Data Mining

Seema, Dharmveer Yadav, Pramod Kumar

Abstract— Graph-based data mining represents a collection of techniques for mining the relational aspects of data r epresented as a graph. Two major approaches to graph based data mining are frequent sub graph mining and graph-based relational learning. This article will focus on one particular approach embodied in the Subdue system, along with recent advances in graph-based supervised learning, graph-based hierarchical conceptual clustering, and graph-grammar induction. The need for mining structured data has increased rapidly. One of the best studied data structures in computer science and discrete mathematics are graphs. Graph based data mining has become quite popular in the last few years. This paper introduces the theoretical basis of graph based data mining and surveys the state of the art of graph-based data mining. Brief descriptions of some representative approaches are provided as well.

Index Terms— Introduction,Graph,Tree,Path,Data Models,Tools,Structured Data,Data Mining,Approaches,Study Results.

1 Introduction

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 in-stance 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. Graph mining has a strong relation with the afore mentioned Multi-relational data mining. However, the main objective of graph mining is to provide new principles and efficient algorithms to mine topological substructures embedded in graph data, while the main objective of multi-relational data mining is to provide principles to mine and/or learn the relational patterns, represented by the expressive logical languages. The former is more geometry oriented and the latter more logic and relation oriented in this paper, the theoretical basis of graph-based data mining is explained in the following section. Second the approaches to graph-based data mining are reviewed and some representative approaches are briefly described.

2 Graph Based Data Mining

2.1 Theoretical Approaches of Graph Based Data Mining:

There are five theoretical based approaches of graph-based data mining. They are sub graph categories, sub graph isomorphism, graph invariants, mining measures and solution methods. The sub graphs are categorized into various classes, and the approaches of graph-based data mining strongly depend on the targeted class. Sub graph isomorphism is the mathematical basis of substructure matching and/or counting in graph-based data mining. Graph invariants provide an important mathematical criterion to efficiently reduce the search space of the targeted graph structures in some approaches. Furthermore, the mining measures define the characteristics of the patterns to be mined similarly to conventional data mining. In this paper, the theoretical basis is explained for only undirected graphs without labels but with/without cyclic edges and parallel edges due to space limitations. But, an almost identical discussion applies to directed graphs and/or labeled graphs. Most of the search algorithms used in graph-based data mining come from artificial intelligence, but some extra search algorithms founded in mathematics are also used.

2.2 Recent Developments Carried Out On Graph Based Data Mining:

Researchers have proposed a variety of unsupervised-discovery approaches for structural data. One approach is to use a knowledge base of concepts to classify the structural data. Systems using this approach learn concepts from examples and then categorize observed data. Such systems represent examples as distinct objects and process individual objects one at a time. In contrast, Subdue stores the entire database (with embedded objects) as one graph and processes the graph as a whole. Scientific discovery systems that use domain knowledge have also been developed, but they target a single application domain. An example is Mechem, which relies on domain knowledge to discover chemistry hypotheses. In contrast, Subdue performs general-purpose, automated discovery with or without domain knowledge and hence can be applied to many structural domains. Logic-based systems have dominated relational concept learning, especially inductive logic programming (ILP) systems. However, first-order logic can also be represented as a graph and, in fact, is a subset of what graphs can represent. Therefore, learning systems using graphical representations potentially can learn richer concepts if they can handle the larger hypothesis space. FOIL, the ILP system discussed in this article, executes a top-down approach to learning relational concepts (theories) represented as an ordered sequence of function-free definite clauses. Given extensional background knowledge including relations and examples of the target concept relation, FOIL begins with the most general theory. Then it follows a set-covering approach, repeatedly adding a clause that covers some positive examples and few negative examples. Then, FOIL removes the positive examples covered by the clause and iterates the process on the reduced set of positive examples and all negative examples

until the theory covers all the positive examples. To avoid over complex clauses, FOIL ensures that a clause's description length does not exceed the description length of the examples the clause covers. In addition to the applications discussed here, as well as applications in numerous recursive and no recursive logical domains, FOIL has been applied to learning search-control rules and patterns in hypertext.

2.3 Data Models

Data miting combines techniques from statistics, databases, machine learning, and pattern recognition to extract (mine) concepts, concept interrelations from large business databases. A data mining software package nearly always includes:
(i) association rules, (ii) classification, (iii) prediction methods, (iv) clustering methods and (v) exploration methods for complex data types.
The choice of the best models should be made carefully for consistency and compatibility in tested data. DM tools to identify the best model should include the costs of making bad decisions (eg, poor grading, mistakes, etc. [1,2,3]). Unfortunately, most of the DM tools to identify the best data model uses the so-called Global Accuracy method (GA) does not include costs of baddecisions. A models without considering the cost of the errors often leads to strange results. Of course, the costs arising from poor predictions may be significantly different depending on the business area. For instance, in a promotional mailing: the cost of sending junk mail to a client that doesn't respond is far less than the lost-business cost of not sending it to a client that would haveresponded (false negative error FN) .

3 Data Mining Tools

There are a lot of available DM tools on the market. However, according to the report by Gartner. SPSS has been named a leader among eight vendors, which received the highest scores not only in the completeness of vision, but also in ability to execute. Factors which have decided to achieve the leading position by SPPS are: an approach to a client, the scope of the analytical tools, management of a analytical environment.

Table 1. Evaluation of Data Mining Tools.

Vendor	2006	2008	2010	2006-2010
SPSS	0,72	0,67	0,82	0,74
SAS	0,68	0,59	0,95	0,74
Tink Analytics	0,29	0,47	0,61	0,45
Portrait_Software	0,33	0,48	0,77	0,53
Angoss	0,00	0,42	0,63	0,35
Infor_crm	0,00	0,37	0,32	0,23
Unica	0,29	0,27	0,21	0,26
Kxen	0,33	0,00	0,77	0,37

(Values of 2006-2010) to execute = (Values of 2006+2008+2010)/3.

4 Approaches of Graph Mining

4.1 Greedy Search Based Approach

Two pioneering works appeared in around 1994, both of which were in the framework of greedy search based graph mining. Interestingly both were originated to discover concepts from graph representations of some structure, e.g. a conceptual graph similar to semantic network and a physical system such as electric circuits. One is called "SUBDUE". SUBDUE deals with conceptual graphs which belong to a class of connected graph. The vertex set V (G) is R [C where R and C are the sets of labeled vertices representing relations and concepts respectively. The edge set E(G) is U which is a set of labeled edges. Though the original SUBDUE targeted the discovery of repeatedly appearing connected subgraphs in this specie type of graph data, i.e., concept graph data, the principle can be applied to generic connected graphs. SUBDUE starts looking for a subgraph which can best com press an input graph G based on Minimum Description Length (MDL) principle. The found subgraph can be considered a concept. This algorithm is based on a computationally constrained beam search. It begins with a subgraph comprising only a single vertex in the input graph G, and grows it incrementally expanding a node in it. At each expansion it evaluates the total description length (DL), I(Gs)+I (GjGs), of the input graph G which is defined as the sum of the two:

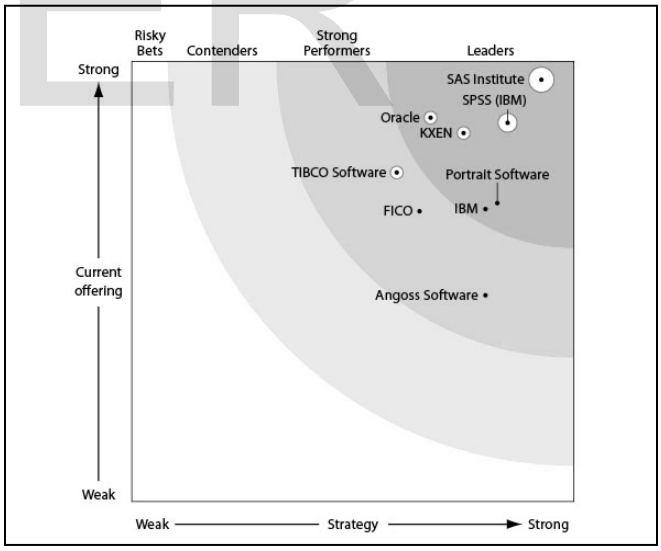

Figure: The Forrester Wave: Predictive Analytics and Data Mining Solutions.

DL of the subgraph, I(Gs), and DL of the input graph, I (GjGs), in which all the instances of the subgraph are replaced by single nodes. It stops when the subgraph that minimizes the total description length is found.

4.2 ILP Based Approach

The first system to search for the wider class of frequent by

substructure in graphs named WARMR was proposed in 1998. They combined ILP method with Apriori-like level wise search to a problem of carcinogenesis prediction of chemical compounds. The structures of chemical compounds are represented by the first order predicates such as atomel(C;A1; c), bond(C;A1;A2;BT), aromatic ring(C; S1) and alcohol(C; S2). The first two state that A1 which is a carbon atom bond to A2 where the bond type is BT in a chemical compound C. The third represents that substructure S1 is an aromatic ring in a chemical compound C, and the last represents that S2 is an alcohol base in C. Because this approach allows variables to be introduced in the arguments of the predicates, the class of structures which can be searched is more general than graphs. However, this approach easily faces the high computational complexity due to the equivalence checking under μ-subsumption (an NP-complete operation) on clauses and the generality of the problem class to be solved. To alleviate this difficulty, a new system called FARMAR has recently been proposed. It also uses the level wise search, but applied less strict equivalence relation under substitution to reduced atom sets. FARMAR runs two orders of magnitudes faster. However, its result includes some propositions having different forms but equivalent in the sense of the μ-subsumption due to the weaker equivalence criterion. A major advantage of these two systems is that they can 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.

4.3 Inductive Database Based Approach

A work in the framework of inductive database having practical computational efficiency is MolFea system based on the level-wise version space algorithm. 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. The monotonic and anti- monotonic mining measures described in define borders in the version space. To define the borders, the minimal a maximal elements of a set in terms of generality are introduced.

4.4 Mathematical Graph Theory Based Approach

The mathematical graph theory based approach mines a complete set of subgraphs under mainly support measure. The initial work is AGM (Apriori-based Graph Mining) system. The basic principle of AGM is similar to the Apriori algorithm for basket analysis. Starting from frequent graphs where each graph is a single vertex, the frequent graphs having larger sizes are searched in bottom up manner by generating candidates having an extra vertex.

5 Our Study Results

Our study results indicate that Subdue CL, the graph-based relational concept learner, is competitive with logic-based relational concept learners on a variety of domains. This comparison has identified a number of avenues for enhancements. Subdue CL would benefit from the ability to identify ranges of numbers. We could accomplish this by utilizing the system's existing capability to find similar but not exact matches of a substructure in the input graph. Numeric values within the instances could be generalized t o then compassing range. A graph-based learner also needs the ability to represent recursion, which plays a central part in many logic-based concepts. More research is needed to dentify representational enhancements for describing recursive structures — for example, graph grammars. Our future work will also focus on extending Subdue to handle other forms of learning, such as clustering. We are continuing our testing of Subdue in real-world applications. In biochemistry, for example, we are applying Subdue to data from the Human Genome Project to find patterns in the DNA sequences that indicate the presence of a gene-transcription-factor site. Unlike other approaches to finding patterns in gene data, Subdue uses a graph to represent structural information in the sequence. We hope that the discovered patterns will point to genes in uncharted areas of the DNA sequence. In another area of chemistry, we are applying Subdue CL to the Predictive Toxicology Challenge data. This data contains the structural descriptions of more than 300 chemical compounds that have been analyzed focarcinogenicity. Each compound (except for about 30 held out for future testing) is labeled as either cancer-causing or not. Our goal is to find a pattern in the cancerous compounds that does not occur in the noncancerous compounds. So far, Subdue CL has found several promising patterns, which are currently under evaluation in the University of Texas at Arlington's Department of Chemistry. In addition, we are applying Subdue to a number of other databases, including the Aviation Safety Reporting System database, US Geological Survey earthquake data, and software call graphs. Subdue has discovered several interesting patterns in the ASRS database. Burke Burkart of UTA's Department of Geology evaluated Subdue's results on the geology data and found that Subdue correctly identified patterns dependent on earthquake depth, often the distinguishing factor among earth quaktypes.

These and other results show that Subdue discovers relevant knowledge in structural data and that it scales to large databases

6 Conclusion

There are many other studies related to graph mining. An approach is proposed to derive induced subgraphs of graph data and to use the induced subgraphs as attributes on decision tree approaches. The method can be used to find frequent induced subgraphs in the set of graph data. A method to completely search homomorphically equivalent subgraphs which are the least general over a given set of graphs and do not include any identical triplet of the labels of two vertices and the edge direction between the vertices within each subgraph. It

show that the computational complexity to find this class of subgraphs is polynomial for 1/2 locally injective graphs where the labels of any two vertices pointing to another common node or pointed from another common vertex are not identical. However, many graphs appearing in real-world problems such as chemical compound analysis are more general, and hence the polynomial characteristics of this approach do not apply in real cases. In addition, this approach may miss many interesting and/or useful subgraph patterns since the homomorphically equivalent subgraph is a small subclass of the general subgraph. In this paper, the theoretical basis of the graph-based data mining was explained from multiple points of views such as subgraph types, subgraph isomorphism problem, graph invariants, mining measures and search algorithms. Thus, representative graph-based data mining approaches were shown in the latter half of this article. Even from theoretical perspective, many open questions on the graph characteristics and the isomorphism complexity remain.

ACKNOWLEDGMENT

We are grateful to many people for their help in writing this paper. First of all, we would like to Pramod Kumar, Ashwani Yadav and the anonymous reviewers for their work and valuable comments that have significantly improved the quality of our initial manuscript. Our thanks to, Vandana Mam, Anju Bala, Pooja Siroha and especially Kuldeep Kumar for the care with which they reviewed the original draft; and for conversations that clarified our thinking on this and other matters. We would also like to thank Urvashi Bakshi for their encouragement and instruction.

REFERENCES

[1] MRDM'01: Workshop multi-relational data mining. In conjunction with PKDD'01 and ECML'01, 2002. http://www.kiminkii.com/mrdm/.
[2] R. Agrawal and R. Srikant . Fast algorithms for mining association rules. In VLDB'94: Twentieth Very Large Dada Base Conference, pages 487{499, 1994.
[3] J. Cook and L. Holder. Substructure discovery using minimum description length and background knowledge. J. Artificial Intel. Research, 1:231{255, 1994.
[4] L. De Raedt and S. Kramer. The levelwise 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] A. Debnath, R. De Compadre, G. Debnath, A. Schusterman, and C. Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. J. Medicinal Chemistry, 34, 1991.
[6] L. Dehaspe and H. Toivonen. Discovery of frequent data log patterns. Data Mining and Knowledge Discovery, 3(1):7{36, 1999.
[7] T. Gaertner. A survey of kernels for structured data. SIGKDD Explorations, 5(1), 2003.
[8] W. Geamsakul, T. Matsuda, T. Yoshida, H. Motoda, and T. Washio. Classifier construction by graph-based induction for graph-structured data. In PAKDD'03: Proc. of 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining, LNAI2637, pages 52{62, 2003.
[9] P. Geibel and F. Wysotzki. Learning relational concepts with decision trees. In ICML'96: 13th Int. Conf. Machine Learning, pages 166{174, 1996.
[10] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of the ACM, 39(11):58{64, 1996.
[11] A. Inokuchi, T.Washio, and H. Motoda. Complete mining of frequent patterns from graphs: Mining graph data. Machine Learning, 50:321{354, 2003.
[12] I. Jonyer, L. Holder, and D. Cook. Concept formation using graph grammars. In Workshop Notes: MRDM 2002 Workshop on Multi-Relational Data Mining, pages 71{792, 2002.
[13] H. Kashima and A. Inokuchi. Kernels for graph classification. In AM2002: Proc. Of Int. Workshop on Active Mining, pages 31{35, 2002.
[14] R. Kondor and J. Laferty. Difusion kernels on graphs and other discrete input space. In ICML'02: Nineteenth International Joint Conference on Machine Learning, pages 315{322, 2002.
[15] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In ICDM'01: 1st IEEE Conf. Data Mining, pages 313{320, 2001.
[16] M. Liquiere and J. Sallantin. S tructural machine learning with galois lattice and graphs. In ICML'98: 15th Int. Conf. Machine Learning, pages 305{313, 1998.
[17] H. Mannila and H. Toivonen. Discovering generalized episodes using minimal occurrences. In 2nd Intl. Conf. Knowledge Discovery and Data Mining, pages 146{151, 1996.
[18] B. Mckay. Nauty users guide (version 1.5). Technical Report Technical Report, TR-CS-90-02, Department of computer Science, Australian National University, 1990.
[19] A. Mendelzon, A. Mihaila, and T. Milo. Querying the world wide web. Int. J. Digit. Libr., 1:54{67, 1997.
[20] S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. J. Logic Programming, 19(20):629{679, 1994.
[21] S. Nijssen and J. Kok. Faster association rules for multiple relations. In IJCAI'01: Seventeenth International Joint Conference on Artificial Intelligence, volume 2, pages 891{896, 2001.
[22] A. Srinivasan, R. King, and D. Bristol. An assessmen t of submissions made to the predictive toxicology evaluation challenge. In IJCAI'99: Proc. of 16th International Joint Conference on Artificial Intelligence, pages 270{275, 1999.
[23] V. Vapnik. The Nature of Statistical Learning Theory. Springer Verlag, New York. 1995.
[24] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In ICDM'02: 2nd IEEE Conf. Data Mining, pages 721{724, 2002.
[25] K. Yoshida, H. Motoda, and N. Indurkhya. Graph- based induction as a unified learning framework. J. of Applied Intel, 4:297{328, 1994.

- *Seema is currently pursuing masters degree program in Computer Science and Engineering in Gurgaon College of Engineering, India, Mb.No.: 8802473502. E-mail: sikhujakhar@yahoo.in*
- *Dharmveer Yadav ,Pramod Kumar are currently working in Gurgaon College of Engineering as HOD,Assistant Professor, India, Mb.No.:9416996656,9718167583,E-mail: dharmvryadav@gmail.com , Pramod2323@gmail.com .*