Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 12, December 2013, pg.406 – 411

RESEARCH ARTICLE

Discovering Classification Rules from Multi Relational Database Having Multiple Class Values

Purvi R. Patel[1], Mahesh Panchal[2], Chetna Chand[3]

[123]Computer Department & Gujarat Technological University, Gujarat, India

[1] purvipatel.211@gmail.com; [2] mkhpanchal@gmail.com; [3] chetnachand88@gmail.com

Abstract— Today in the real world data are stored in a structured format known as relational database. An increasing number of data mining applications involve the analysis of complex and structured types of data and require the use of expressive pattern languages. So, to classify object in one relation, we have to collect information from another relation. Multi-relational classification is the process of building a classifier based on information stored in multiple relations. Multi relational classification is used to predict behavior and unknown patterns automatically. This paper presents the various approaches of classification for relational database such as Tuple ID Propagation, Selection Graph, Multi-View Learning etc. and proposed work for multi class classification using Tuple ID Propagation in multi relational database.

Keywords— Multi relational classification; Tuple ID propagation; Selection Graph; Multi View Learning

I. INTRODUCTION

The main objective of data mining technique is to extract the information from large amount of data. There are lots of existing data mining techniques available, which look for pattern in single table, but in real world, data are stored in structured format known as relational database. Such a database consist of multiple relation which are linked together conceptually via entity relationship links[1].

A. *Multi relational classification*

Multi relational classification build a model for predicting class labels of target objects using information in multiple relation and goal is represented by the class label in the training data sets. Multi relational classification aims at building a classification model that utilize information in different relations. In a database for multi relational classification, there is one target relation, Rt, whose tuples are called target tuples and are associated with class labels. The other relations are called non-target relations. Each relation has one primary key which uniquely identifies tuples in the relations and several foreign keys where a primary key in one relation can be linked to the foreign key in another.

Research direction map for relational database.

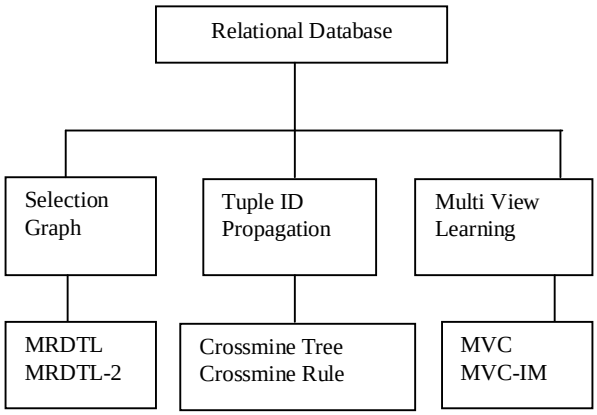

Fig. 1 : Research direction map for relational database.

II. LITERATURE REVIEW

In this section, I present the different approaches of multi relation classification based on relational database such as Selection Graph, Tuple ID Propagation and Multi View Learning.

A. *Selection Graph*

When multi relational patterns are expressed in terms of graphical language, then we called as selection graph. Selection graph model is used in SQL database language to directly deal with tables.

i) MRDTL (multi relational decision tree algorithm) [7]

MRDTL constructs a decision tree for classifying a target attribute from a target table in a given database. MRDTL adds decision nodes to the tree through a process of successive refinement until some termination criterion is met. The choice of the decision node to be added at each step is guided by a suitable impurity measure such as information gain. MRDTL starts with the selection graph containing a single node at the root of the tree, which represents the set of all objects of interest in the relational database. This node corresponds to the target table T_0. The algorithm iteratively considers every possible refinement that can be made to the current selection graph S with respect to the database D and selects the optimal refinement and its complement. Each candidate refinement is evaluated in terms of the split of the data induced by it with respect to the target attribute, as in the case of the propositional version of the decision tree learning algorithm The hypothesis resulting from the induction of the relational decision tree algorithm can be viewed as a set of SQL queries associated with the selection graphs that correspond to the leaves of the decision tree.

ii) MRDTL-2[9]

MRDTL has two significant limitations from the standpoint of multi-relational data mining from large, real-world data sets:

(a) Slow running time
(b) Inability to handle missing attribute values

MRDTL-2 extends an MRDTL which includes enhancements that overcome two significant limitations of MRDTL.

iii) Limitations:

- Incorporation of sophisticated methods for handling missing attributes values into MRDTL-2.
- Incorporation of sophisticated pruning methods or complexity regularization techniques into MRDTL-2 to minimize overfitting and improve generalization.
- More extensive experimental evaluation of MRDTL-2 on real-world data sets.

- Development of ontology-guided multi-relational decision tree learning algorithms to generate classifiers at multiple levels of abstraction (based on the recently developed prepositional decision tree counterparts of such algorithms.
- Development of variants of MRDTL for classification tasks where the classes are not disjoint, based on the recently developed propositional decision tree counterparts of such algorithms.
- Development of variants of MRDTL that can learn from heterogeneous, distributed, autonomous data sources based on recently developed techniques for distributed learning
- Application of multi-relational data mining algorithms to data-driven knowledge discovery problems in bioinformatics and computational biology.

B. *Tuple ID Propagation*

Tuple ID propagation is a technique for performing virtual join, which greatly improves efficiency of multirelational classification. Instead of physically joining relations, they are virtually joined by attaching the IDs of target tuples to tuples in nontarget relations. Tuple ID propagation is flexible and efficient, because IDs can easily be propagated between any two relations, requiring only small amounts of data transfer and extra storage space. By doing so, predicates in different relations can be evaluated with little redundant computation [1]. CrossMine - CrossMine, which includes a set of novel and powerful methods for multi relational classification, including

- Tuple ID propagation, an efficient and flexible method for virtually joining relations, which enables convenient search among different relations.
- New definitions for predicates and decision-tree nodes, which involve aggregated information to provide essential statistics for classification.
- A selective sampling method for improving scalability with regard to the number of tuples[1][5].

i)CrossMine Tree[20]

Initially works on all target tuples and set the target relation to active. At each step, it searches for the best attribute A in all active relations and relations joinable with any active relation, and uses A to divide the target tuples. CrossMine-Tree works on each partition of tuples, and recursively builds a tree for that partition. A tree node is not further divided if the number of target tuples is less than Minimum number of tuples.

ii)CrossMine Rule

Repeatedly search for a best predicate and append it to the current rule, until no gainful predicate can be found. A relation is active if it appears in the current rule. Every active relation is required to have the correct propagated IDs on every tuple before searching for the next best predicate[5][14][20].

iii) Limitations:

- Improving scalability by directly use database operation to achieve Tuple ID propagation.
- Sometimes too many IDs are propagated to each tuple in a relation, which makes it hard to limit the time/space complexity of the algorithm.

C. *Multi View Learning*

The multi-view learning problem with n views can be seen as n inter-dependent relations and are thus applicable to multi-relational learning.

i)MVC[12]

Multi View Classification approach first classifies the multiple databases into different local groups. Next, local patterns are learned from each of these groups. Finally, a weighting process is initiated to synthesize the obtained local patterns.

There are five stages in this method.

1. Information Propagation stage :

Use Tuple ID Propagation to Propagate the IDs from target relation to non-target relation.

2. Aggregation stage :
   Summarize information embedded in multiple tuples and squeeze them in to one row.

3. View Learner Construction stage :
   Construct various hypotheses on the target concept.

4. View Validation stage :
   Validate various hypotheses generated in above stage.

5. View Combination stage :
   The resulting multiple view learners are incorporated in to a meta learner to construct the final classification model.

ii)MVC-IM[11][13]

Multi View Classification for imbalanced data. The main goal of the MVC-IM method is to learn a model which can better predict both the majority and minority class in imbalanced data. MVC-IM method extends the MVC method by introducing two methods:

1. View Validation :
   To produce a combine classifier, which is superior to the individual view learner, each inducted view learner has sufficient knowledge on the minority class.

2. View Combination :
   The voting combination technique is employed to obtain better knowledge on both majority and minority classes. For this, different techniques are used such as Stacking, Bagging and Boosting.

iii) Limitations:

- To study applying data preprocessing techniques such as feature selection in order to further improve the performance of the MVC algorithms.
- It would also be interesting to examine the influence of different model combination techniques and view validation strategies.
- Study different "goodness" heuristic measurements and their impact on these algorithms.
- Evaluating the method against learning tasks with more than two classes will be interesting to investigate.
- Study how the total tuples and imbalanced ratio in each resulting view impacts the result of the final combination model.
- Also, it would be very interesting to further investigate relational schemas with composite keys.
- MVC can be extended to include selection of the right classifier at the table level.
- Consequently, no guidelines are available to select the best classifier for a particular type of data.
- In future, experimentation with different view combination techniques, such as majority voting and weighted voting can be future investigated.

D. *Motivation from Literature*

Today, in the real world all data are stored in the structured format. This data are used for better decision making process. It is not necessary that the all-time the decisions are taken in the binary form such as Yes or No. Sometimes many situations are occurring in which more than two types of decisions come which is a challenging task. The methods found in literature can handle binary class data well. It motivates to propose the method Cross Mine Tree – MC for the multiclass data in multi relational database.

III. PROPOSED APPROACH

The proposed work extracts the classified rules using Tuple ID Propagation from Multi class Multi relational database. The proposed algorithm is CrossMine Tree-MC(CrossMine Tree – Multi class) for multi class classification in multirelational database.

The Steps for proposed Algorithm are as follows :

1. First , Data preprocessing is done on target relation. This process is shown in figure 2:

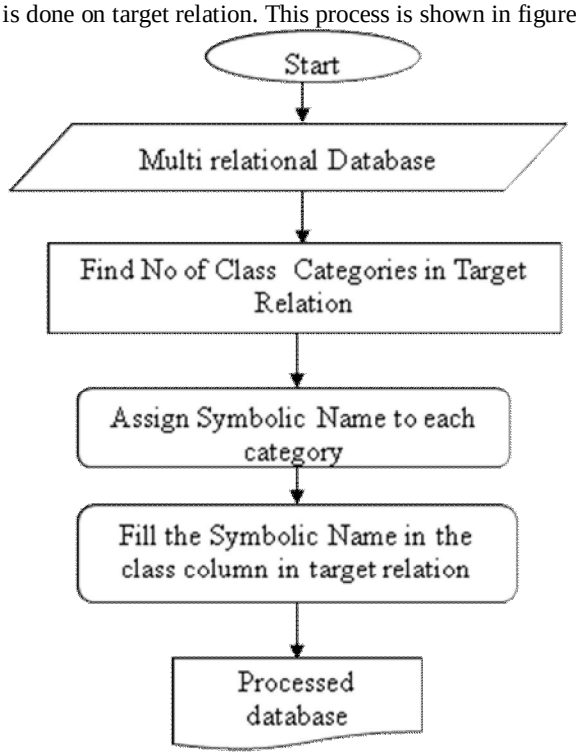

Figure 2 : Data Preprocessing Flow Chart

2. Take the multi relational database as input.
3. Tuple Id propagation is performed.
4. Count the number of target tuples in the target relation.
5. If no. of target tuples < MIN_SUP then Return. (MIN_SUP=Minimum No. of tuples required for classification)
6. Evaluate all attributes(A) of the target relation and calculate the Information Gain of each attribute. Information gain of class label is calculated by

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

Information gain of attribute is calculated by

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$

Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on *A*). That is,

$$Gain(A) = Info(D) - Info_A(D).$$

7. If Information Gain(A)<MIN_INFO_GAIN then return.
8. The attribute with highest INFO_GAIN is become a root node of the tree.
9. Then divide the target relation according to root node.
10 . Repeat until all attributes are evaluated and finally set the class label with higher frequency.

IV. CONCLUSIONS

Multirelational data mining deals with knowledge extraction from relational database containing multiple tables. This paper presents the different classification approaches based on relational database

including selection graph, Tuple ID Propagation and Multi View Learning. All of these methods are implemented for two classes. But in real world, many situations are occurred in which multiple decisions are take place. The proposed work is used for this type of situation.

V. REFERENCES

[1] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd ed., Morgan Kaufmann Publishers, 2006.

[2] Saˇso Dˇzeroski Jozˇef Stefan "MultiRelational Data Mining: An Introduction" InstituteJamova 39, SI1000,Ljubljana, Slovenia

[3] Dzeroski, S., Lavtac, N. (2001). "Relatioanl data mining", Berlin: Springer.

[4] S.Muggleton. Inverse Entailment and Progol. "New Generation Computing, Special issue on Inductive Logic Programming", 1995.

[5] Xiaoxin Yin, Jiawei Han and Jiong Yang, "Efficient Multi-relational Classification by Tuple ID Propagation" Department of Computer Science, University of Illinois at Urbana-Champaign

[6] Neelamadhab Padhy , Rasmita Panigrahi , "Multi Relational Data Mining Approaches: A Data Mining Technique" ,International Journal of Computer Applications (0975 – 8887) Volume 57– No.17, November 2012

[7] Anna Atramentov, Vasant Honavar, "Speeding Up Multi-Relational Data Mining"

[8] Amit Thakkar, Y P Kosta "Survey of Multi Relational Classification (MRC) Approaches & Current Research Challenges in the field of MRC based on Multi-View Learning" , International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-6, January 2012

[9] Anna Atramentov, Hector Leiva, and Vasant Honavar "A Multi-relational Decision Tree Learning Algorithm - Implementation and Experiments"

[10] Dr. M. Thangaraj, C.R.Vijayalakshmi, "A Study on Classification Approaches across Multiple Database Relations", International Journal of Computer Applications (0975 – 8887) Volume 12– No.12, January 2011

[11] Hongyu Guo and Herna Viktor "Mining Imbalanced classes in multi relational classification"

[12] Hongyu Guo and Herna Viktor "Multi relational classification using multi view learning"

[13] Hongyu Guo, and Herna Viktor, "Learning from Skewed Class Multi Relational Database", School of information technology and engineering, University of Ottava, Canada

[14] Xiaoxin Yin, Jiawei Han and Jiong Yang, Philip S. Yu, "CrossMine : Efficient Classification Across Multiple Database Relation"

[15] Ayesha Shaikh, Ankita Kapadia, "Competent Multi Relational Classifier using Filter based Feature Selection Method on CrossMine Algorithm"

[16] Shraddha Modi, Amit Thakkar, Amit Ganatra, *A* "Survey on Approaches of Multirelational Classification Based On Relational database"

[17] Zhen Peng, Lifeng Wu, Xiaoju Wang, "Research on Multi-Relational Classification Approaches"

[18] Amal S. Ghanem, Svetha Venkatesh , Geoff West "Learning in Imbalanced Relational Data"

[19] Prof. Saurabh Tandel, Prof. Vimal Vaghela, Dr. Nilesh Modi, Dr. Kalpesh Vandra, "Multi Relational Data Mining Classification Procession- A Survey"

[20] M. Thangaraj, Ph.D, C.R. Vijyalakshmi, "Performance Study on Rule based Classification Techniques across multiple Database Relations", Madurai Kamraj University, Tamil Nadu, India. March - 2013

