Classification Approaches on Relational Databases

[1]T.Hemalatha,[2]J.Samatha,[3]Ch. Vani Priyanka,[4]A.Lavanya, [5]Ch.Ranjith Kumar

[1]Asst Professor , Dept. of Information Science and Technology, K.L.University, Guntur, India.
[2,3,4,5]B-Tech Student, Dept. of IST, K. L. University, Guntur, India.

Abstract
Relational databases are the most popular repository for structured data, and are thus one of the richest sources of knowledge in the world. In a relational database, multiple relations are linked together via entity-relationship links. Classification is an important task in data mining and machine learning, which has been studied extensively because of its usefulness development of classification across multiple database relations, becomes important. Multi relational classification is the procedure of building a classifier based on information stored in multiple relations and making predictions with it. There are many popular approaches for finding patterns in data. This paper provides an insight into various classification methods including ILP (Inductive Logic Programming), Relational database, emerging patterns and associative approaches. Their characteristics and comparisons in detail have also been provided.

Keywords
Tuple id propagation , Crossmine, Classification

1.Introduction

Relational databases are the popular format for structured data, and also one of the richest sources of knowledge in the world. There are many real world applications involving decision making process based on information stored in relational databases such as credit card fraud detection and loan application. Approaches that can perform in-depth analysis on relational data are of crucial importance in such applications. Therefore, multirelational data mining has become a field with strategic importance. Multi-relational data mining (MRDM) aims to discover knowledge directly from relational data. There have been many approaches for classification, such as neural networks and support vector machines. However, they can only be applied to data in single flat relations. It is counterproductive to convert multi-relational data into single flat table because such conversion may lead to the generation of huge relation and lose of essential semantic information. The important MRDM task is Multi-Relational Classification (MRC) which aims to build a classification model that utilizes information in different relations. Databases are rich with hidden information that can be used for intelligent decision making. Classification is a form of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous valued functions. The classification across multiple database relations is divided into two steps with the same propositional classification

- To learn classification model from examples (Learning step or Training Phase) and
- To classify and test using the model. Based on the methods of knowledge representation, this paper focuses on the relational classification with four main categories such as

i). Inductive Logic Programming (ILP) based MRC
ii). Emerging Patterns based MRC
iii). MRDM(Multi Relational Data Mining)
iv). Associative MRC

An extensive survey of literature was made to identify various research issues in this filed.

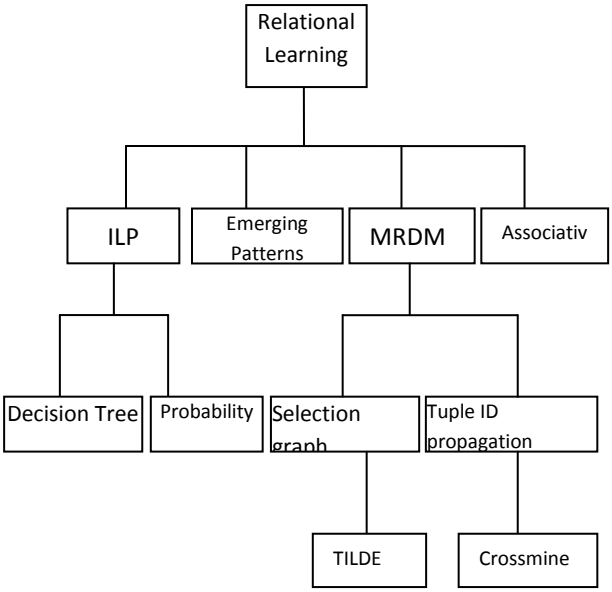


Fig. 1 Research Direction's Map

Manuscript received April 5, 2013
Manuscript revised April 20, 2013

2. INDUCTIVE LOGIC PROGRAMMING-BASED CLASSIFICATION

Inductive logic program in broadest sense the mathematical logic for computer programming It is characterized by the use of logic for the representation of multi relational data. It predicts the class labels based on background knowledge. They mainly include two categories – Decision tree relational classification, Probability classification approach (Probabilistic Relational Model and Stochastic Logic Program).

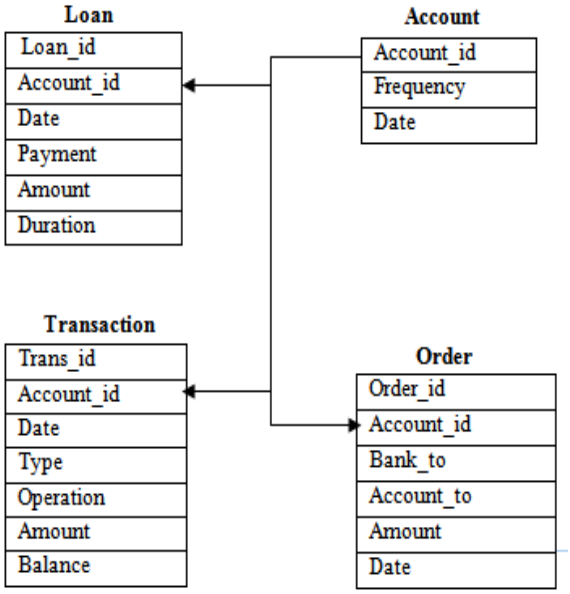


Fig. 2 The financial database

2.1. Decision tree relational classification approaches

A decision tree is a flow chart like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Each node in a 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 of node and sorted based on their feature values. The Decision tree construction does not require any domain knowledge and is appropriate for exploratory knowledge discovery.

Decision trees used in data mining are of two main types:

- Classification tree analysis is when the predicted outcome is the class to which the data belongs.
- Regression tree analysis is when the predicted outcome can be considered a real number

There are two major classification algorithms for inducing relational decision trees TILDE and SCART. These classification algorithms upgraded from the above trees. The major difference in comparison to the propositional method is its dependence on the tests along the path from root to the current node. The TDID algorithm of SCART first tests the termination condition. If it is yes, a leaf is constructed with an appropriate prediction. Otherwise a test is selected among the possible tests for the node at hand. It split the examples into subsets according to the outcome of the test. The tree construction proceeds recursively on each of the subsets.

2.2. Probability relational classification approaches

For dealing with the noise and uncertainty encountered in most real-world domains, probability is introduced into LBRC to integrate the advantages of both logical and probabilistic approaches to knowledge representation and reasoning. At present, the method mainly includes inductive Logic Programming and Bayesian Networks, ILP and Stochastic Grammars. Probabilistic Relational Model (PRM) is an extension of Bayesian networks for handling relational data. A PRM describes a template for a probability distribution over a database. The template includes a relational component, that describes the relational schema for the domain, and a probabilistic component, that describes the probabilistic dependencies that hold in the domain. A PRM, together with a particular universe of objects, define a probability distribution over the attributes of the objects and the relations that hold between them. Stochastic Logic Programs (SLPs) have been a generalization of Hidden Markov Models, stochastic context-free grammars, and directed Bayes nets. A stochastic logic program consists of a set of labeled clauses p: C, where p is a probability label described the probability information of the corresponding relational pattern and C is a logic clause for extended dependent relationship between data. And by learning the data, the clause set covers each specific example and probabilities record the dependence relationships.

3. EMERGING PATTERN BASED CLASSIFICATION

Emerging Patterns (Eps) are classes of regularities whose support significantly changes from one class to another and the main idea is to exploit class characterization provided by discovered emerging patterns for class labeling. In a border based approach is adopted to discover the EPs discriminating between separate classes. Classification by Aggregating Jumping

Emerging Patterns is proposed in (JEP-Classifier) , Classification by aggregating emerging patterns (CAEP) in , are eager-learning based approaches. JEP-Classifier uses Jumping EPs (JEPs) whose support increases from zero in one dataset to non-zero in the other dataset whereas CAEP uses general EPs. For datasets with more than two classes CAEP uses the classes in a symmetric way, whereas JEP-Classifier uses them in an ordered way. The paper proposed a CP-tree data structure to register method which improves the efficiency of EP discovery by adopting the counts of both positive and negative class. CAEP and JEP classifier are the two relatives of decision through emerging patterns (DeEPs). Instance based classifiers using Eps called DeEPs,
To achieve much better accuracy and efficiency than the previously proposed Ep based classifiers. High accuracy is achieved because the instance based approach enables DeEps to pin point all Eps to relevant to a test instance, sum of which are missed by the eager learning approaches. High efficiency is obtained using a series of data reduction and concise data representation techniques.

4. RELATIONAL DATABASE CLASSIFICATION

Relational Database Classification (RDC) includes
i) selection graph based relational classification
ii) tuple ID propagation based relational classification.

4.1. Selection graph based relational classification

The most characteristic of MRDM is its most intense combination with relational database. Selection graph model can use database language of SQL to directly deal with relational tables of database. Selection graph based MRC, from a multi-relational data mining frame, get out of ILP approaches and transform the relationship between the tables into intuitive selection graph that is easy to be represented by SQL. That is to say, the query by SQL can complete MRC. MRDTL (Multi-relational decision tree learning) in the frame is based on selection graph and has lots of similarity with classic decision tree algorithm by a series of refinement to add decision tree node until meeting an end and the leaf nodes getting class label.

4.2. TupleID propagation based relational classification

Tuple ID propagation is a technique for performing virtual join among the tables, which greatly improves efficiency of multi relational classification. Multi-relational decision tree learning algorithm (MRDTL) constructs a decision tree whose nodes are selection graphs is an extension of logical decision tree induction algorithm Top down Induction of Logical Decision Trees. It adds decision nodes to the tree through a process of successive refinement until some termination criterion is met. By using suitable impurity measure e.g. information gain, the choice of decision node to be added at each step is determined. MRDTL which improved the calculation efficiency and information loss of MRDTL.

Loan

loan-id	account-id	...	class
1	124		+
2	124		+
3	108		-
4	45		-
5	45		+

Account

Account-id	frequency	date	IDs	*class labels*
124	monthly	270291	1,2	2+,0-
108	weekly	230995	3	0+,1-
45	monthly	120994	4,5	1+,1-
67	weekly	010195	-	0+,0-

Fig. 3 An example of Tuple ID propagation (some attributes of Loan re not shown).

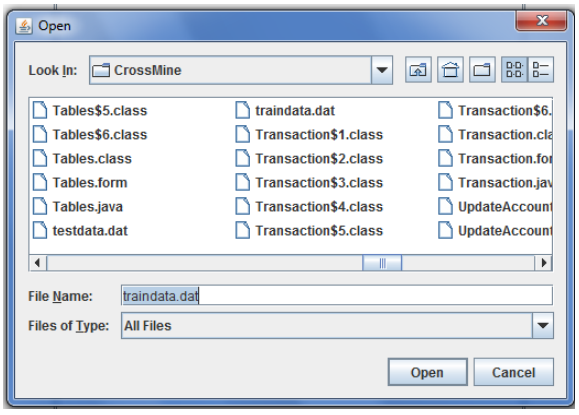
Tuple ID propagation is flexible and efficient because IDs can be easily propagated between any two relations, requiring only small amount of data transfer and extra storage space. Multirelational naïve bayes classifier Mr-SBC is an integrated approach of first-order classification rules with naive Bayesian classification, in order to separate the computation of probabilities of shared literals from the computation of probabilities for the remaining literals. However, while searching first–order rules, only tables in a foreign key path can be considered and other join paths are neglected. It handles categorical as well as numerical data through a discretization method. In general, Cross Mine is a divide-and conquer algorithm, which searches for the best way to split the target relation into partitions, and then recursively works on each partition. Cross Mine uses tuple ID propagation to efficiently search for good predicates or tree nodes among different relations, in order to build accurate classifiers based on relational information. The paper proposed two methods for

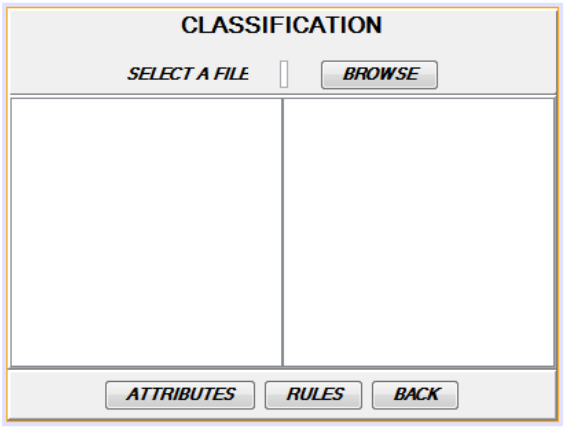
classification:Crossmine rule and Crossmine tree. Cross Mine-Tree and Cross Mine-Rule use different methods to split the target relation. Cross Mine-Tree searches for the best tree node in each step, and splits the target tuples according to the attribute of the tree node. In this way, it recursively builds a decision tree for classification. In contrast, Cross Mine-Rule generates a good rule in each step, and splits the target tuples according to whether they satisfy the rule. Then, it focuses on the tuples not satisfying the rule to generate more rules. Cross Mine-Tree is usually more efficient because it splits the target tuples into small partitions, while Cross Mine-Rule is usually more accurate because each rule is built based on a large set of tuples. We will first introduce the overall procedures of Cross Mine- Tree and Cross Mine-Rule, then describe the detailed methods for searching for predicates or attributes by tuple ID propagation. Classification with aggregation of Multiple Features (CLAMF) method is proposed in which is an adaptation of the sequential covering algorithm and classifies the multi relational data using aggregation involving single and multiple features. In temporal databases, classification with multi feature aggregation could provide very interesting rules that are much more meaningful to the end–user by allowing temporal trends. For eliminating the statistical skew in Graph-NB, the paper proposed an extended SRG and a new counting method to construct new multi-relational naïve Bayesian classifier. This paper here makes two chief contributions to the multirelational data mining community. First, a novel approach, namely, the MRC strategies devised for multirelational mining. Employing knowledge discovery methods which can be chosen from a wide range of existing propositional mining algorithms, learning directly from relational databases, and excluding the extensively "flattening" preprocessing make the MRC algorithm appropriate for mining useful patterns from many real-world databases. Another contribution is that, the paper here suggests the benefits of incorporating sets of features (views) when dealing with multirelational data. In other words, the adoption of multi-view learning framework may shed light on the issue of making good use of the rich feature space presented in structured domains, where attributes to describe an object are usually large and often highlight from different aspects.

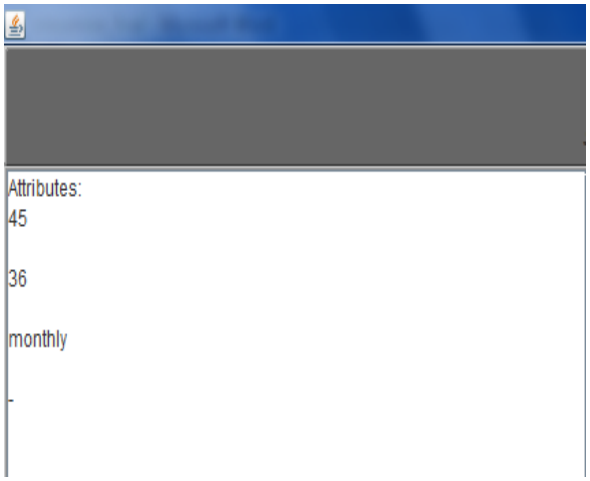
5. ASSOCIATIVE CLASSIFICATION

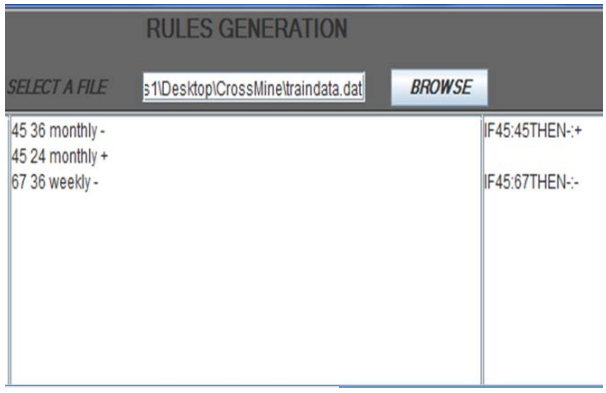
Associative classification is a promising technique to build accurate classifiers. However in a large correlated data sets, association rule mining may yield huge rule set. Several algorithms have been proposed for associative classification such as Classification based on Multiple Association Rule (CMAR), Classification based on Predictive Association Rules (CPAR).CMAR determines the class label by a set of rules. To improve both accuracy and efficiency, it employs a data structure called Classification Rule- tree, to compactly store and retrieve a large number of rules for classification. To speed up the mining of complete set of rules, it adopts a variant of Frequent-Pattern growth method. CPAR combines the advantages of both associative classification and traditional rule-based classification. It adopts a greedy algorithm to generate rules directly from training data. All the above algorithms only focus on processing data in a single table and applying these algorithms in multi relational environment will result in many problems. The paper extends Apriori to mine the association rules in multiple relations. The paper is also based on deductive databases. These two approaches cannot be applied in relational databases directly. They have high computational complexity, and the pattern they find is hard to understand. Apriori algorithm attempts to find subsets which are common to at least a minimum number of data sets. Apriori uses a bottom up approach where frequent subsets are extended one subset at time. It reflects the association between class labels and other item sets, and used to generate classification rules. The rules discovered have more comprehensive characterization of databases.

7. Results









Generating rules and finding the accuracy

1nd set Rule

IFacc_id:124THENclass_lables:+
IFacc_id:108THENclass_lables:-
IFacc_id:45ANDduration:12ANDduration:24THENclass_lables:+
IFacc_id:45ANDduration:36THENclass_lables:-
IFacc_id:67THENclass_lables:-
Number of Rules Generated in the Iteration:5

2nd set Rule

IFacc_id:124THENclass_lables:+
IFacc_id:108THENclass_lables:-
IFacc_id:45ANDduration:12ANDduration:24THENclass_lables:+
IFacc_id:45ANDduration:36THENclass_lables:-
IFacc_id:67THENclass_lables:-
IFacc_id:67ANDfrequency:monthlyANDacc_id:124THENclass_lables:+
IFacc_id:45ANDduration:12ANDduration:36THENclass_lables:-
IFacc_id:45ANDduration:24THENclass_lables:+
IFacc_id:67ANDduration:12ANDduration:36THENclass_lables:+
IFacc_id:67ANDduration:24THENclass_lables:-
IFacc_id:67ANDfrequency:weeklyTHENclass_lables:-
Number of Rules Generated in the Iteration:11

3nd set Rule

IFacc_id:124THENclass_lables:+
IFacc_id:108THENclass_lables:-
IFacc_id:45ANDduration:12ANDduration:24THENclass_lables:+
IFacc_id:45ANDduration:36THENclass_lables:-
IFacc_id:67THENclass_lables:-
IFacc_id:67ANDfrequency:monthlyANDacc_id:124THENclass_lables:+
IFacc_id:45ANDduration:12ANDduration:36THENclass_lables:-
IFacc_id:45ANDduration:24THENclass_lables:+
IFacc_id:67ANDduration:12ANDduration:36THENclass_lables:+
IFacc_id:67ANDduration:24THENclass_lables:-
IFacc_id:67ANDfrequency:weeklyTHENclass_lables:-
IFacc_id:124THENclass_lables:+
IFacc_id:108ANDduration:12ANDduration:24THENclass_lables:-
Number of Rules Generated in the Iteration:13
Accuracy Rate:65Accuracy rate::

8. Conclusions

Multi-relational data mining deals with knowledge discovery from relational databases consisting of multiple tables. With the development of data mining techniques, multi relational data mining has become a new research area. This paper presents the several kind of classification methods across multiple database relations including ILP based, Relational database based, Emerging Pattern based and Associative based approaches. Out of these four approaches, the Relation Based Classification seems to have drawn a lot of attention and exploration because of its convenience, usefulness and suitability. Still, the other approaches can be explored to a fare amount of detail.

References

[1] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery.
[2] Rule Induction with CN2: Some Recent Improvements.
[3] H. Garcia-Molina, J.D. Ullman, and J. Widom, Database Systems: The Complete Book. Prentice Hall.
[4] J. Gehrke, R. Ramakrishnan, and V. Ganti, "Rainforest: A Framework for Fast Decision Tree Construction of Large Data Sets," Proc. 1998 Int'l Conf. Very Large Data Bases.
[5] N. Lavrac and S. Dzeroski, Inductive Logic Programming: Techniques and Applications.
[6] H. Liu, H. Lu, and J. Yao, "Identifying Relevant Databases for Multidatabase Mining," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining.
[7] T.M. Mitchell, Machine Learning.
[8] S.Muggleton, Inductive Logic Programming.
[9] S. Muggleton, "Inverse Entailment and Progol," New Generation Computing, special issue on inductive logic programming.
[10] S. Muggleton and C. Feng, Efficient Induction of Logic Programs.
[11] J. Neville, D. Jensen, L. Friedland, and M. Hay, "Learning Relational Probability Trees," Proc. 2003 Int'l Conf. Knowledge Discovery and Data Mining, 2003
[12] A. Popescul, L. Ungar, S. Lawrence, and M. Pennock, "Towards Structural Logistic Regression: Combining Relational and Statistical Learning," Proc. Multi-Relational Data Mining Workshop, 2002.
[13] J.R. Quinlan, C4.5: Programs for Machine Learning.
[14] J.R. Quinlan and R.M. Cameron-Jones, "FOIL: A Midterm Report.
[15] B. Taskar, E. Segal, and D. Koller, "Probabilistic Classification and Clustering in Relational Data," Proc. 2001 Int'l Joint Conf. Artificial Intelligence, 2001.
[16] X. Wu and S. Zhang, "Synthesizing High-Frequency Rules from Different Data Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 2, pp. 353-367, 2003.

T.HEMALATHA received her B-Tech degree in Computer Science and En from Jawaharlal Nehru Technology University, Hyderabad, in 2006, the M-Tech degree in Computer Science from SRM University, Chennai, in 2009. She is working as an Assistant Professor, with Department of Information Science Technology, Koneru Lakshmaiah University, Vijayawada, from 2009 to till now. Her research interests include Cluster technologies, Classification of Data mining techniques.

J.Samatha is pursuing B.Tech in Koneru lakshmaiah college of engineering. She got selected in HCL Company. Her area of interest is Data mining.

A.Lavanya is pursuing B.Tech in Koneru lakshmaiah college of engineering. She got selected in TCS Company. Her area of interest is Data mining.

CH.Vani priyanka is pursuing B.Tech in Koneru lakshmaiah college of engineering. She got selected in HCL Company. Her area of interest is Data mining.

CH.Ranjith kumar is pursuing B.Tech in Koneru lakshmaiah college of engineering. His area of interest is Data mining.