A REVIEW OVER CLASSIFICATION APPROACHED IN DATA MINING FOR MULTI-RELATIONS

Jitendra Patel, Anurag Jain
Department of Computer Science & Engineering
Radharman Institute of technology & Science, Bhopal, India

ABSTRACT

Data mining is rapid growth field in order to mine the information from the huge data set. Data mining techniques are results of long process of research and product development and include artificial neural networks, decision trees and genetic algorithms. data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. When the data set gets very large then there is need of minimization the size of data set. It will help to reduce the complexity in dataset and minimize the search space the classification is one of the best approach to categorized the data. There are many classification methodology has been available like clustering, bysein approach, design tree etc. Recent research having lower predictive accuracy which lead to tends, combine existing log-linear model with probabilistic techniques. This paper is a survey on all these approaches with their basic working principal.

KEYWORDS: *Data Mining, Classification, Decision tree*

I. INTRODUCTION

Classification 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.

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.

II. STEPS OF DATA MINING

The KDD process is interactive and iterative, involving numerous steps with many decisions made by the user.

- First is developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer’s viewpoint.
- Second is creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
- Third is data cleaning and preprocessing. Basic operations include removing noise if Appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes.
- Fourth is data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or invariant representations for the data can be found.
- Fifth is matching the goals of the KDD process (step 1) to a particular data-mining method. For example, summarization, classification, regression, clusters, and so on.

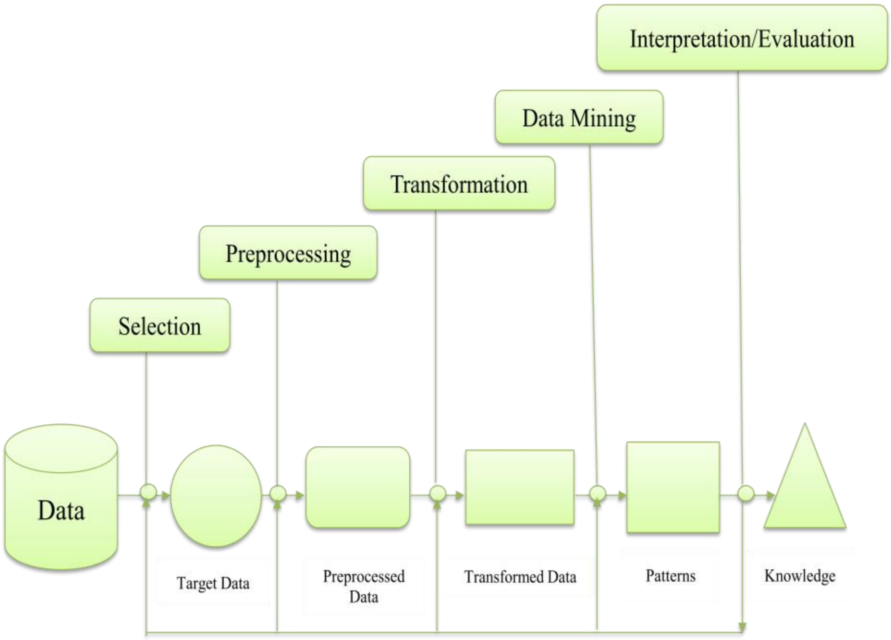

Figure 1 Data mining steps

- Sixth is exploratory analysis and model and hypothesis selection: choosing the data mining Algorithm and selecting method(s) to be used for searching for data patterns. This process includes deciding which models and parameters might be appropriate (for example, models of categorical data are different than models of vectors over the real) and matching a particular data-mining method with the overall criteria of the KDD process (For example, the end user might be more interested in understanding the model than its Predictive capabilities).

- Seventh is data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps.
- Eighth is interpreting mined patterns, possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models.
- Ninth is acting on the discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

III. CLASSIFICATION TECHNIQUES

It could be some time but not necessarily advisable predictive modeling is seen as a "black box" that makes predictions about the future based on information from the past and present. Some designs are better than others in terms of accuracy. Some designs are better than others in terms of understanding. For example, models from better understanding of the incomprehensible decision trees, rule induction, and regression models, neural networks. The classification is a type of predictive models. More specifically, the ranking is the appointment process of new objects or predefined categories: given a set of marked files, build a model such as the decision tree, and predicting future records labels is called for classes. This section discusses the classification techniques: used in Data Mining

Decision Tree

Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. An example of a decision tree for the training set. Decision tree is a predictive model which, as its name suggests, can be seen as a tree. Specifically each branch of the tree is a classification of matter and leaves of trees and data sections with classified. For example, if we classify customers who violently (not to renew their contracts) in the industry decision tree may seem cell phone

Using the decision tree as an example, the instance At1 = a1, At2 = b2, At3 = a3, At4 = b4∗ would sort to the nodes: At1, At2, and finally At3, which would classify the instance as being positive (represented by the values “Yes”). The problem of constructing optimal binary decision trees is an NP complete problem and thus theoreticians have searched for efficient heuristics for constructing near-optimal decision trees. The feature that best divides the training data would be the root node of the tree. There are numerous methods for finding the feature that best divides the training data such as information gain. While myopic measures estimate each attribute independently,

IV. BAYESIAN NETWORKS

Data extraction, through the simplest definition, automate the detection of relevant models in a database. For example, a model that may indicate that men are married with children are twice as likely to order a sports car of some married men without children. If you are the marketing manager for the automotive industry, and this model surprising to some extent, it can be very valuable. However, data mining is not magic. For many years, statisticians and databases manually "mined", the search for statistically significant patterns. Technical data and statistical machine learning extraction used to build a strong anticipation that customer behavior patterns. Today, mining and technology automates the process, and integrates with enterprise data stores, and provide relevant information to business users in a way. The main mining products are now more than a car models that use powerful algorithms. Instead, they meet the technical and commercial issues wider, such as integration in complex IT environments today.

A Bayesian Network (BN) [4] is a graphical model for probability relationships among a set of variables features. The Bayesian network structure S is a directed acyclic graph DAG) and the nodes in S are in one-to-one correspondence with the features X. The arcs represent casual influences among the features while the lack of possible arcs in S encodes conditional independencies. Moreover, a feature (node) is conditionally independent from its non-descendants given its parents (X1 is conditionally independent from X2 given X3 if P (X1|X2, X3) =P (X1|X3) for all possible values of (X1, X2, X3) [5].

Consequently, it is naive to believe Baez, representing the strength of the naive Bayes model is lower than that of decision trees. If it uses the nominal data model, can realize the limits of the single linear class. When using the digital data, the more complex (non-linear) can be represented boundary. Otherwise, the naive Bayes model has many advantages: It is very simple, effective and powerful sound and easy to interpret. It is especially for small data sets, because it combines a small complex with a probabilistic model appropriate flexibility. You must be Discretized suits just basic model to separate data and digital data. Instead, we can learn from the continuous model by estimating the density instead of distributions. However, assume Bayesian networks form a continuous year of distribution, normal distribution, which is often not realistic. Usually individualize the best solution because it also simplifies the model and seed output is more powerful for overfitting.

V. K-NEAREST NEIGHBOUR CLASSIFIERS

Nearest neighbor classifiers [4] are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points,

X=(x1, x2, …………,x_n) and

Y= (y1,y2,…………...y_n) is

$$(x+a)^n = \sqrt{}\sum_{i=1}^{n}(x_i - y_i)^2$$

The unknown sample is assigned the most common class among its k nearest neighbors. When k=1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space.

K nearest neighbors works represent a completely different approach to classification. They do not build an express universal model, but only locally rounded and implicitly. The main idea is to classify a new object by examining the value of class data points K. category selected the most similar class may be the most common among neighbors or distribution of the class in the district.

VI. CLUSTERING

Clustring and classification, both the basic functions in data mining. Classification is used mainly as a way of learning under the supervision block to learn uncensored. The objective of the meeting is descriptive, this classification is predictive. Because the goal of the meeting is to find a new set of new categories and groups are important in themselves, and their evaluation is essential.. As we mentioned before, classification can be taken as supervised learning process, clustering is another mining technique similar to classification. However clustering is a unsupervised learning process. Clustering[4] is the process of grouping a set of physical or abstract objects into classes of similar objects, so that objects within the same cluster must be similar to some extent, also they should be dissimilar to those objects in other clusters. In classification which record belongs which class is predefined, while in clustering there is no predefined classes. In clustering, objects are grouped together based on their similarities. Similarity between objects is defined by similarity functions; usually similarities are quantitatively specified as distance or other measures by corresponding domain experts. Most clustering applications are used in market segmentation. By clustering their customers into different groups, business organizations can provide different personalized services to different group of markets. For example, based on the expense, deposit and draw patterns of the customers, a bank can clustering the market into different groups of people. For different groups of market, the bank can provide different kinds of loans for houses or cars with different budget plans. In this case the bank can provide a better service, and also make sure that all the loans can be re-claimed. After a concept of "similarity" between the samples and the purpose of the whole can be formulated as an optimization problem to obtain the maximum similarity within the block and reduce the similarity between the group. If the feature vectors corresponding to all of the digital samples, the concept of similarity can be defined in terms of distances. But if the characteristics of factions, it is much more difficult to define a general idea of reasonable similarity.

- Density-Based Clustering
- Grid-Based Clustering :
- Model-Based Clustering
- Categorical Data Clustering

These are the basic approach by which the clustering will be apply on the raw dataset

VII. RELATED WORK AND PROBLEM IDENTIFICATION

There are studies various research paper and journal and know about data classification. All methodology and process are not described here. But some related work in the field of association classification discuss by the name of authors and their respective title.

S N	Author Name	Title	Publication Year	Methodology	Demerits
1	Dewan Md. Farid, Li Zhang, Chowdhury Mofizur Rahman, M.A. Hossain, Rebecca Strachan	Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks	2013 Elsevier	The author has introduced 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.	Here the proposed work does not work on dynamic feature set. So 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.
2	Geetha Manjunath, M. Narasimha Murty Dinkar Sitaramb	Combining heterogeneous classifiers for relational databases	2013 IEEE	The author has presented a practical, two-phase hierarchical meta classification algorithm for relational databases with a semantic divide and conquer approach	Classification accuracy can be Improved by eliminating some non-contributing tables. For this, the author has plan to associate an entropy metric with individual database tables and select the right subgraph from the Join Graph that minimizes the information loss.
3	Tahar Mehenni, Abdelouahab Moussaouib	Data mining from multiple heterogeneous relational databases using decision tree classification	2012 Elsevier	The author has proposed a classification approach across multiple heterogeneous relational databases. More specifically, given a set of inter-related databases	It can be 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.
4	Marko Debeljak, Aneta Trajanov, Daniela Stojanova, Florence Leprince, Sa so D zeroski,	Using relational decision trees to model out-crossing rates in a multi-field setting	2012 Elsevier	The approach proposed by authors a new methodology to predict the level of adventitious presence on a multi-field setting, where the influence of more than	The approaches can be apply on various data sets in order to compare study

				one GM field is considered at the same time. The structure	

Recent research having lower predictive accuracy which lead to tends, combine existing 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 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 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.

There are some limitation and problem of classification algorithm. Furthermore, we have also evaluated the traditional NB classifier using all the 10 datasets and achieved an average accuracy rate of 77.3% using 10-fold cross validation, while the proposed NB classifier (Algorithm 2) obtained an average accuracy rate of 86.7%. 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.

Now we suggest to used genetic algorithm along with probabilistic approach for the optimization of classification rate of association classification. In this case the results are improved because the genetic algorithm is a heuristic function. The heuristic function gives an optimal result. Whereas D-S theory approaches apply over historical data and give better result. Now we suggest to adopted optimization of classification of association rule with the help of genetic algorithm and D-S theory approach.

Multiple relational classification algorithm modified by GA and D-S theory, improve the predictive accuracy rate of classification in comparison of Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks

VIII. CONCLUSION

Data mining techniques are results of long process of research and product development and include artificial neural networks, decision trees and genetic algorithms. data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. This technology provides a wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. There is large number of classification approaches has been used to minimize the complexity. It seems to be that there are some limitation and problem of classification algorithm. This paper gives the general review on the classification algorithms. It also seems to be that the limitations can be resolve so that this paper also contains the proposal to improve the classification efficiently.

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

AUTHOR

Jitendra Kumar Patel has received the Bachelor in information technology from the University RGPV Bhopal, in 2012. He is currently pursuing the Masters. degree with the Department of Computer Science & Engineering, from RITS Bhopal. His research interests include Data Mining and DBMS.