ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

A Roadmap to an Enhanced Graph Based Data mining Approach for Multi-Relational Data mining

D.Kavinya[1]

Student, Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India [1]

ABSTRACT: Multi-relational data mining (MRDM) is a subfield of data mining which focuses on knowledge discovery from relational databases comprising multiple tables. The approaches to MRDM are categorized into five groups. They are greedy search based approach, inductive logic programming (ILP) based approach, inductive database based approach, mathematical graph theory based approach and kernel function based approach. Representation is a fundamental as well as a critical aspect in the process of discovery. These five approaches are grouped into two forms of representation, namely the graph-based representation (GBR) and the logic-based representation (LBR). In this paper, some major studies in each category are described, and representative methods among the studies are sketched. The comparison of Graph-based and logic-based multi-relational data mining according to the factors such as Structural Complexity, Semantic Complexity, Background Knowledge intended to condense the hypothesis space, and Background Knowledge intended to augment the hypothesis space is described and based on the comparison results the necessity of enhancing the Graph-based representation is enforced.

KEYWORDS: Structural Complexity, Semantic Complexity, Background Knowledge

I. INTRODUCTION

During the past decade, the field of data mining has emerged as a novel field of research, investigating interesting research issues and developing challenging real-life applications. The objective data formats in the beginning of the field were limited to relational tables and transactions where each instance is represented by one row in a table or one transaction represented as a set. However, the studies within the last several years began to extend the classes of considered data to semi-structured data such as HTML and XML texts, symbolic sequences, ordered trees and relations represented by advanced logics. One of the most recent research topics associated with structured data is multi-relational data mining whose main scope is to find patterns in expressive logical and relational languages from complex, multi-relational and structured data. The main aim of mining semi-structured data, symbolic sequences and ordered trees is to extract patterns from structured data.

Representation is a fundamental as well as a critical aspect in the process of discovery and two forms of representation, namely the graph-based representation and the logic-based representation, have been used for MRDM. Logic-based MRDM popularly known as Inductive Logic Programming (ILP), is the intersection of Machine Learning and Logic Programming. ILP is characterized by the use of logic for the representation of multi relational data. ILP systems represent examples, background knowledge, hypotheses and target concepts in Horn clause logic. Graph-based approaches are characterized by representation of multi- relational data in the form of graphs. Graph-based MRDM systems have been extensively applied to the task of unsupervised learning, popularly known as frequent subgraph mining and to a certain extent to supervised learning. Graph-based approaches represent examples, background Knowledge, hypotheses and target concepts as graphs. These approaches include mathematical graph theory based approaches like FSG and gSpan, greedy search based approaches like Subdue or GBI, and kernel function based approaches. The core of all these approaches is the use of a graph-based representation and the search for graph patterns which are frequent or which compress the input graphs or which distinguish positive and negative examples.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

II. APPROACHES OF GRAPH BASED MINING

The approaches to graph-based data mining are categorized into four groups. They are

- Greedy search based approach
- Inductive Logic Programming (ILP) based approach,
- Inductive Database based approach,
- Mathematical graph theory based approach
- Kernel function based approach

A. GREEDY SEARCH BASED APPROACH

Which has been used in the initial works in graph-based data mining[1]. This type belongs to heuristic search and direct matching. The greedy search is further categorized into depth-first search (DFS) and breadth-first search (BFS). DFS was used in the early studies since it can save memory consumption. Initially the mapping f_{s1} from a vertex in a candidate subgraph to a vertex in the graphs of a given data set is searched under a mining measure. Then another adjacent vertex is added to the vertex mapped by $\boldsymbol{f_{s1}}$, and the extended mapping $\boldsymbol{f_{s2}}$ to map these two vertices to two vertices in the graphs of a given data set is searched under the mining measure. This process is repeated until no more extension of the mapping $\boldsymbol{f_{sn}}$ is available where $\boldsymbol{n}$ is the maximal depth of the search of a DFS branch. A drawback of *this DFS* approach is that only an arbitrary part of the isomorphic subgraphs can be found when the search must be stopped due to the search time constraints if the search space is very large. Because of the recent progress of the computer hardware, more memory became available in the search. Accordingly, the recent approaches to graph-based data mining are using BFS. *An advantage of BFS* is that it can ensure derivation of all isomorphic subgraphs within a specified size of the subgraphs even under greedy search scheme. However, the search space is so large in many applications that it often does not fit in memory. To alleviate this difficulty, *beam search method* is used in the recent greedy search based approach where the maximum number of the BFS branches is set, and the search proceeds downward by pruning the branches which do not fit the maximum branch number. Since this method prunes the search paths, the search of the isomorphic subgraphs finishes within tractable time while the completeness of the search is lost.

Advantages

- It can perform approximate matching to allow slight variations of subgraphs
- It can embed background knowledge in the form of predefined subgraphs
- Facilitate the global understanding of the complex database by forming hierarchical concepts and using them to approximately describe the input data.

Disadvantages

- The search is completely greedy, and it never backtracks.
- Since the maximum width of the beam is predetermined, it may miss an optimum Gs.

B. INDUCTIVE LOGIC PROGRAMMING (ILP)

The "induction" is known to be the combination of the "abduction" to select some hypotheses and the "justification" to seek the hypotheses to justify the observed facts. Its main advantage is the abilities to introduce background knowledge associated with the subgraph isomorphism and the objective of the graph-based data mining. It can also derive knowledge represented by"First order predicate logic" from a given set of data under the background knowledge. The general graph is known to be represented by "First orders predicate logic". The first order predicate logic is so generic that generalized patterns of graph structures to include variables on the labels of vertices and edges are represented[2]. ILP is formalized as follows:

Given the background knowledge B and the evidence (the observed data) E where E consists of the positive evidence E^+ and the negative evidence E^-, ILP finds a hypothesis H such that the following normal semantics conditions hold.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

1. Posterior Satisability: $B \wedge H \wedge E^{-} \not\models \square$
2. Posterior Sufficiency: $B \wedge H \models E+$

where □ is false, and hence $\not\models \square$ means that the theory is satisfiable. In case of ILP, intentional definitions are derived from the given data represented by instantiated first order predicates, i.e., extensional definitions. For ILP, the advantage is not limited to the knowledge to be discovered but the ability to use the positive and the negative examples in the induction of the knowledge. A disadvantage is the size of the search space which is very huge in general and computational intractability. The ILP method can be any of heuristic, complete, direct and indirect search according to the background knowledge used to control the search process. When control knowledge is used to prune some search paths having low possibility to find isomorphic subgraphs under a given mining measure, the method is heuristic. Otherwise, it is complete. When some knowledge on predetermined subgraph patterns are introduced to match subgraph structures, the method is indirect since only the subgraph patterns including the predetermined patterns or being similar to the predetermined patterns are mined. In this case the subgraph isomorphism is not strictly solved.

Advantages

- Class of structures which can be searched is more general than graphs
- Discover frequent structures in high level descriptions. These approaches are expected to address many problems,

because many context dependent data in the real world can be represented as a set of grounded first order predicates which is represented by graphs

Disadvantages

- Easily face the high computational complexity

C. INDUCTIVE DATABASE BASED APPROACH

This method performs the complete search of the paths embedded in a graph data set where the paths satisfy monotonic and anti-monotonic measures in the version space. The version space is a search subspace in a lattice structure. Given a data set, a mining approach such as inductive decision tree learning, basket analysis and ILP is applied to the data to pregenerate inductive rules, relations or patterns. The induced results are stored in a database. The database is queried by using a query language designed to concisely express query conditions on the forms of the pregenerated results in the database. This framework is applicable to graph-based mining. Subgraphs and/or relations among subgraphs are pregenerated by using a graph-based mining approach, and stored in an inductive database. A query on the subgraphs and/or the relations is made by using a query language dedicated to the database.

Advantage

- The operation of the graph mining is fast as the basic patterns of the subgraphs and/or the relations have already been pregenerated.

Disadvantage

- Large amount of computation and memory is required to pregenerate and store the induced patterns.

D. MATHEMATICAL GRAPH THEORY BASED METHOD

These are complete search and direct methods. In case of Apriori algorithm which is the most representative for the basket analysis [3], all frequent items which appear more than a specified minimum support "minsup" in the transaction data are enumerated as frequent itemsets of size 1. This task is easily conducted by scanning the given transaction data once. Subsequently, the frequent itemsets are joined into the candidate frequent itemsets of size 2, and their support values are checked in the data. Only the candidates having the support higher than minsup are retained as the frequent itemsets of size 2. This process to extend the search level in terms of the size of the frequent itemsets is repeated until no more frequent itemsets are found. This search is complete since the algorithm exhaustively searches the complete set of frequent item sets in a level-wise manner. In case of the graph-based data mining, the data are not the transactions, i.e., sets of items, but graphs, i.e., combinations of a vertex set V (G) and an edge set E(G) which include topological information. Accordingly, the above level-wise search is extended to handle the connections of vertices and edges.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Similarly to the Apriori algorithm, the search in a given graph data starts from the frequent graphs of size 1 where each consists of only a single vertex. Subsequently, the candidate frequent graphs of size 2 are enumerated by combining two frequent

Vertices. Then the support of each candidate is counted in the graph data, and only the graphs having higher support than the minsup are retained. In this counting stage, the edge information is used. If the existence and the label of the edge between the two vertices do not match, the graph of size 2 is not counted as an identical graph. This process is further repeated to incrementally extend the size of the frequent graphs in a level wise manner, and finishes when the frequent graphs are exhaustively searched. In inductive database and constraint-based mining, the algorithm can be extended to introduce monotonic measures such as "maxsup" [4].

Advantage

- This technique significantly increases the matching efficiency
- can derive complete set of frequent subgraphs over a given minsup in a very efficient manner in both computational time and memory consumption

Disadvantage

- Consumes much memory space to store massive graph data.

E. KERNEL FUNCTION BASED APPROACH

This is a heuristic search and indirect method in terms of the subgraph isomorphism problem and used in the graph classification problem. It is not dedicated to graph data but to feature vector data[5]

A kernel function K defines a similarity between two graphs G_x and G_y. For the application to graph-based data mining, the key issue is to find the good combinations of the feature vector XG and the mapping Ø: XG → H to define appropriate similarity under abstracted inner product <Ø(XGx), Ø (XGy) >. A recent study proposed a composition of a kernel function characterizing the similarity between two graphs G_x and G_y based on the feature vectors consisting of graph invariants of vertex labels and edge labels in the certain neighbor area of each vertex [6]. This is used to classify the graphs into binary classes.

Advantage

- Though the similarity is not complete and sound in terms of the graph isomorphism, the graphs are classified properly based on the similarity defined by the kernel function.
- Can provide an efficient classifier based on the set of graph invariants.

Some experiments report that the similarity evaluation in the structure characterizing the relations among the instances provides better performance in classification and clustering tasks than the distance based similarity evaluation [7].

III. COMPARISONS OF GBR AND LBR

When graph-based and logic-based multi-relational data mining have been compared according to, the ability to discover structurally large concepts, the ability to discover semantically complicated concepts (or the ability to utilize background knowledge which augments the hypothesis space), and the ability to effectively utilize background knowledge (which condenses the hypothesis space), it has been found that

- GBR performs significantly better than LBR in the first case, i.e. while learning structurally large concepts.
- LBR outperformed GBR in the second case, i.e. while learning semantically complicated concepts.
- In the third case i.e. while utilizing background knowledge the performance of the systems is found to be comparable.
- GBR has much less expressiveness than that LBR which can accept and utilize the background knowledge.
- The less expressive representation used by GBR leads to an efficient exploration of the hypothesis space.
- LBR which uses a more expressive representation not only performs more efficiently but also can learn concepts which may not be expressed by GBR mechanism of explicit instantiation.

IV. CONCLUSION AND SUGGESSION

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

In this paper, some major studies in the various approaches to MRDM are discussed in detail and their features and drawbacks highlighted. Finally, the comparison of Graph-based and logic-based multi-relational data mining according to the factors such as Structural Complexity, Semantic Complexity, Background Knowledge intended to condense the hypothesis space, and Background Knowledge intended to augment the hypothesis space is discussed. The use of a less expressive representation (GBR) will facilitate an efficient search and lead to a superior performance while learning structurally large concepts. The use of a weaker representation (LBR) will limit the learning of semantically complicated concepts but encourages the effective use of background knowledge. It would be possible to introduce the syntax and semantics in graph-based representations to express semantically complicated concepts using ordered graphs, hyper-graphs or graph rewriting rules. In this case, a graph-based system will tend to outperform a logic-based system.

Based on the comparison results the necessity of enhancing the Graph-based representation is enforced and hence the scope of this paper is to develop an enhanced Graph- based method to mine the multi-relational databases which has a desirable expressive representation and the ability to utilize the background knowledge effectively.

REFERENCES

1. J. Cook and L. Holder. Substructure discovery using minimum description length and background knowledge. J. Artificial Intel. Research, 1:231-255, 1994.

2. S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. J. Logic Programming, 19(20):629-679, 1994.

3. A.Inokuchi, T.Washio, and H. Motoda. Complete mining of frequent patterns from graphs: Mining graph data. Machine Learning, 50:321-354, 2003.

4. L. De Raedt and S. Kramer. The level wise version space algorithm and its application to molecular fragment finding. In IJCAI'01: Seventeenth International Joint Conference on Artificial Intelligence, volume 2, pages 853-859, 2001.

5. H. Kashima and A. Inokuchi. Kernels for graph classification. In AM2002: Proc. of Int. Workshop on Active Mining, pages 31-35, 2002.

6. R. Kondor and J. Laberty. Diffusion kernels on graphs and other discrete input space. In ICML'02: Nineteenth International Joint Conference on Machine Learning, pages 315-322, 2002.

7. T. Gaertner. A survey of kernels for structured data. SIGKDD Explorations, 5(1), 2003.