ISSN No. 0976-5697

Volume 1, No. 4, Nov-Dec 2010

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Performance based Multi-relational Data Mining Technique

Santosh Shakya*
Bhopal Institute of Technology & Science
Bhopal, M.P., India
santosh.shakya3@gmail.com

Gopal Patidar
Jai Narain College of Technology
M.P., India
jnct_gopal@yahoo.in

Abstract: Multi-relational learning has become popular due to the limitations of propositional problem definition in structured domains and the tendency of storing data in relational databases. As patterns involve multiple relations, the search space of possible hypotheses becomes intractably complex. Many relational knowledge discovery systems have been developed employing various search strategies, search heuristics and pattern language limitations in order to cope with the complexity of hypothesis space. In this work, we propose a relational concept learning technique, which adopts concept descriptions as associations between the concept and the preconditions to this concept and employs a relational upgrade of association rule mining search heuristic, APRIORI rule, to effectively prune the search space? The proposed system is a hybrid predictive inductive logic system, which utilizes inverse resolution for generalization of concept instances in the presence of background knowledge and refines these general patterns into frequent and strong concept definitions with a modified APRIORI-based specialization operator. Two versions of the system are tested for three real-world learning problems: learning a linearly recursive relation, predicting carcinogenicity of molecules within Predictive Toxicology Evaluation (PTE) challenge and mesh design. Results of the experiments show that the proposed hybrid method is competitive with state-of-the-art systems.

Keywords: Multi-Relational Learning, ILP, Association Rule-Mining, APRIORI.

I. INTRODUCTION

Multi-Relational Learning- Initial knowledge acquisition systems have been developed to learn from propositional representation of problem domains. In propositional (attribute value) learning, every target instance and the background knowledge related to that instance is represented by a single record in a table. This type of representation is infeasible to specify the relations between the subparts of the instance and one-to-many relations between the instance and its subparts.

The inadequacy in representation results in incomplete learned concept descriptions.

Due to the impracticality of single-table data representation, multi-relational databases have become widespread in all computer-based processes. This has led to the need for multi-relational learning systems that directly apply to relational representations of structured problem domains. There are three key approaches in constructing relational learning systems:

A. The system is composed of three parts: pre-processing, hypothesis construction and post-processing. In the preprocessing phase, the problem definition in relational form is transformed into propositional one. Then, one of the attribute-value learning systems, suitable for the data mining task, is applied. Finally, the induced if-then rules are transformed in relational form. One of the ILP systems using this approach is the LINUS framework that utilizes an embedded deductive hierarchical database (DHDB) interface in data transformation and one of three propositional learning systems among ASSISTANT, NEWGEM and CN^2 according to the problem domain in induction phase. Due to the limitations of attribute-value representation mentioned, information loss is possible in transformation and propositional patterns are not as easily understandable as relational ones in a structured problem domain. Therefore, this method is not preferable.

B. Attribute-value learning systems have been upgraded to the multi-relational counterparts in every branch of data-mining.

C. New concept description systems have been introduced, in order to fulfill the task of defining unknown relations with the help of known background knowledge as logical programs. Most relational upgrades of data mining systems and concept learning systems employ first-order predicate logic as representation language for background knowledge and data structures/patterns. The learning systems, which induce logical patterns or programs valid for given background knowledge, have been gathered under a research area, called Inductive Logic Programming (ILP), a subfield of Machine Learning and Logic Programming [1] and [2]. The propositional data structures used in data mining area (decision trees, if-then classification rules and association rules) have been extended to relational form in multi-relational data mining (MRDM) systems.

Concept learning aims at developing search techniques that efficiently traverse target concept description space consisting of logical Horn clauses. There are various approaches designed to solve this problem:

D. Top-down approach using information gain as search heuristics

E. Top-down approach utilizing higher-order rule schemas to constrain search

F. Bottom-up approach constraining search by generalizing from concept instances using inverse resolution operators

G. Bottom-up approach avoiding search using relative least general generalization (RLGG) operator.

The first relational learning algorithm to use information gain based search heuristics was FOIL. It uses an AQ-like covering approach and it inherits the top-down search strategy from MIS, which is an early concept

learning system. Recently, many systems that extend FOIL in various aspects have been introduced.

The search heuristics, information gain and higher-order rule schemas, have no proof-theoretic basis; therefore the search space of possible concept descriptions is not complete. The resolution rule that forms the basis of the logic programming paradigm is a sound and complete inference rule. Inverting this inference rule results in induction of refutation trees in a bottom-up fashion and systems employing inverse resolution operators have a proof-theoretic search strategy.

MARVIN is the first ILP system inducing Horn clauses using an inverse resolution generalization operator. The hypothesis language of the system does not contain clauses with existential quantified variables and the system can not introduce new predicates. There is no search heuristics to direct the search; instead the oracle evaluates the quality of induced clauses.

II. BACKGROUND

A. Inductive Logic Programming- When human brain reasons about events, it tries to prove and deduce the result with the help of assembled background knowledge about the event domain. So, how is background knowledge acquired and collected in human brain? Apart from the knowledge obtained from ancestors, human being collects some particular patterns recurring in different situations for similar future events. This ability of generalization from specific observations, called induction, influenced the development of Inductive Logic Programming (ILP) as a branch of Artificial Intelligence.

ILP basically studies learning concept definitions or regularities from specific instances in terms of prior known relations in clausal logic framework. Generally, ILP learner is presented a set of training examples and background knowledge in form of logic clauses, and induces concepts or frequent patterns as logical expressions. The term hypothesis is also used for induced concept/pattern description. Inductive learning is in fact searching for complete and consistent concept descriptions in the space limited by description language of the ILP system. The current state of art in ILP is achieving to find qualified logical hypothesis efficiently, i.e. in minimal learning time. Current learning systems employ constraints on the search space via language, search strategy or user feedback in the sake of efficiency.

Predictive Inductive Learning- In predictive ILP, the task is learning concept/class descriptions, that correctly classify instances (and non-instances) of a specific concept, in terms of the background knowledge about the problem domain. Predictive learning can be applied to any classification or prediction problem, such as predicting carcinogenic activity of chemical compounds based on their chemical structures. In this problem, the concept instance space is chemical compounds, the concept is whether a compound is carcinogenic or not and the task is finding correct classification rules that map positive instances to carcinogenic class and negative ones to non-carcinogenic class. The problem setting of the predictive ILP learning task introduced as follows:

Given:

[a] Target class/concept C,
[b] A set E of positive and negative example of the class/concept C,
[c] A finite set of background facts/clauses B,
[d] Concept description language L (language bias).

Find:

[e] A finite set of clauses H, expressed in concept description language L, such that H together with the background knowledge B entail all positive instances E+ and none of the negative instances E-. In other words, H is complete and consistent with respect to B and E, respectively.

In this problem setting, completeness and consistency are the quality criteria for selecting the induced hypotheses; however the definitions of these terms require the hypotheses %100 fit the given instances, which is too strict for hypothesis to have predictive power. There may be errors in the background knowledge and training concept instances; or training examples can be sparse to reflect the general regularities hidden in the concept. Since success of a predictive learning system lies in the ability to generalize for unseen concept instances correctly, predictive ILP systems should employ more relaxed quality criterion that allow some training examples remain misclassified [2].

B. Descriptive Inductive Learning- Descriptive data mining differs from the predictive data mining such that the search is not directed by a target concept. A descriptive ILP system does not know which class or concept it is looking for in underlying database; instead it searches for interesting frequent patterns with no single target attribute, i.e. the consequent of the rules can be any attribute or relation in the data. In other words, the data mining system explores relationships between the tendency of domain subjects in doing an action/having a property (buying a specific product/ having cancer genetic effect) and domain-related features of the subjects (being female/having a specific molecular structure). In descriptive data mining, the main objective is to find useful/interesting and understandable patterns. Therefore, the pattern representation language and the interestingness criterion play the main role in the success of a descriptive data mining system [2] and [3].

C. Relational Association Rule Mining- Association rule mining aims at discovering hidden structures, also called patterns, in data. In Boolean association rule discovery, there is one object type and one database table describing different features of this object type. The patterns mined are feature sets that are common for number of objects exceeding a frequency threshold. For instance, in the market-basket problem, the objects are baskets, each item is one feature of the basket and the patterns are the frequent item sets common in baskets. In relational association rule mining, there are more than one object types and the patterns are not only feature sets but also they consist of relations between objects. Relational association rule mining can be described as discovering recurrent relational patterns in a relational database [2].

APRIORI-APRIORI utilizes an important property of frequent item sets in order to prune candidate item set space: All subsets of a frequent item set must be large. The contra positive of this property says that if an item set is not frequent then any superset of this set is also not frequent. It can be included that the item set space should be traversed

from small size item sets to large ones in order to discard any superset of infrequent item sets from scratch. In order to apply this reasoning, APRIORI reorganizes the item set space as a lattice based on the subset relation, as shown in Figure. The item set lattice in Figure is composed of possible large item sets for items I_1, I_2, I_3. The directed lines in the lattice represent the subset relationships, and the frequent item set property says that any set in a path below an item set is infrequent if the original item set is infrequent. For instance, if the item I_1 is not found frequently in transaction baskets, then the item sets { I_1, I_2}, { I_1, I_3} and { I_1, I_2, I_3} are not frequent, either.

In APRIORI, an item set is called a candidate if all its subsets are frequent item sets. An item set is large/frequent if it is candidate and the number of occurrences of this item set in transactions is greater than the support threshold value [1], [4], and [5].

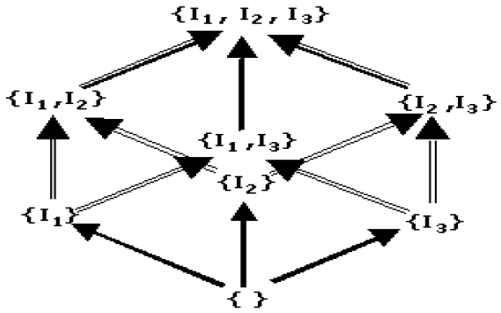

Figure: 1 The APRIORI lattice with three items

WARMR- A relational association rule miner that discovers frequent Datalog queries, WARMR, is presented. WARMR takes a Datalog relational database and a support threshold as input and outputs Datalog queries that are frequent in the input database. Since first-order predicate language allows the use of variables and multiple relations in patterns, the patterns are more expressive than the propositional ones; besides, the size of the pattern space is huge.

A relational association rule miner should determine a formalism to syntactically constrain the query language to a set of meaningful queries. For instance, the formalism/declarative bias should exclude queries that bind incompatible argument types, like unifying a person and a product type variables in "sibling(X, Y), buys(john, Y)" (Y is a person in "sibling" predicate and is a product in "buys" predicate).

In the language formalism of WARMR, WRMODE, a set of all possible ground and non-ground atoms is explicitly presented to the system. Each variable argument of each atom in the set is marked by means of three mode-labels +, - and ±; where + means that the variable is strictly input/bound, i.e. has to appear earlier in the query; - means that the variable is strictly output/unbound, i.e. must not appear earlier; ± means that the variable can be both input and/or output. Input-output modes of the variables in the formalism constrain the refinement of queries in a way that the modes determine which atoms can be added to a query.

The main advantage of the WARMR system is its flexibility offered to the user in determining the search space of possible patterns and adding background knowledge to the database. These settings are fully isolated from the implementation. However, the mode declaration in the formalism is too hard for a normal user to state which patterns he/she really wants to discover and it is not practical for large databases. It should not be overlooked that the user in descriptive data mining does not know what he/she wants to find and even does not have deep knowledge about relations in database [8]and [9].

III. PROPOSED TECHNIQUE

We propose in this thesis a concept learning ILP technique, which employs relational association rule mining techniques. The technique proposed utilizes inverse resolution for generalization of concept instances in the presence of background knowledge and refines these general patterns into frequent and strong concept definitions with a relational upgrade of the APRIORI refinement operator.

Proposed System I- The system proposed employs a coverage algorithm in constructing concept definition. It first selects a positive concept example, based on the order of concept instances in the database. The most general clauses, with two predicates, that entail the positive example are generated and then the concept rule space is searched with an APRIORI-like specialization operator. The specialization operator utilizes the frequency property of definite clauses in order to effectively prune the search space. Among the frequent and strong rules produced, the system selects the best clause using a criterion called f-measure, which is discussed later, and repeats the rule search for the remaining concept instances that are not in the coverage of the hypothesis clauses. The proposed system will be explained in three basic sections: generalization, refinement and evaluation.

```
For each pair of the clauses k and m in the previous tree level l-1,
do the following:
a.If clausek and clausem have same group number, continue.
b.If clausek and clausem are both recursive, continue.
c.Compute the union clauses of clausek and clausem.
d.For each possible union,
i.If tree(level) does not contain the union and the
frequency of the union is above the support threshold, then
1.If the union is a fully connected clause,
add it to the level l; otherwise discard it.
2.Generate clauses by unifying existential variables in
the body clause.
3.For each clause generated, check whether it is frequent
and connected. If it is qualified, add it to the level l.
ii.Else continue.
```

Figure: 2 The pseudo code of the refinement operator in Proposed System I.

Proposed System II- In order to capture clauses that have relations not directly bound to the head predicate, the proposed system allows fully existential/unbound predicates in the body of clauses in the generalization step. Therefore, the first level of the search lattice expands exponentially as the number of facts related to the current concept instance increases. Since the size of each level l of the ARPIORI search lattice is order of two squared the size of the level l-1, the size of the search lattice is order of n^k (where $k= 2^{(d-1)}$)in the worst case, where n is the size of the first level and d is the depth of the search tree. For large scale data mining tasks like discovering structure-activity relationships (SAR) that relate molecular structure with specific ability of

molecules, the background knowledge database is generally composed of 20000 records or more, which results in an intractable problem for the proposed system I. There is a tradeoff between the complexity and the completeness of the algorithm. In our less complex and then less complete solution, the system tightens the limits of the language bias in the sake of efficiency. The proposed system II does not allow clauses with body relations not directly bound to the head predicate in the language. We change the generalization and refinement operators of the proposed system I in the second version. The concept learning time is determined by the number of two predicate clauses in the first level of the search space. Therefore, the generalization operator of the proposed system II differentiates in the way that it does not allow two predicate clauses that have body predicates not bound to the head. Besides the limitation on the body predicates, the generalization operator selectively substitutes one-location existential variables (variable that is not bound in the head and exists only once) to terms in the body predicates.

In the Proposed System I, an extra specialization step is employed, via unifying the existential variables, to capture inner structures not directly related to the head predicate, in other words n-depth relational patterns. In this implementation, the system does not allow fully existential body predicates in the generalization step.

- Initialize the set of concept instances set I
- Initialize the hypothesis H = □
- Do until all the concept instances are covered by the hypothesis (I = □):
1.Select the first positive concept instance p from I.
2.Generalize positive instance in presence of background knowledge and call the set of generalizations G.
3.Initialize level d := 1
4.Initialize the set of candidate clauses C1 := G
5.Initialize the set of frequent queries F := { }
6.While Qd not empty and d ≤ maxdepth
a. Find frequency of all clauses C □ Cd
b. Discard the clauses with frequency below minfreq from Cd.
c. Update F := F □ Cd
d. Compute new candidates Qd+1 from d+1,e. Increment d by 1.
7.Discard the clauses with confidence below minconf from F.
8. Select the best clause cbest from F using the f-measure criterion.
9.Compute the set of concept instances Ic covered by the best clause.
10.Update H := H □ cbest
11.Update I := I – Ibc
-Return H.

Figure: 3 The Proposed System I

Therefore, an extra generalization step is employed after combining the clauses in order to discover the first-order features, which are sets of body literals interacting by local variables. In the generalization step, new local variables are introduced by unifying common constants of the same type in various argument positions of body predicates. The generalization after specialization constitutes a breach in the application of the APRIORI rule since any generalization of the specialization of two clauses can be more general than one of or both of the clauses. Therefore, this prevents the top-down specialization of the APRIORI lattice. As a result, it is possible to bypass some frequent definite clauses because of this gap.

Finally, the filtering step, which checks whether the candidate clause is connected or not, is not employed in this version since the clauses are guaranteed to be connected via head variables.

There is a trade-off between two versions of the proposed technique. The proposed system I allows the clauses including body predicates that are not directly connected to the head predicate in the search space, whereas the second one does not in the sake of efficiency. The efficiency of the second system lies in not allowing fully existential body predicates in the generalization step. It is only possible by adding literals to the body of the clauses in inner steps and bounding these literals to only body predicates. This extra specialization step also results in performance overhead as introducing existential variables in the first level.

IV. RESULTS

We trained the system with varying values of the support threshold while the confidence threshold and the maximum number of predicates are fixed to 0.6 and 5, respectively. The support values and the corresponding predictive accuracies of the resultant hypotheses are plotted.

A very low or very high support threshold results in low accuracy. If the minimum support value is too high (> 0.15), then the rules involving patterns that rarely occur are not generated. Besides, rules that partition the concept instances into many small subsets may be generated, many of which are not correct if it is set too low (< 0.1). Therefore, the minimum support value in the range of [0.1, 0.15] should be preferred. After selecting the optimum value 0.15 for the support value, we obtained and tested theories for different confidence values in the range [0.3, 0.8]. Normally, the rules having 100% confidence value should appear in the final hypothesis; however, there can be noise in the data that should be tolerated. There is no regular behavior of the accuracy in confidence threshold values below 0.5 and this tells us that confidence below a minimum value is not a criterion in discriminating performance of the rules.

However, accuracy steadily decreases as the confidence increases from the value of 0.6. This can be explained by the decrease in the coverage of the hypothesis. As can be seen from the graph, the confidence threshold value 0.6 results in the best predictive accuracy. To set the minimum confidence to 0.6 means that the system tolerates 40% noise at maximum. However, 40% noise for the carcinogenesis data is an aggregated value since the data is obtained from the long-lasting bioassays. We can conclude that not the noise but the missing data about molecular structures and properties in the background knowledge pulls down the minimum confidence value.

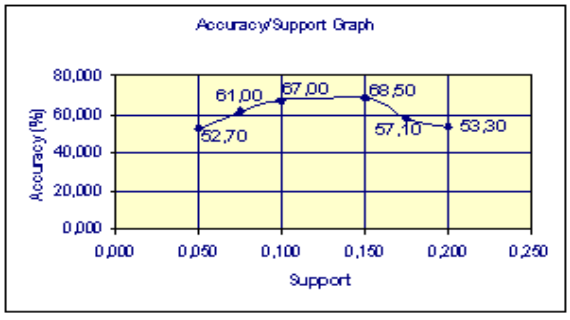

Figure: 4 Predictive Accuracy/Support Graph

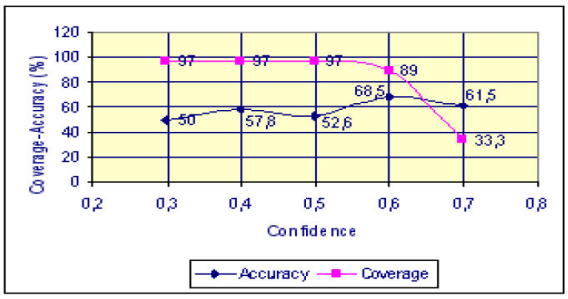

Figure: 5 Predictive Accuracy/Confidence Graph

V. CONCLUSION

In this thesis, many aspects of multi-relational data mining are examined and discussed. The aim is to combine rule extraction methods in ILP and efficient search strategies of data mining. As an outcome, two versions of a concept learning tool, a modified combination of WARMR and Inverse Resolution absorption operator, is produced. We come up with promising test results that are comparable with the performance of current state of-the-art knowledge discovery systems, such as PROGOL.

Additionally, this thesis introduces a new method that induces modes of predicate arguments, referenced or non-referenced, via inputting basic domain knowledge from the user. The non-referenced arguments are in fact one-location existential variables; an n-argument predicate results in 2^n different combinations of instantiations if non-referenced variables are allowed for all argument positions of the predicate, resulting in exponential grow of the search space. The proposed mode induction algorithm speeds up the learning process, by determining which predicate arguments must be referenced in the clause or which ones can be ignored, not referenced. However, the methodology requires normalized data set in which utility and structural predicates are isolated.

VI. REFERENCES

[1] Yingqin Gu, Hongyan Liu, Jun He, Bo Hu and Xiaoyong Du1, "A Multi-relational Classification Algorithm based on Association Rules", IEEE 2009 Conference on Web Information Systems and Mining.

[2] Alexessander Alves, Rui Camacho and Eugenio Oliveira, "Discovery of Functional Relationships in Multi-relational Data using Inductive Logic Programming" , Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04).

[3] Harish Sethua and Alexander Yates, "Using Text Analysis to Understand the Structure and Dynamics of the World Wide Web as a Multi-Relational Graph", IEEE 2010.

[4] Wei Zhang, "Mining Multi-Level Multi-Relational Frequent Patterns Based on Conjunctive Query Containment", IEEE-2010.

[5] Amanda Clare, Hugh E. Williams and Nicholas Lester, "Scalable Multi-Relational Association Mining", Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04).

[6] Alexessander Alves, Rui Camacho and Eugenio Oliveira, "Discovery of Functional Relationships in Multi-relational Data using Inductive Logic Programming", Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04).

[7] Bao Liang,Xiaoguang Hong,Lei Zhang,Shuai Li, "Extended MRI-Cube Algorithm for Mining Multi-Relational Patterns", IEEE 2008 The 9th International Conference for Young Computer Scientists.

[8] L. P. Castillo and S. Wrobel. Macro-operators in multirelational learning: a search-space reduction technique. In T. Elomaa, H. Mannila, and H. Toivonen, editors, Proceedings of the 13th European Conference on Machine Learning, volume 2430 of Lecture Notes in Artificial Intelligence, pages 357–368. Springer-Verlag, August 2002.

[9] L. De Raedt, H. Blockeel, L. Dehaspe, andW. Van Laer. Three companions for data mining in first order logic. In S. Dˇzeroski and N. Lavraˇc, editors, Relational Data Mining, pages 105–139. Springer-Verlag, September 2001.

[10] M. H. Dunham. Data Mining: Introductory and Advanced Topics. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2002.