small weights, which allows the search for aggregate features to focus on the most relevant link paths. Whereas in [9] a hybrid mining algorithms to improve the classification accuracy rates of decision tree (DT) and naïve Bayes (NB) classifiers for the classification of multi-class problems but it's don't have any genetic algorithms, rough set approaches and fuzzy logic, be used to deal with real-time multi-class classification tasks under dynamic feature sets.

V. CONCLUSION

There are some limitation and problem of classification algorithm. now we choose association classification for classification algorithm and we used bee colony optimization algorithm for the optimization of classification rate of association classification. In our case the results are improved because the bee colony optimization is a heuristic function. The heuristic function gives an optimal result. Whereas naive Bayes approaches apply over historical data and give better result. Multiple relational classification algorithm modified by bee colony optimization algorithm, improve the predictive accuracy rate of classification in comparison of naïve Bayes classifiers for multi-class classification tasks

REFERENCES

[1] Yingqin Gu1,2, Hongyan Liu3, Jun He1,2, Bo Hu1,2 and Xiaoyong Du1,2 "A Multi-relational Classification Algorithm based on Association Rules" pp.4-9 2009 IEEE.

[2] W. Li, J. Han, and J. Pei, "CMAR: Accurate and efficient Classification Based on Multiple Class-Association Rules", Proceedings of the ICDM, IEEE Computer Society, San Jose California, 2001, pp. 369-376.

[3] X. Yin, and J. Han, "CPAR: Classification based on Predictive Association Rules", Proceedings of the SDM, SIAM, Francisco California, 2003.

[4] Xiao-Lin Li , Xiang-Dong He "A hybrid particle swarm optimization method for structure learning of probabilistic relational models" in transaction of Elsevier Information Sciences 283 (2014) 258–266

[5] Bahareh Bina, Oliver Schulte , Branden Crawford, Zhensong Qian, Yi Xiong "Simple decision forests for multi-relational classification " in transaction of Elsevier Decision Support Systems 54 (2013) 1269–1279

[6] Geetha Manjunath , M. Narasimha Murty , Dinkar Sitaram "Combining heterogeneous classifiers for relational databases" in transaction of Elsevier Pattern Recognition 46 (2013) 317–324

[7] Tahar Mehenni , Abdelouahab Moussaoui "Data mining from multiple heterogeneous relational databases using decision tree classification" in transaction of Elsevier Pattern Recognition Letters 33 (2012) 1768–1775

[8] Marko Debeljaka , Aneta Trajanova, Daniela Stojanovaa, Florence Leprincec, Sa D zeroski "Using relational decision trees to model out-crossing rates in a multi-field setting" in Ecological Modelling 245 (2012) 75– 83

[9] Dewan Md. Farid , Li Zhang "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks" in transaction of Elsevier of Expert Systems with Applications 41 (2014) 1937–1946

[10] Rajdev Tiwari, Manu Pratap Singh "Correlation-based Attribute Selection using Genetic Algorithm" International Journal of Computer Applications (0975 – 8887) Volume 4– No.8, August 2010.

[11] Kalyanmoy Deb, "Introduction to Genetic Algorithms", Kanpur Genetic Laboratory (Kangal), Depart of Mechanical Engineering, IIIT Kanpur 2005.

[12] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth "From Data Mining to KDD in Databases" pp. 0738-4602 1996.

[13] Xun Zhu1, Hongtao Deng2, Zheng Chen3 "A Brief Review On Frequent Pattern Mining" PP-4-11 2011 IEEE.

[14] Thair Nu Phyu "Survey of Classification Techniques in Data Mining" Vol I Imecs 2009, March 18 - 20, 2009, Hong Kong.

[15] Zhen- Hui Song &Yi Li, "Associative classification over Data Streams", IEEE, PP.2-10, 2010.

[16] S.P.Syed Ibrahim1 K. R. Chandran2 M. S. Abinaya3 "Compact Weighted Associative Classification" IEEE pp.8-11, 2011.