International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249–8958, Volume-1 Issue-3, February 2012

A Survey on Approaches of Multirelational Classification Based On Relational Database

Shraddha Modi, Amit Thakkar, Amit Ganatra

Abstract*: Classification is an important task in data mining and machine learning, in which a model is generated based on training dataset and that model is used to predict class label of unknown dataset. Today most real-world data are stored in relational databases. So to classify objects in one relation, other relations provide crucial information. Relational databases are the popular format for structured data which consist of tables connected via relations (primary key/ foreign key). So relational databases are simply too complex to analyse with a propositional algorithm of data mining. To classify data from relational database need of multi relational classification arise which is used to analyze relational database and used to predict behaviour and unknown pattern automatically which include credit card fraud detection, disease diagnosis system, financial decision making system, information extraction and face recognition applications. This paper presents survey of different approaches to classify data from multiple relations, which includes Flattening based approach, Upgrading approach and Multiple view based approach.

Index Terms*: Inductive logic programming, Multi relational classification, Multiple view, Multi-view, Relational database, Selection graph, Tuple id propagation.

I. INTRODUCTION

Multi relational classification is an important part of Multi relational data mining which involve applications like credit card fraud detection, disease diagnosis system, financial decision making system, information extraction and face recognition applications. Most existing data mining methods are propositional, and hence unfortunately do not work with current relational data, which contain multiple tables. Current databases need to store data that is simply too complex to be expressed as a single type of entity, or could contain relationships between entities. Propositional data mining in these cases is replaced by Multi-Relational Data Mining (MRDM). MRDM works with data stored in a multi-relational database and the relationships between entities. In this type of database, all the relationships between the entities are explicitly given and are expressed through the use of foreign keys and primary key.

There are many algorithms available for classification but they are applied only on single/flat file. So if we want to deal with relational database, either we have to upgrade propositional algorithms or we have to convert relational database into single file. It is not trivial to extend techniques that mine propositional data so that they work efficiently and accurately on multi-relational databases. Converting relational data into single flat file is not fruitful option because result of this process can be huge, contain much duplicate information and still loose essential information from data. Lots of work has been done in the flattening approach. The most widely studied method for relational classification is inductive logic programming (ILP). Mainly techniques for relational classification which upgrades propositional algorithms are ILP based relational classification, Relational database based relational classification, Association or frequent pattern based relational classification. Another approach called multiple view based relational classification is also used to predict behaviour of unknown pattern from relational database which neither upgrade propositional algorithm nor flatten the database. These techniques are described briefly in next section.

Manuscript published on 28 February 2012.
***** Correspondence Author (s)
Shraddha Modi*, U and P.U. Patel Department of Computer Engineering, Chandubhai S Patel Institute of Technology, Changa, India, (e-mail: shraddhamodi@yahoo.co.in).
Amit Thakkar*, Department of Information Technology, Chandubhai S Patel Institute of Technolog, Changa, India, (e-mail: amitthakkar.it@ecchanga.ac.in).
Amit Ganatra*, U and P.U. Patel Department of Computer Engineering, Chandubhai S Patel Institute of Technology, Changa, India, (e-mail: amitganatra.ce@charusat.ac.in).

II. MULTIRELATIONAL CLASSIFICATION

For relational classification, relational database is needed which consist of tables connected through primary key/foreign key relationship. Multi-relational classification can directly look for patterns that involve multiple relations from a relational database. So we can say relational database R is a collection of tables R = $\{R_1,R_2,...R_n\}$. A table R_i consists of a set of tuples T_R and has at least one key attribute, either the primary key attribute and/or the foreign key attribute. Foreign key attributes link to key attributes of other tables. This link specifies a join between two tables. Foreign key relationship may be directed or undirected between tables. For relational classification, we have one target relation R_t and other background relations $R_{b1},R_{b2},....R_{bn}$. Each tuple $x \in TR_t$ includes a unique primary key attribute x.k and a categorical variable (target variable) y. The aim of relational classification is to find a function F(x) which maps each tuple x of the target relation R_t to the category y such that:

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$y = F(x, R_t, R_{b1}, R_{b2}, \ldots R_{bn}), \; x \in TR_t$$

III. RELATED WORK

The relational classification across multiple database relations is divided into two steps with the same propositional classification to learn classification model from examples and to classify and test using the model. Relational classification makes use of additional information of related objects using multiple relations.

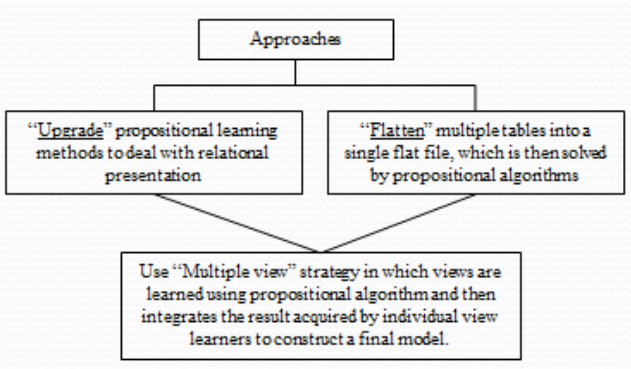

Figure 1. **Approaches to Multirelational classification**

There are mainly two categories of relational classification as shown in Fig. 1., either propositional learning algorithm should be upgraded to deal with relational format (upgrading) or relational format should be converted into flat format and then applying propositional learning algorithm (flattening). Another approach called multiple view approach is also used which combines upgrading and flattening approach because it uses propositional algorithm so no need to upgrade algorithm and combines the result of different views. All these approaches are described next.

A. Flattening Approach

In this approach multiple relations are converted into single flat file and then propositional algorithms are used for classification. Several approaches for classification in multi-relational setting have been proposed in the literature, but often the problem is solved by moulding a relational database into a single table format, such that traditional attribute-value algorithms are able to work on. This approach is also known as propositionalization. Two techniques have been proposed for propositionalization. The former is based on the principle that it is possible to consider a single relation reconstructed by performing a relational join operation on the tables. But it produces an extremely large and impractical to handle table with lots of data being repeated. A different technique is the construction of a single central relation that summarizes and/or aggregates information which can be found in other tables. Also this approach has some drawbacks, since information about how data were originally structured is lost.

LINUS algorithm uses flattening approach in which background knowledge is used to introduce new attributes into the learning process. Following the same lines, propositionalization approaches such as DINUS and SINUS were designed. Also RELAGGS propositionalization method is introduced which shows superior predictive performance over the DINUS and SINUS algorithms. In the RELAGGS strategy, new features are constructed for the target relation through an aggregation process, where function dependencies in the database are considered [19]. So in RELAGGS aggregate operators are used to transform multiple relations into a single table in order to be able to use propositional algorithms for the learning.

B. Upgrading Approach

1) ILP based relational classification: The ILP paradigm employs logic programs to describe patterns. The logic programs need to be induced from a database of logical facts, hence the name Inductive Logic Programming. ILP searches for hypothesis from examples based on background knowledge that will be used to classify unknown data. They achieve good classification accuracy in multi relational classification [1]. A major shortcoming of ILP is the computational demand that results from the large hypothesis spaces searched [2]. ILP based approach mainly include three categories: Decision tree relational classification, Instance based relational classification (RIBL and kernel) and Probability classification approach (PRM and SLP).

a) Relational classification using decision tree: The idea behind propositional decision tree algorithms is that, given an input with well-defined attributes, we can classify the input entirely based on making choices about each attribute. In tree each node has a question to be answered. Once we reached to the leaves of the tree, we get the answer. Each of the questions corresponds to a node in the decision tree, and each node has branches for each possible answer. Eventually, the algorithm reaches a leaf node that contains the correct classification of the input or the correct decision to make. Decision trees are quite easy because there is no need of domain knowledge to create tree. Multi-relational decision tree learning algorithm constructs a decision tree whose nodes are multi-relational patterns. CART and C 4.5 are famous propositional decision tree algorithms. To deal with relational data, these algorithms are upgraded and named TILDE [3] and SCART [4]. The expressivity of these trees is shown to be larger than that of the flat logic programs which are typically induced by classical ILP systems [3]. TILDE is more efficient than most traditional ILP approaches due to the divide-and-conquer nature of decision tree algorithm [13]. In [4], the system constructs a tree containing a literal (an atomic formula or its negation) or a conjunction of literals in each node, and assigns either a discrete class or a numerical value to each leaf.

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.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

b) Relational classification using instance: In [8], instance based methods offer solutions to the often unsatisfactory behaviour of current inductive logic programming (ILP) approaches in domains with continuous attribute value and in domains with noisy attributes and/or examples. RIBL is a generalization of the propositional k-nearest neighbour algorithm and a lazy learning method because it does not create an explicit model of the knowledge learned. In [9], Kernel methods have gained a great deal of popularity in the machine learning community as a method to learn indirectly in high-dimensional feature spaces.

c) Relational classification using probability: 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 models (PRMs) extend Bayesian networks with the concepts of objects, their properties, and relations between them [5]. A PRM [6] 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, defines a probability distribution over the attributes of the objects and the relations that hold between them. The relational component describes entities in the model, attributes of each entity, and references from one entity to another. The probabilistic component describes dependencies among attributes, both within the same entity and between attributes in related entities. An edge between attributes represents a probabilistic dependence of one attribute on the other attribute. An improved version of PRMs is represented by SRMs (Statistical Relational Models). Differently from PRMs, SRMs have a different semantics and are able to capture tuple frequencies in the database.

A stochastic logic program [7] consists of a set of labelled clauses p: C, where p is a probability label which describes the probability information of the corresponding relational pattern and C is a logic clause for extended dependent relationship between data. An SLP [6] defines a sampling distribution over logic programming proofs; as a consequence, it induces a probability distribution over the possible ground facts for a given predicate.

2) Relational database based relational classification: It includes Selection graph based relational classification and Tuple ID propagation based relational classification. Selection graphs can be represented graphically as labelled directed graphs. In [10], Selection graph model can use database language SQL to directly deal with relational tables of database. It transforms the relationship between the tables into intuitive selection graph that is easy to be represented by SQL. Multi relational decision tree learning algorithm (MRDTL) [11] is an upgraded propositional algorithm which adds selection graphs as the nodes to the tree through a process of successive refinement until some termination criterion is met. 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. The paper [12] proposed MRDTL-2 algorithm, which improves the calculation efficiency and information loss of MRDTL. MRTDL-2 also uses selection graph to represent rules which visually depict the SQL statements used to describe the rules.

Tuple ID propagation is used to perform virtual join between non-target relations and the target relations. Tuple Id propagation approach greatly improves efficiency with compare to physical join because of small amount of data transfer between tables. The paper [13] proposed CrossMine for relational classification. CrossMine uses Tuple ID propagation, in which it propagates data into related tables through foreign key relationships instead of performing a physical-join in the database. The idea of tuple ID propagation is to virtually join the relations with minimal cost, and then find good predicates in the joined relation. CrossMine obtains high scalability by avoiding the high cost of physical joins [13]. CrossMine uses a sequential covering algorithm to construct rules for classification. In many sequential covering algorithms, the negative examples are never removed in the rule building process, which makes the algorithm inefficient for databases with large numbers of tuples. To address this issue, CrossMine employs a selective sampling method to reduce the number of negative tuples when the numbers of positive and negative tuples are unbalanced. This helps CrossMine to achieve high scalability w.r.t. the number of tuples in databases. In [14], tuple ID propagation is used to speed up the join operation for Graph-NB algorithm which is upgraded from Naïve Bayesian classifier. It also uses semantic relationship graph to avoid unnecessary join among tables.

3) Association or frequent pattern based relational classification: Association rule mining discovers interesting relations between variables in large databases. These interesting relations are represented in rule form. And these rules are used in classification. In [15], gSpan is proposed for frequent pattern mining in graphs, which can be applied on multi-relational data. In [16] the authors have proposed an approach for association rule mining in relational databases. In [17] authors presented Warmer, which extends APRIORI to mine association rules in multiple relations. In [18], A Multirelational classification algorithm based on association rules (MrCAR) is proposed. It uses class frequent closed itemsets and reflects the association between class labels and other itemsets, and used to generate classification rules. MrCAR have higher accuracies comparing with the existing multi relational algorithm.

C. Multiple View Approach

In multiple view approach, problem can be characterized using different representations (views), and that learning from these representations separately can lead to better gains than merging them into a single dataset.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Using a relational database as input, Multi-view Relational Classification strategy learns from multiple views (feature set) of a relational database, and then information acquired by view learners are integrated to construct a final classification model. As in [19] the MVC algorithm works as follows: Firstly, the tuple IDs and the target concepts from the target relation (a relation in which each tuple is associated with a class label) are propagated to other relations (background relations), based on the foreign links between them. Secondly, aggregation functions are applied to each background relation in order to handle the one-to-many relationships. Each background relation is then used as training data for a particular multi-view learner. Thirdly, conventional single-table data mining methods are used in order to learn the target concept from each view of the data separately. Lastly, the multi-view learners trained are validated and incorporated into a meta-learner to construct the final model. Since the MVC algorithm is based on the multi-view learning framework, it is able to use any conventional method to mine data from relational databases.

IV. COMPARATIVE ANALYSIS

TABLE I. COMPARATIVE ANALYSIS

	Algorithm	Accuracy (%)	Time (s)	Reference
Flattening Approach	Relaggs	85.1	12.80	[19]
ILP based MRC	FOIL	85.7	2.30	[20]
	TILDE	85.6	1.40	[19]
	RIBL	86.5	-	[19]
Association based MRC	MrCAR	89.3	2.6	[18]
Relational database based MRC	MRDTL	87.5	52.15	[12]
	MRDTL-2	87.5	28.45	[12]
	CrossMine	85.7	1.00	[20]
Multiple view Approach	MVC	86.7	3.00	[20]

Based on the study of approaches related to multirelational classification, above table is created. Table I represent results performed on Mutagenesis dataset which represents accuracy and running time of different approaches.

Flattening approach has lowest accuracy because it has several limitations. Firstly, in flattening approach extensive pre-processing efforts are needed and after pre-processing the resulting flat file may contain large amounts of NULL values. Second, flattening relational data often results in a very big table with large numbers of additional attributes, which causes further scaling challenges and over-fitting problems for propositional algorithms.

In upgrading approach, ILP based approaches usually show unsatisfactory performance when handling noisy or numeric-value business data [19]. For logic based classification, relational data should be transformed into logic programs in pre-processing stage, which determines the relatively weak relation in database. It also Suffer from scalability problems with regard to the number of relations in database [13]. ILP has lack of results with respect to efficient theoretical complexity and the algorithms typically require more computational power. But as shown in above table I, ILP based approaches achieve good classification accuracy because it includes additional information such as background knowledge in the form of computer programs.

Association based MRC mainly deals with rules and rules are easiest way to express knowledge but many rules produced by standard classification systems are difficult to understand because these systems often use only domain independent biases and heuristics [1]. As shown in table I., MrCAR has higher accuracy then existing multi relational classification algorithms because the rules discovered by MrCAR have more comprehensive characterization of databases. Also associative classification makes association rule mining techniques applicable to classification tasks. The user can decide to mine both association rules and a classification model in the same data mining process which is very important benefit.

In relational database based MRC, tuple id propagation approach which is used in CrossMine improves efficiency because of small amount of data transfer and avoids the high cost of physical joins so reduce cost of space and time also. But in tuple id propagation sometimes too many IDs may be propagated to each tuple in a relation, which makes it hard to limit the time/space complexity of the algorithm. As shown in table I, MRDTL algorithm requires highest running time because it uses selection graph. And selection graph is a major bottleneck in terms of the running time of the algorithm [23]. MRTDTL-2 is proposed to improve the calculation efficiency and information loss of MRDTL which decrease the running time but not as well as other algorithms.

Multiple view approach uses multi view learning which describes the setting of learning from data where observations are represented by multiple independent sets of features and learning from these representations separately can lead to better gains than merging them into a single dataset. MVC algorithm excludes the need to transform multiple inter-connected tables into a universal relation. In Relational database each entity thus intuitively corresponds to a different concept or view of the problem domain. However in some cases, two attributes in different tables may be related. That is, the combination of these two seemingly unrelated attributes may provide us new knowledge about the problem domain. Currently, algorithm based on multiple view approach does not take this scenario into considerations [19].

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

V. CONCLUSION

Multi-relational data mining (MRDM) approaches look for patterns that involve multiple tables (relations) from a relational database. Relational classification has got more and more great concern from the large number of international researchers. In this paper, we presented different approaches based on relational classification like ILP based relational classification, Relational database based relational classification, Association or frequent pattern based relational classification, and multiple view based relational classification. Also comparative study for different approaches has been done with their benefits and limitations.

REFERENCES

1. 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, DOI: 10.5120/1740-2366, January 2011
2. Raymond J. Mooney, Prem Melville, Lappoon Rupert Tang, "Relational Data Mining with Inductive Logic Programming for Link Discovery", Appears in the Proceedings of the National Science Foundation Workshop on Next Generation Data Mining, Nov. 2002
3. Blockeel, H. , Top-down induction of first order logical decision trees, Artificial Intelligence Journal, vol.101,pp.285-297, 1998
4. Kramer S. , Widmer G. , "Inducing Classification and Regression Tress in First Order Logic", Relational Data Mining, pp.140-159, Springer, 2001
5. Lise Getoor, Nir Friedman, Daphne Koller, Avi Pfeffer, Ben Taskar, "Probabilistic Relational Models", In An Introduction to Statistical Relational Learning, August 2007
6. Getoor L. , "Multi-Relational Data Mining was using probabilistic Models Research Summary", In Proc. Of 1st workshop in MRDM, 2001
7. Stephen Muggleton, "Learning Stochastic Logic Programs", In Proceedings of the AAAI-2000 Workshop on Learning Statistical Models from Relational Data, Technical Report WS-00-06, pp. 36-41, 2000
8. Emde W. , Wettschereck D. , "Relational instance based learning", In Proceedings of the 13th Int. Conference on Machine Learning, Morgan Kaufmann, San Mateo, CA, 122-130, 1996
9. Cumby C. , Roth D. , "On kernel methods for relational learning", In Proceedings of 20th International Conf. On Machine Learning (ICML-2003,) Washington, 2003
10. Zhen Peng, Lifeng Wu, Xiaoju Wang, "Research on Multi-Relational Classification Approaches", International Conference on Computational Intelligence and Natural Computing, IEEE, 2009
11. Hector Ariel Leiva, Shashi Gadia, Drena Dobbs, "MRDTL: A multi-relational decision tree learning algorithm", Proceedings of the 13th International Conference on Inductive Logic Programming, 2002
12. Atramentov A. , Leiva H. , Honavar V. , "A Multirelational Decision Tree Learning Algorithm - Implementation and Experiments", ILP LNCS, Vol.2835, pp. 38-56, 2003
13. Yin X, Han J, Yu PS, "CrossMine: Efficient Classification across Multiple Database Relations". In Proceedings of 20th Int. Conf. on Data Engineering (ICDE'04), 2004
14. Liu H. , Yin X. ,Han J. , "A Efficient Multirelational Naïve Bayesian Classifier Based on Semantic Relationship Graph", In MRDM'05 Proceedings of 4th international workshop on MRDM, 2005
15. X. Yan, J. Han, "gSpan: Graph-based substructure pattern mining". In Proc. 2002 Int. Conf. on Data Mining (ICDM'02), Maebashi, Japan, Dec. 2002.
16. L. Dehaspe, H. Toivonen, "Discovery of Relational Association Rules", Springer-Verlag, 2000.
17. Dehaspe L. , Raedt D. , "Mining Association Rules in Multiple Relations", In Proceedings of the ILP, Springer- Verlang, London UK, pp.125-132, 1997
18. Yingqin Gu, Hongyan Liu, Jun He, Bo Hu, Xiaoyong Du, "MrCAR: A Multi-relational Classification Algorithm Based on Association Rules", IEEE, Web Information Systems and Mining, 2009. WISM 2009. International Conference, 31 December 2009
19. Hongyu Guo, Herna L. Viktor, "Multirelational Classification: A Multiple View Approach", ACM, KNOWLEDGE AND INFORMATION SYSTEMS, Volume 17, Number 3, 287-312, DOI: 10.1007/ s10115-008-0127-5, 2008
20. Hongyu Guo, Herna L. Viktor, "Mining relational databases with multi-view learning", ACM, DOI:10.1145/1090193.1090197, 2005
21. Sašo Džeroski, "Multi-relational data mining: an introduction", Published in ACM SIGKDD Explorations Newsletter Homepage archive, Volume 5 Issue 1, July 2003

AUTHOR PROFILE

Shraddha Modi is a M.Tech student in the U and P.U. Patel Department of Computer Engineering at Chandubhai S Patel Institute of Technology, Changa, India. She graduated from Hemchandracharya North Gujarat University, Patan, India with a B.E. degree in Computer Engineering in 2009. Currently in dissertation, her research work focuses on Multi relational Classification in Multi Relational Data Mining. She expects to obtain the M.Tech degree in 2012.

Amit Thakkar has received his B.E degree in Information Technology from Gujarat University, Gujarat, India in 2002 and master Degree from Dharmsinh Desai University, Gujarat, India in 2007. He has joined his Ph.D in the area of Multi relational Classification at Kadi Sarvavishvidhalaya University, Gandhinagar, India in June 2010.
Since 2002 he has been with faculty of Engineering and Technology, Charotar University of Science and Technology, Changa, Gujarat, Where he is currently working as an Associate Professor in the Department of Information Technology. He has published more than 20 research papers in the field of data mining and web technology. His current research interest includes Multi relational Data Mining, Relational Classification and Associate Classification.

Amit Ganatra has received his B.E degree in Computer Engineering from Gujarat University, Gujarat, India in 2000 and master Degree from Dharmsinh Desai University, Gujarat, India in 2004. He has joined his Ph.D in the area of Multiple Classifier System(Information Fusion) at Kadi Sarvavishvidhalaya University, Gandhinagar, India in August 2008.
Since 2000 he has been with faculty of Engineering and Technology, Charotar University of Science and Technology, Changa, Gujarat, Where he is currently working as an Associate Professor in the Department of Computer Engineering. He has published more than 50 research papers in the field of data mining and Artificial Intelligence. His current research interest includes Multiple Classifier System, Sequence Pattern Mining.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)