A Study on Classification Approaches across Multiple Database Relations

Dr. M. Thangaraj
Associate Professor
Madurai Kamaraj University,Madurai,
Tamil Nadu,India

C.R.Vijayalakshmi
Dean
Nadar Saraswathi College of Arts & Science, Theni
Tamil Nadu, India

ABSTRACT
Classification is an important task in data mining and machine learning, which has been studied extensively and has a wide range of applications. Lots of algorithms have been proposed to build accurate and scalable classifiers. Most of these algorithms can only applied to single "flat" relations, whereas in the real world most data are stored in multiple tables. As converting data from multiple relations into single flat relation usually causes many problems, development of classification across multiple database relations becomes important. In this paper, we present the several kinds of classification method across multiple database relations including Inductive Logic Programming (ILP), Relational database, Emerging Pattern, Associative approaches and their characteristics, the comparisons in detail.

General Terms
Classification, Decision tree, Emerging pattern.

Keywords
Multi-relational classification, Inductive logic programming, Selection graph, Tuple ID propagation

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, the multi relational data mining has become a field with 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.

Research Direction's Map:

The classification across multiple database relations is divided into two steps with the same propositional classification i). to learn classification model from examples ii). 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). ILP based MRC (LBRC), ii). Relational database based MRC (RBRC), iii). Emerging Patterns based MRC iv). Associative MRC. The Figure.1 shows the four categories of classification across multiple database relations. An extensive survey of literature was made to identify various research issues in this filed. The following five sections present different methods and the research directions in each area.

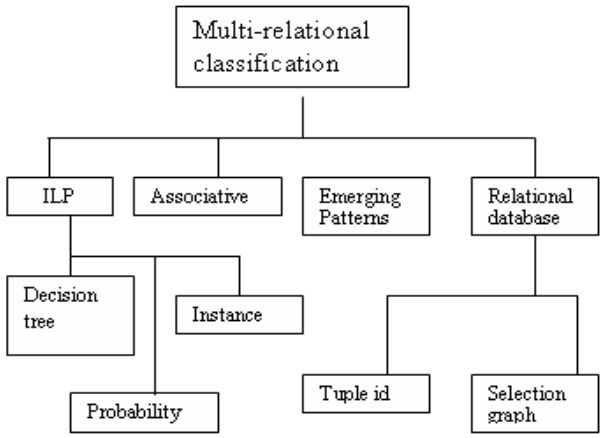

Fig.1. Research Direction's Map

2. ILP–BASED RELATIONAL CLASSIFICATION
Logic–based MRDM popularly known as Inductive Logic Programming [14], [13], [44], is the intersection of machine learning and logic programming. It is characterized by the use of logic for the representation of multi relational data. The LBRC search for syntactically legal hypotheses constructed from predicates that can predict the class labels of examples based on background knowledge. They achieve good classification accuracy in multi relational classification. They mainly include three categories – Decision tree relational classification, Instance based relational classification (RIBL and kernel) and Probability classification approach (PRM and SLP).

2.1 Decision tree relational classification approaches
Decision trees are trees that classify instances by sorting them based on feature values. 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. In general decision tree classifiers have good accuracy. There are two major classification algorithms for inducing relational decision trees TILDE [3] and SCART [32], that are upgraded from the two famous algorithms for inducing propositional decision trees (CART and C 4.5). 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 [31], [21], [22], [42], [7]. A PRM [20] 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) [39] 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.

2.3 Instance Relational Classification

Relational Instance Based Learning (RIBL) as first introduced in Emde and Wettschereck (1996), is a first-order instance-based learner that applies the basic principles of Instance Based Learning (IBL) to an instance space consisting of first-order descriptions [15], [30]. RIBL uses a *k*-nearest-neighbor algorithm that determines the value of *k* through cross-validation on the training set. It differs from the basic algorithm through its first-order learning task, similarity measure and weight learning scheme for predicates and arguments. RIBL turns the basic IBL problem into a first-order version by allowing each instance to be described by numerical, discrete attributes and by attributes of type object. RIBL was applied to practical problem of diterpene structure evaluation.

In [28], it computes similarity of instances with non-flat components (lists, terms). It predicts mRNA molecules and to automatically discover uncharacterized mRNA signal structure classes. In paper [29], the learning examples are stored in a relational database. It is based on relational algebra representations. It defines relational distances whose building blocks are distances between tuples of relations and distances between sets.

Kernel functions can project data in non-linear space into high dimensional linear hyper sphere feature spaces to classify the data according to the distances. The paper [43] exploits the notion of foreign keys to perform the leap from a flat attribute value representation to a structured representation that underlines relational learning. It uses direct sum kernel and kernel, which is derived by application of the R-Convolution kernel. The paper [19], [9] defines a frame for applying Kernel in a variety of structured data. The methods [18], [8] have been used in multi relational classification learning.

3. ASSOCIATIVE CLASSIFICATION

Associative classification [26] uses association mining techniques that search for frequently occurring patterns in large databases. The patterns may generate rules, which can be analyzed for use in classification. Several algorithms have been proposed for associative classification such as Classification based on Multiple Association Rule (CMAR) [36], Classification based on Predictive Association Rules (CPAR) [46]. 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 combine 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 [10] extends Apriori to mine the association rules in multiple relations. The paper [40] 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. A Multi-relational classification algorithm based on association rules is proposed in MrCAR [23]. It uses class frequent closed itemsets. It 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. The rules discovered by MrCAR have more comprehensive characterization of databases.

4. EMERGING PATTERN BASED CLASSIFICATION

The discovery of emerging patterns (EPs) is a descriptive data mining task defined for pre-classified data. Emerging patterns

are classes of regularities whose support significantly changes from one class to another. In [11], 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) [37], Classification by aggregating emerging patterns (CAEP) in [12], 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. CAEP is better for cases with few or even no jumping EPs whose supports meets a reasonable threshold , whereas JEP-Classifier is better when there are many jumping EPs.

The paper [16] 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.

To achieve much better accuracy and efficiency than the previously EP-based classifiers, an instance based classifier using EPs (DeEPs) is proposed in [35], [34]. This approach achieves high accuracy, because the instance-based approach enables DeEPs to pinpoint all EPs relevant to a test instance, some of which are missed by the eager-learning approaches. It also achieves high efficiency by using a series of data reduction and concise data-representation techniques. CAEP, JEP-Classifier, are the two relatives to DeEPs. DeEPs have considerable advantages on speed, and dimensional scalability over CAEP and the JEP-Classifier, because of its efficient ways to select the sharp and relevant EPs and to aggregate the discriminating power of individual EPs. Another advantage is that DeEPs can handle new training data without the need to retrain the classifier which is, commonly required by the eager learning based classifiers. This feature is extremely useful for practical applications where the training data must be frequently updated.

ConsEPMiner [1], which adopts a level wise, generate and test approach to discover EPs, which satisfy several constraints. All these methods assume that data to be mined are stored in a single table.

Mr.-EP [50], which discovers EPs from data scattered in multiple tables of a relational database. Generated EPs can capture the differences between objects of two classes which involve properties possibly spanned in separate data tables.

In [5], two EPs- based relational classifiers Multi-Relational Classification based on Aggregating Emerging Patterns (Mr-CAEP) and Multi Relational Probabilistic Emerging Patterns Based Classifier (Mr-PEPC) are proposed. Mr-CAEP upgrades the EP-based classifier CAEP from the propositional setting to the relational setting. It computes the membership score of an object to each class. The score is computed by aggregating a growth rate based function of the relational EPs covered by the object to be classified. In Mr-PEPC, relational emerging patterns are used to build a naïve Bayesian classifier which classifies any object by maximizing the posterior probability.

5. RELATIONAL DATABASE-BASED CLASSIFICATION

RBRC includes i) selection graph based relational classification ii) tuple ID propagation based relational classification. Selection graph model can use database language SQL to directly deal with relational tables of database. 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) [33] 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 -2 [2] which improved the calculation efficiency and information loss of MRDTL.

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. Multi-relational naïve bayes classifier Mr-SBC [4] 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.

CrossMine [49], [47] is a divide and conquer algorithm, which uses rules for classification. It searches for the best way to split the target relation into partitions, and then recursively works on each partition. It also employs selective sampling method, which makes it highly scalable with respect to the number of relations. Graph-NB [38] which upgrades Naïve Bayesian classifier, and use the semantic relationship graph (SRG) to describe the relationship and to avoid unnecessary joins among tables. To improve the accuracy, a pruning strategy named "cutting off" strategy is used to simplify the graph to avoid examining too many weakly linked tables.

The paper [48] proposed two methods for classification: CrossMine-Rule is a rule-based classifier and CrossMine-Tree, is decision tree based classifier. The comprehensive experiments demonstrate the high scalability and accuracy of CrossMine. The Relational decision tree (RDC) [25] is an extension of MRDTL algorithm with the usage of tuple ID propagation. For dealing with the missing attribute, a naïve bayes model for each attribute in a table is built based on the other attributes excluding the class attribute. The missing values are filled with the most likely predicted value by the naïve bayes predictor. It achieves higher efficiency and is more efficient in running time than MRDTL-2.

Classification with aggregation of Multiple Features (CLAMF) method is proposed in [17], 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 [45] proposed an extended SRG and a new counting method to construct new multi-relational naïve Bayesian classifier.

A multiple view strategy is proposed in [24], which enable us to classify relational objects by applying conventional data mining methods, while there is no need to flatten multiple relations to a universal one. It employs multiple view learners to separately capture essential information embedded in individual relation. The acquired knowledge is incorporated into a meta learning mechanism to construct the final model.

The paper [41] is based on two pruning strategy. Firstly, get rid of some attributes based on the foil gain, and make use of relationship between the accuracy of the attribute to give them the second pruning. In the second step, the remaining attributes are used to classify the data. This method guarantees the accuracy and also saves much time. The Semantic Relationship Graph for Multi-relational Bayesian Classification (SRG-BC) is proposed in [6], which integrates relation selection and feature selection into the multi-relational Bayesian classifier and uses optimized SRG to describe the relationship between tables in the database. Based on this optimized SRG, not only the search space becomes smaller, but also the accuracy is much improved.

In [27], novel approach proposed to conduct both Feature and Relation Selection for efficient multi-relational classification. In this approach symmetrical uncertainty is used to measure correlation between attributes in a table or cross tables. It also measures the correlation between a table and a class attribute. Based on the correlations, it selects relevant attributes and tables from the database.

6. CONCLUSION AND PERSPECTIVE

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, Associative based approaches.

1. Patterns discovered by ILP systems are expressed as logic programs. It is an important subset of first order logic. The first order logic clearly corresponds to concepts of relational databases. LBRC can express more complex patterns. But it needs to transform relational data into logic programs in preprocessing stage, which determines the relatively weak relation in database. This conversion leads to lot of inconvenience in real application.

2. The associative classification approach helps to solve the understandability problem that may occur with some classification methods. Indeed, many rules produced by standard classification systems are difficult to understand because these systems often use only domain independent biases and heuristics, which may not fulfill user's expectation.

3. The Emerging Pattern based classifiers take advantages from the fact that EPs provide features which better discriminate classes that association rules do. The scalability factor is the major issue.

4. Most of today's structured data is stored in relational databases and the representation of RBRC is also relational in nature. So that it need not transform the structured data to any other form. Furthermore, the major issues to be solved are how to directly use database operation to achieve tuple ID propagation based classification and the scalability. For certain multi-relational classification tasks, some tables may also be redundant. Eliminating redundancy among tables is another challenging task.

The Relational Classification challenges are RC approaches are mainly from ILP technology, which is developed from propositional classification and also how to extend other proposition methods to LBRC. RBRC opens up a new way for relational classification research. At present, the focus of the selection graph based relational classification is on MRDM with decision tree inductive methods.

7. REFERENCES

[1] Appice, A., Ceci M., Malgieri C., Maleraba D. 2007. Discovering relational emerging patters, AI*AI 2007, LNCS (LNAI), Vol. 4733, 206-217, Springer, Heidelberg.

[2] Atramentov, A., Leiva, H., and Honavar, V. 2003. A Multi-relational Decision Tree Learning Algorithm-Implementation and Experiments, ILP LNCS, Vol.2835, pp. 38-56.

[3] Blockeel, H. 1998. Top-down induction of first order logical decision trees, Artificial Intelligence Journal, vol.101, pp.285-297.

[4] Ceci, M., Appice, A., and Malerb, D.2003. Mr-SBC: a Multi-Relational Naïve Bayes Classifier, Knowledge Discovery in Databases PKDD 2003, LNAI, vol.2838, pp.95-106.

[5] Ceci, M., Appice, A., Maleraba, D. 2008. Emerging Pattern Based Classification in Relational Data Mining, DEXA 2008, LNCS, vol.5181, pp.283-296.

[6] Chen, H., Liu, H., Han, J., Yin, X. 2009. Exploring Optimization of Semantic Relationship Graph for Multi-relational Bayesian Classification, In Decision Support System, Vol.48, pp.112-121.

[7] Cheng, Q. 2007. PRM based multi relational association rule mining, Thesis Report, Simon Fraser University.

[8] Cumby, C., Roth, D. 2003. On kernel methods for relational learning, In Proceedings of 20th International Conf. on Machine Learning (ICML-2003,) Washington.

[9] De Raedt, L. 2008. Kernels and Distances for Structured Data: Logical and Relational Learning, 289-324, Springer.

[10] Dehaspe, L., Raedt, D. 1997. Mining Association Rules in Multiple Relations, In Proceedings of the ILP, Springer-Verlang, London UK, pp.125-132.

[11] Dong, G., Li, J. 1999. Efficient mining of emerging patterns: Discovery trends and differences, In International Conference on Knowledge Discovery and Data Mining, pp. 43-52. ACM Press, New York.

[12] Dong, G., Zhang, X., Wong, L., and Li, J. 1999. CAEP: Classification by aggregating emerging patterns, In Proceedings of the Second International Conference on Discovery Science, Tokyo, Japan, pages 30-42.

[13] Dzeroski, S. 2003. Multi-relational data mining: an introduction, [J]. SIGKDD Explorations, vol. 5(1):1-16.

[14] Dzeroski, S., Lavtac, N. 2001. eds, Relational data mining, Berlin: Springer.

[15] Emde, W., Wettschereck, D. 1996. Relational instance–based learning, In Proceedings of the 13th Int. Conference on Machine Learning, Morgan Kaufmann, San Mateo, CA, 122-130.

[16] Fan, H., Ramamonanarao, K. 2002. An efficient single scan algorithm for mining essential jumping emerging patterns for classification, In Pacific-Asia Conference on Knowldege Discovery and Data Mining , pp.456-462.

[17] Frank, R., Moser, F., Ester, M. 2007. A Method for Multi-Relational Classification Using Single and Multi-Feature Aggregation Functions, In Proceedings of 11th European Conf. on PKDD, Springer, Verlag Berlin Heidelberg.

[18] Gaertner, T., Flach, P., Kowalczyk, A. 2002. Multi-instance kernels, In Proceedings of 19th International Conf. on Machine Learning, pp.179-186.

[19] Gaertner, T., Lloyed, J., Flach, P. 2004. Kernels and distances for structured data, In Machine Learning, vol.57, No.3, pp.205-332.

[20] Getoor, L. 2001. Multi-Relational Data Mining was using probabilistic Models Research Summary, In Proc. Of 1st workshop in MRDM.

[21] Getoor, L., Friedman, N., Koller, D., and Pfeffer, A. 2001. Learning Probabilistic Relational Models, pp.307-355, Springer Verlage, New York.

[22] Getoor, L., Friedman, N., Koller, D., Taskar, B. 2001. Learning Probabilistic Models of Relational Structure, ICMl'01 Proceedings of 8th International Conference on Machine Learning.

[23] Gu,Y., Liu, H., He, J. 2009. MrCAR: A Multi relational Classification Algorithm based on Association Rules, Int. Conf. on Web Information Systems and Mining, pp.256-260.

[24] Guo, H., Herna, L., Viktor. 2008. Multirelational classification: a multiple view approach, Knowl. Inf. Systems, vol.17, pp.287–312, Springer-Verlag London.

[25] Guo, JF., Li, J., Bian, WF. 2007. An Efficient Relational Decision Tree Classification Algorithm, In proceedings of 3rd ICNC, vol.3.

[26] Han, J., Kamber, M. 2007. Data Mining: Concepts and Techniques", 2nd Edition, Morgan Kaufmann.

[27] H, J. Liu,H.,et at, 2010. Selecting Effective Features and Relations For EfficientMulti-Relational Classification, Computational Intelligence, Vol 26, No.3.

[28] Horva, T., Wrobel, S., Bohnebeck, U. 2001. Relational Instance-Based Learning with Lists and Terms, Machine Learning, vol.43, pp.53–80.

[29] Kalousis, A., Woznica, A., and Hilario, M. 2006. A unifying framework for relational distance-based learning founded on relational algebra, Technical Report, University of Geneva..

[30] Kirsten, M., Wrobel, S., Horvath, T. 2002.Distance Based Approaches to Relational Learning and Clustering: Relational Data Mining, Morgan Kaufmann (2005) 6, pp.213-232, springer, Heidelberg.

[31] Koller, Pfeffer, A. 1998. Probabilistic frame-based systems, In Proceedings of the 15th National Conference on Artificial Intelligence, pp. 580–587, Madison, WI.

[32] Kramer, S., Widmer, G. 2001. Inducing Classification and Regression Tress in First Order Logic: Relational Data Mining, pp.140-159, Springer.

[33] Leiva, HA. 2002. A multi-relational decision tree learning algorithm, ISU-CS-TR, Lowa State University, pp.02-12.

[34] Li, J., Dong, G., Ramamohanarao, K., Wong, L. 2004. A new instance-based lazy discovery and classification system", Machine Learning, vol.54, No.2, pp0. 99-124.

[35] Li, J., Dong, Ramamohanarao, K. 2000. DeEPs: Instance-based classification by emerging patterns, Technical Report, Dept of CSSE, University of Melbourne.

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

[37] Li, J., Dong, G., and Ramamohanarao, K.1999. JEP-Classifier. Classification by Aggregating Jumping Emerging Patterns, Technical report, Univ of Melbourne.

[38] Liu, H., Yin, X., and Han, J. 2005. A Efficient Multi-relational Naïve Bayesian Classifier Based on Semantic Relationship Graph, In MRDM'05 Proceedings of 4th international workshop on MRDM.

[39] Muggleton, SH. 2000. 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.

[40] Nijssen S., Kok, J. 2001. Faster Association Rules for Multiple Relations, In Proceedings of the IJCAI, pp.891-896.

[41] Pan Cao, Wang Hong-yuan. 2009. Multi-relational classification on the basis of the attribute reduction twice, Communication and Computer,Vol. 6, No.11. pp: 49-52.

[42] Taskar B, Segal E, Koller D, "Probabilistic Classification and Clustering in Relational Data", In Proceedings of International Conf. Artificial Intelligence, vol.2, 2001.

[43] Woznica, A, Kalousis A, and Hilario M, " Kernel-based distances for relational learning," In Proceedings of the workshop on Multi-Relational Data Mining at KDD -2004.

[44] Wrobel S, "Inductive Logic Programming for Knowledge Discovery in Databases: Relational Data Mining", Berlin: Springer, pp.74-101, 2001.

[45] Xu GM, Yang BR, Qin YQ, " New multi relational naïve Bayesian classifier", Systems Engineering and Electronics, vol. 30, No.4, pp 655-655, 2008.

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

[47] Yin X, Han J, and Yu PS, "CrossMine: Efficient Classification across Multiple Database Relations". In Proceedings of 20th Int. Conf. on Data Engineering (ICDE'04), 2004.

[48] Yin X, Han J, and Yu PS, "Efficient Classification across Multiple Database Relations: A CrossMine Approach", IEEE Transactions on Knowledge and Data Engineering, Vol 16, No.6, 2006.

[49] Yin, X., Han, J., Yang, J. 2003. Efficient Multi-relational Classification by Tuple ID Propagation, In Proceedings of KDD workshop on MRDM.

[50] Zhang, X., Dong, G., Ramamohanarao, K. 2000. Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets, In Proceedings of 6th SIGKDD international conference on Knowledge Discovery and Data Mining, pp. 310-314.