Survey of Multi Relational Classification (MRC) Approaches & Current Research Challenges in the field of MRC based on Multi-View Learning

Amit Thakkar, Y P Kosta

Abstract*— 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. Many of these applications cannot be solved using traditional data mining algorithms. This observation forms the main motivation for the multi-disciplinary field of Multi-Relational Data Mining (MRDM). Unfortunately, existing "upgrading" approaches, especially those using Logic Programming techniques, often suffer not only from poor scalability when dealing with complex database schemas but also from unsatisfactory predictive performance while handling noisy or numeric values in real-world applications. However, "flattening" strategies tend to require considerable time and effort for the data transformation, result in losing the compact representations of the normalized databases, and produce an extremely large table with huge number of additional attributes and numerous NULL values (missing values). As a result, these difficulties have prevented a wider application of multi relational mining, and post an urgent challenge to the data mining community. To address the above mentioned problems, this article introduces a multiple view approach—where neither "upgrading" nor "flattening" is required— to bridge the gap between propositional learning algorithms and relational databases and current research challenges in the field of Multi relational classification based on Multi View Learning.

***Index Terms*—Multi Relational Data Mining, Propositional Learning, Multi Relational Classification, Relational Learning.**

I. INTRODUCTION

Most real-world data are stored in relational databases. So to classify objects in one relation, other relations provide crucial information. Traditional mechanism cannot convert relational data into a single table without expert knowledge or loosing essential information. Multi-relational classification automatically classifies objects using multiple relations. Vast amounts of real world data are routinely collected into and organized in relational databases. Most of today's structured data is stored in relational databases. Thus, the task of learning from relational data has begun to receive significant attention in the literature. Unfortunately, most methods only utilize "flat" data representations. Thus, to apply these single-table data mining techniques, we are forced to incur a computational penalty by first converting the data into this "flat" form. Patterns of activity that, in isolation, are of limited significance for classification but, when combined/related, will improve the performance of system. Multi relational classification aims at discovering useful patterns across multiple inter-connected tables (relations) in a relational database. Traditional machine learning approaches assume a random sample of homogeneous data from single relation but real world data sets are multi-relational and heterogeneous. Current solution does not scale well and cannot realistically be applied when considering database containing huge amount of data.

II. CATEGARIZATION OF DATA MINING (DM) TECHNIOQUE

Data Mining Technique can be broadly divided into two category Propositional Data Mining and Multi Relational Data Mining[4].

A. *Propositional Data Mining (Unique Table Approach)*

Most classifiers works on a single table (attribute-value learning) with a fixed set of attributes,so their use is restrictive in DM applications with multiple tables. It is possible to construct, by hand, a single table by performing a relational join operation on multiple tables using propositional logic. For one-to-one and many-to-one relationships, one can join in the extra fields to the original relation without problems. For one to many relationships, there are two ways to handle them. The first one is just compute the join, but this leads to data redundancy, missing values, statistical skew, and loss of meaning. A single instance in the original database is mapped onto multiple instances in the new table, which is problematic. The second way is aggregate the information in different tuples representing the same individual into one tuple after computing the join. This removes the problems mentioned above, but causes loss of information because details originally present have been summarized away. [9]

B. *Multi Relational Data Mining (MRDM)*

The database consists of a collection of tables (a relational database). Records in each table represent parts, and individuals can be reconstructed by joining over the foreign key relations between the tables.

Manuscript Received on December 10, 2011

Amit Thakkar, Department of Information Technology, Charotar Institute of Technology (Faculty of Technology and Engineering), Charotar University of Technology Changa, Anand, Gujarat, India, 388421 amitthakkar.it@charusat.ac.in

Y P Kosta Marwadi Education Foundation's Group of Institutions, Rajkot-360 003, Gujarat, India,ypkosta@yahoo.com

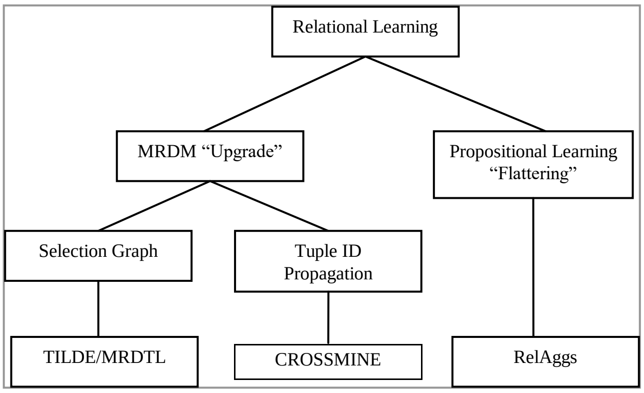


Fig 1: Classification of Relational Learning Approaches

MRDM Based on Selection Graph

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.[2].

MRDM Based on TupleID propogation

Tuple ID propagation model joins relational tables through propagating tuple ID. Tuple ID propagation is to propagate class ID from target table to other tables in relational database. The method do not implement so many physical connections as ILP technique but only once virtually joining, not convert into logic program so as to reduce the costs of time and space. [3]

In essence, tuple ID propagation is a method for virtually joining non-target relations with the target one, and it is a simple but efficient method. It is much less costly than physical join in both time and space. Suppose the primary key of the target relation is an attribute of integers, which represents the ID of each target tuple. Tuple ID propagation is a flexible and efficient method. IDs and their associated class labels can be easily propagated from one relation to another relation. By doing so, the next computing tasks can do with little redundant, and the required space is also small than the physical join.

CrossMine is a scalable and accurate approach for multi-relational classification.[3] Its basic idea is to propagate the tuple IDs (together with their associated class labels) in the target relation to other relations. In the relation to whom the IDs are propagated, each tuple t is associated with a set of IDs, which represent the target tuples that are joinable with t. Besides propagating IDs from target relation to non-target relations, one can propagate the IDs transitively to additional non-target relations to search for good predicates among many relations.

C. Weaknesses of Existing Approaches

Existing approaches either, "upgrade" propositional learning methods to deal with multiple interlinked relations or "flattening" multiple tables into a single flat file.Unfortunately, existing "upgrading" approaches, especially those using Logic Programming techniques, often suffer not only from poor scalability when dealing with complex database schemas but also from unsatisfactory predictive performance while handling noisy or numeric values in real-world applications. Contrary to those "upgrading" algorithms, "flattening" methods aim to directly use propositional learning algorithms by transforming multiple relations of a relational database into a universal flat file. [9] However, "flattening" strategies tend to require considerable time and effort for the data transformation, result in losing the compact representations of the normalized databases, and produce an extremely large table with huge number of additional attributes and numerous NULL values (missing values). As a result, these difficulties have prevented a wider application of multirelational mining, and post an urgent challenge to the data mining community.

III. MINING RELATIONAL DATA MINING WITH MULTI-VIEW LEARNING

To address the above mentioned problems, this paper introduces a multiple view approach where neither "upgrading" nor "flattening" is required to bridge the gap between propositional learning algorithms and relational databases. On the one hand, our approach enables traditional data mining methods to utilize information across multiple relations in a relational database. Hence, many efficient and accurate propositional learning algorithms make a wider choice of mining methods available for multirelational data mining applications. On the other hand, the strategy excludes the need to transform multiple inter-connected tables into a universal relation. Therefore, the above mentioned shortcomings resulted from the "flattening" process can be avoided. The approach was inspired by a promising new strategy, i.e. multi-view learning. Multi-view learning describes the problem of learning from multiple independent sets of features, i.e. views, of the presented data.

In fact, a multi-view learning problem with *n* views can be seen as *n* strongly *uncorrelated feature sets* which are distributed in the multiple relations of a relational database.

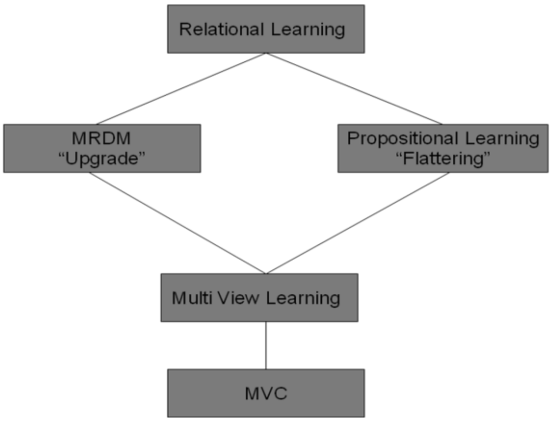


Fig 2: Multi View Learning

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The framework of Multi View Learning has been applied successfully to many real-world applications such as information extraction, text classification, and face recognition. We argue that this approach offers a technique which can be used to learn concepts in relational databases. In the multi-view setting, each view can be expressed in terms of a set of disjoint features of the training data. The target concept is learned on each of these separate views using the features present. The results from the views are then combined, each contributing to the learning task.[9]

The multi-view learning problem with n views can be seen as n inter-dependent relations and are thus applicable to multi-relational learning. This is the basic scenario in multi relational learning problem. As an example, consider the loan problem in the PKDD 99 discovery challenge [5], where the banking database consists of eight relations. Each relation describes different characteristics of a client. For example, the Client relation contains a customers' age, the Account relation identifies a customers' banking account information, and the Card relation refers to customers' credit card details. In other words, each relation from this database provides different types of information or views of the concept to be learned, i.e. whether the customer is good or not. This problem is therefore a perfect candidate for multi-view learning.

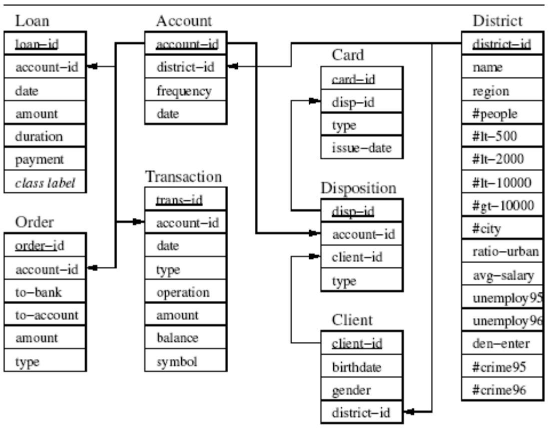


Fig 3: PKDD 99 Discovery Challenge Sample Database

The Multi View Classification (MVC) approach employs the multi-view learning framework to operate directly on multi-relational databases with conventional data mining methods. The approach works as follows.

1) Information Propagation Stage: The Information Propagation Stage, first of all, constructs training data sets for use by a number of view learners, using a relational database as input. The *Information Propagation Element* propagates essential information from the target relation to the background relations, based on the foreign key links. In this way, each resulting relation contains efficient and various information, which then enables a propositional learner to efficiently learn the target concept.

2) Aggregation Stage: After the *Information Propagation*, the *Aggregation Stage* summarizes information embedded in multiple tuples and squeeze them into one row. This procedure is applied to each of the data sets constructed in the *Information Propagation Stage*. In this stage, aggregation functions are applied to each background relation (to which the essential information from the target relation were propagated). By applying the basic aggregation functions in SQL, new features are created to summarize information stored in multiple tuples. Each newly constructed background relation is then used as training data for a particular view learner.

3) Multiple Views Construction Stage: In the third phase of the MRC algorithm, the *Multiple Views Construction Stage* constructs various hypotheses on the target concept, based on the multiple training data sets given by the *Aggregation Stage*. Conventional single-table data mining methods (view learners) are used in order to learn the target concept from each view of the database separately. In this stage, a number of view learners, which differ from one another, are trained.

4) View Validation Stage: All view learners constructed in the *Multiple Views Construction Stage* is then evaluated in the the *View Validation Stage*. The trained view learners need to be validated before being used by the meta learner. This processing is needed to ensure that they are sufficiently able to learn the target concept on their respective training sets. In addition, strongly uncorrelated view learners are preferred.

5) View Combination Stage: In the last step of the MRC strategy, the resulting multiple view learners (from the *View Validation Stage*) are incorporated into a meta learner to construct the final classification model. The meta learner is called upon to produce a function to control how the view learners work together, to achieve maximum classification accuracy. This function, along with the hypotheses constructed by each of the view learners, constitutes the final model.

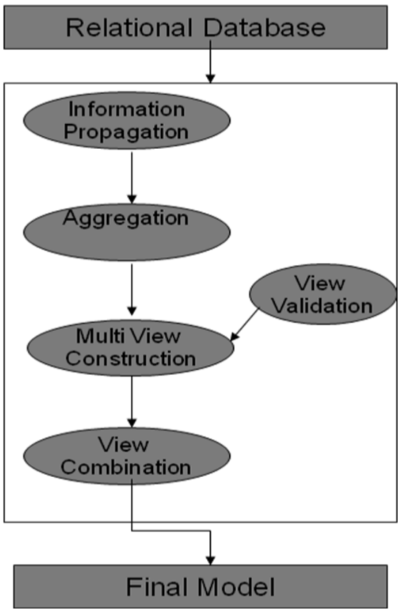
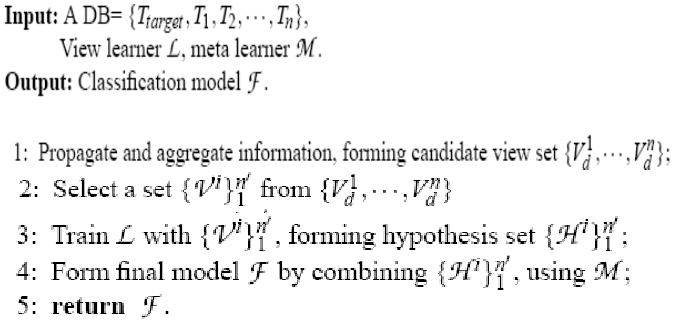


Fig 4 Multi View Learning Algorithm

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.

Popular ensemble methods such as boosting, bagging and stacking, focus on constructing different hypotheses from subsets of learning instances. In contrast, an important characteristic of the multi-view learning is that this approach is more interested in learning independent models from disjoint features of the training data. That is, multi-view learning learns separate views from various disjoint-feature-based aspects of the data, leading to independent views on the target concept.



```
Input: A DB= $\{T_{target}, T_1, T_2, \cdots, T_n\}$,
       View learner $\mathcal{L}$, meta learner $\mathcal{M}$.
Output: Classification model $\mathcal{F}$.

Propagate and aggregate information, forming candidate view set $\{V_d^1, \cdots, V_d^n\}$;
Select a set $\{\mathcal{V}^i\}_1^{n'}$ from $\{V_d^1, \cdots, V_d^n\}$
Train $\mathcal{L}$ with $\{\mathcal{V}^i\}_1^{n'}$, forming hypothesis set $\{\mathcal{H}^i\}_1^{n'}$;
Form final model $\mathcal{F}$ by combining $\{\mathcal{H}^i\}_1^{n'}$, using $\mathcal{M}$;
return $\mathcal{F}$.
```

IV. COMPARATIVE ANALYSIS OF MRDM APPROACHES

Table 1: Comparison of Different MRDM Approaches

Parameters	Approaches Based on Relational Databases		
	Selection Graph (2002)	Tuple ID Propagation (2004)	Multi View Learning (2008)
Scalability	Low	Low	High
Expressiveness	High	High	Low
Time to Learn	Less	More	Less
Numerical Attributes	Yes	Yes	Yes
Normalized Structure	No	No	Yes
Direct mining on real world Database	No	No	Yes
Incorporation of Traditional Single Table Algorithm	No	No	Yes
Integration of Advance Techniques	No	No	Yes
Learning Using Heterogeneous Classifier	No	No	Yes
Learning on Heterogeneous Data	No	No	Yes
Support of Incremental Design	Low	Low	High

Above comparative analysis shows some strong points of Multi View Learning (MVL) compare to other approaches. The first is that the relational database is able to keep its compact representation and normalized structure. The second is that it uses a framework that is able to directly incorporate any traditional single-table data mining algorithm. The third is that the multi-view learning framework is highly efficient for mining relational databases in term of running time. It is very easy to integrate the advanced techniques developed by the ensemble and heterogeneous learning communities Efficient and Practical solution for classifying enterprise data in relational databases. The algorithm should be scalable in the size of the database tables as well as the number of such tables. It should work directly out of existing databases since it is not feasible (From both time and space perspective) to transform data even in individual tables to different forms of representations. It should be able to work without requiring collation or replication of data from all tables. Since there may be different authorization controls for different database tables, the algorithm should be modular with an ability to execute different parts of the algorithm by different users/stages. The algorithm should effectively leverage the semantic grouping that is implicit in the design of RDBMS.

Experimental Study and Discussion [9]

This section provides the results obtained for the MRC algorithm on benchmark real-world databases. These results are presented in comparison, in terms of predictive performance achieved and running time needed, with four other well-known multirelational data mining systems, namely CrossMine method (Yin et al., 2004), TILDE first-order logical trees (Blockeel and Raedt, 1998), and RelAggs algorithm (Krogel, 2005). Six learning tasks derived from four standard real-world databases were used to evaluate our algorithm. The four benchmark databases, namely the Financial, Mutagenesis, Thrombosis, and Warehouse databases, come from different application domains, have variant relational structures, consist of different numbers of tuples in the entire database and in the target relation, and present varying degree of class distribution in the target relation.

Table 1: Summary of Dataset Used

Database	# tuples in the target	No of Related relations	Target class distribution	No of tuples in task
Mutagenesis	188	3	125:63	15,218
Financial	682	7	606:76	76,264
EMCL	7,329	8	3705:3624	197,478
Thrombosis	770	4	695:75	4,780

We present the predictive accuracy obtained for each of the six learning tasks in Table II. For each data set in Table II, the highest results are highlighted in *bold*. In addition, in the parentheses of this table, we provide the accuracy *gains (denoted by "+")* or *lost (denoted by "-")* of each approach, compared to that of the MRC algorithm with view validation applied. To evaluate the performance of the MRC strategy in terms of run time, we also provide the running time needed (in seconds) for each learning tasks in Table III, where the best results for each data set are also highlighted in *bold*. The predictive performance results, as presented in Table II, show that the MRC algorithm appears to consistently reduce the error rate for almost all of the data sets, when compared to the CrossMine, and TILDE methods. In addition, our results, as shown in Table II, also indicate that, in many cases the error rate reduction achieved by the MRC approaches is large.

In terms of running time needed, one can see from the experimental results (shown in Table III) that the MRC methods achieved very promising outcomes, when compared to the other four well-known algorithms. The MRC algorithm meaningfully reduced the running time needed for most of the cases.

Table 2:Accuracies obtained using different methods of MRC

Database	RelAgges	CrossMine	TILDE	MRC
Mutagenesis	85.1	85.7	85.6	**86.7**
Financial	92.1	90.3	88.9	**93.4**
EMCL	**88.0**	85.3	53.7	87.6
Thrombosis	100	90.0	90.4	**100**

Table 3: Runing time required by different method of MRC

Database	RelAgges	CrossMine	TILDE	MRC
Mutagenesis	12.80	**1.00**	1.40	3.26
Financial	89.54	11.60	1051.90	**5.59**
EMCL	1703.58	570.90	1108.60	**418.37**
Thrombosis	**0.72**	75.70	75.70	1.03

V. RESEARCH CHALLENGES

Following are the Current Research Challenges in the Field of MRC based on Multiple View Learning.

- The major challenges come from, the large high dimensional search spaces due to many attributes in multiple relations and the high computational cost in feature selection and classifier construction due to the high complexity in the structure of multiple relations. [17].
- The idea of using heterogeneous learners will further increase understanding of the multiple views learning scheme.
- To study applying data preprocessing techniques such as feature selection in order to further improve the performance of the MVC algorithms.
- Also, prior work has shown that more complex aggregation functions can improve the generalization accuracy of relational learning. It would also be interesting to investigate this for MVC.
- 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.
- In addition, another area for future work is the study which extends this approach to deal with relational data stored in the form of graph and social network.
- To investigate the behavior of the multiple view learning frameworks, while developing more sophisticated view construction techniques. In other words, the view construction procedure will search the entire feature space in order to determine how to better group the features into different views.
- Another area for future work is employing relational data mining algorithms as hypotheses construction methods, rather than generating relational features and then applying single-table learning strategies, while training a set of diverse individual view learners.
- Research has shown that popular ensemble methods such as Bagging, Boosting, and Stacking can significantly improve the predictive performance of an individual model in some cases. Through employing relational mining algorithms as view learners in the multiple view learning framework, will be able to explore the impact of popular ensemble techniques on the relational learning strategies.
- A novel approach is needed which can conduct both Feature and Relation Selection for efficient multi-relational classification.
- Join Graph can be further pruned to improve the classification time, by eliminating tables that may not really contribute much to the overall classification task.
- 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.

VI. CONCLUSION

In this paper we have demonstrated that Multi View Learning is inherently more powerful than other approaches of relational learning. There clearly is a large class of Data Mining problems that cannot be successfully approached using another relational learning without transformation. These problems, which can be characterized by the presence of relational structure within the database they deal with, can successfully be approached by the Multi View Learning.

ACKNOWLEDGMENT

We are thankful to the Omnipotent GOD for making us able to do something. We express our gratitude to the management of CHARUSAT; for providing us research opportunities and their wholehearted support for such activities.

Finally, our acknowledgement cannot end without thanking to the authors whose research papers helped us in making this research.

REFERENCES

1. Dr. M. Thangaraj, A Study on Classification Approaches across Multiple Database Relations, International Journal of Computer Applications (0975 – 8887) Volume 12– No.12, January 2011.
2. Jing-Feng Guo, An Efficient Relational Decision Tree Classification Algorithm, Third International Conference on Natural Computation (ICNC 2007).
3. Yin, X., Han, J., Yang, J., and Yu, P.S., CrossMine: Efficient Classification across Multiple Database Relations, in Proceedings of the 2004 International conference on Data Engineering (ICDE'04), Boston, MA, 2004.
4. Hongyan Liu, Xiaoxin Yin, and Jiawei Han, “d” , MRDM-2005, Chicago, 2005
5. Arno Jan Knobbe, A Ph.D Theis on Multi relational Data Mining SIKS Dissertation Series No. 2004-15.
6. Amir Netz, Integration of Data Mining and Relational Databases, Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000
7. Andreas Heß and Nick Kushmerick, Iterative Ensemble Classification for Relational Data: A Case Study of Semantic Web Services, 2007
8. Anneleen Van Assche, Improving The Applicability Of Ensemble Methods In Data Mining, PhD Thesis, ISBN 978–90–5682–896–7, Katholieke University Leuven – B-3001 Heverlee (Belgium) 2008.
9. Guo, H., Herna, L., Viktor.. Multirelational classification: a multiple view approach, Knowl. Inf. Systems, vol.17, pp.287–312, Springer-Verlag London. 2008
10. PAN Cao, WANG Hong-yuan,,Multi-relational classification on the basis of the attribute reduction twice, Journal of Communication and Computer, ISSN 1548-7709, USA, Nov. 2009, Volume 6, No.11 (Serial No.60)
11. Christophe Giraud-, Relationships among Learning Algorithms and Tasks, A dissertation submitted to the faculty of Brigham Young University in partial fulfillment of the requirements for the degree of Doctor of Philosophy Brigham Young University, 2011
12. Christine Preisach and Lars Schmidt-Thieme, Relational ensemble classification. In ICDM Conference on Data Mining, pages 499–509,Washington, DC, USA, 2006. IEEE Computer Society..
13. Hongyu Guo, Member, IEEE, and Herna L. Viktor, Member, IEEE, Multi-view ANNs for Multi-relational Classification, 2006 International Joint Conference on Neural Networks Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada,July 16-21, 2006
14. Hongyu Guo and Herna L. Viktor, Mining Relational Data through Correlation based Multiple View Validation, KDD’06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.Copyright 2006 ACM
15. Yun Li, Luan Luan, Multi-relational Classification Based on the Contribution of Tables, International Conference on Artificial Intelligence and Computational Intelligence,2009.
16. Zhen Peng and Lifeng Wu, Research on Multi-Relational Classification Approaches, International Conference on Computational Intelligence and Natural Computing, 2009
17. Miao Zou, Tengjiao Wang,A General Multi-relational Classification Approach Using FeatureGeneration & Selection, Advanced Data Mining & Applications, Lecture Notes in Computer Science,2010,Vol 6441/2010,21-33.
18. Jun He,Hongyan Liu, SELECTING EFFECTIVE FEATURES AND RELATIONS FOR EFFICIENT MULTI-RELATIONAL CLASSIFICATION, International Journal of Computational Intelligence, Volume 26,Issue 3,pages 258-281,2010, Wiley Periodicals, Inc
19. Geeta Manjunath,M Narasimha,Dinkar Sitaram, A heterogeneous Naive Bayesian classifier for relational databases, International Conference on Pattern Recognition, pp 3316-2219,2010.
20. Guo, H., Herna, L., Viktor, Learning from Skewed Class Multi-relational Databases, Fundamenta Informaticae - Progress on Multi-Relational Data Mining Volume 89 Issue 1, January 2009
21. Werner Uwents Neural networks for relational learning: an experimental comparison, Journal Machine Learning Volume 82 Issue 3, March 2011.
22. Sarah Daniel Abdelmessih, Classifiers' Accuracy Prediction based on Data Characterization, Multimedia Analysis and Data Mining Competence Center German Research Center for Artificial Intelligence (DFKI GmbH) Kaiserslautern, Germany, August, 2010.

AUTHORS PROFILE

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 Technology & Engineering, at 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 15 research papers in the field of data mining and web technology. He is Institutional member of CSI-India and he has also worked as a reviewer for National Journal and Conferences .His current research interest includes Multi relational Data Mining, Relational Classification and learning Using Multiple Views.