ANALYSIS

21(69), July 1, 2014

ISSN 2278–5469 EISSN 2278–5450

Discovery

A study of mining frequent subgraphs using graph classification algorithms

Naga Jyothi P[1], Rajya lakshmi D[2], Chandra Prakash V[3]

1.Department of IT, GITAM University, Visakhapatnam, Andhra Pradesh, India; E-mail: pbtjyothiraj.33@gmail.com
2.Department of CSE & IT, JNT University, Vizianagaram, Andhra Pradesh, India; E-mail: rdavuluri@yahoo.com
3.Department of CSE, K L University, Vijayawada, Andhra Pradesh, India; E-mail: vchandrap@kluniversity.in

Publication History
Received: 06 May 2014
Accepted: 13 June 2014
Published: 1 July 2014

Citation
Naga Jyothi P, Rajya lakshmi D, Chandra Prakash V. A study of mining frequent subgraphs using graph classification algorithms. *Discovery*, 2014, 21(69), 135-139

ABSTRACT

There are many underlying relationships among real world data in several areas of science and engineering like computer vision, molecular chemistry and biology, pattern recognition, bioinformatics, web exploration and data mining, which can be represented in terms of graphs. Progressively the research had been extended from mining frequent items and sequences to mining structures which takes in the form of trees, graphs, and lattices. Graphs can be used to represent this type of complicated structure. So mining large graph is really big task for the graph classification problem .Graph Classification problem is learning to classify graphs in a graph database into two or more categories. Mining Frequent Sub-Structure or Sub Graph is the major role for classifying the given input graph. An enormous study has been carried under the graph classification problem which resulted in better performance against various application domains. With the fast accumulation of graph data, building highly accurate predictive models data emerges as new challenge in the data mining community. This paper presents a comparative study of major approaches for mining frequent subgraphs which is useful for graph classification namely Subdue, gSpan, gSSC, MIGDAC, and MISMOC.

Keywords: Frequent Sub graph; Graph Classification; Graph Database; Predictive models

1. INTRODUCTION

Conventional machine learning research represents data in a tabular form. This kind of notation is successful in diverse real world domains but the inherent meaning is being lost. First-order logic and graphs are used to express relationships within data [1]. Graphs are emerging as a prevalent form of representation in various fields of neuroscience, chemistry and optical character recognition. Graph is defined as a vertex, edge pair where vertex denotes node, atom, atom type, charge etc., and edge denotes relationship or bond between nodes or atoms. A novel approach such as Graph mining has been introduced to best. Graph based mining has prevalent usage in various application areas.

DEFINITION 1.1: (Connected Graph)

A graph is represented as G= (V,E,L) , where V is a set $V=\{v_1,v_2...v_n\}$,$E\subseteq V \times V$ is a set of edges, L is the set of symbols for vertices and edges. A connected graph is a graph such that there is a path between any pair of vertices.

DEFINITION 1.2: *(Sub Graph)*

Let G' =(V',E',L') and G=(V,E,L) be the connected graphs. G' is a sub graph of G(G'$\sqsubseteq$ G) iff: (1) V' $\subseteq$V ,

(2) E' $\sqsubseteq$E,(3)L'$\sqsubseteq$L. If G' is a subgraph of G, then G is a supergraph of G'.

A. Graph classification

Graph classification comprises of two tasks

Primary task constructs a model to predict the class label of whole graph.

Secondary task is to predict the class label of unknown graph from large dataset. This is called as label propagation.

DEFINITION 1.3: *(The Graph Classification Problem)*

A labelled graph is defined as G=(V,E,α,β) where V is set of vertices, $E \subseteq V \times V$ is the set of edges, α:V ⟶ $\sum_V$, where $\sum_V$ is the alphabet of vertex of labels and β is the edge labelling function where β:E ⟶ $\sum_E$ is the alphabet of edge labels. Given a set of training examples T=$\{<x_i,y_i>\}_{i=0}^{L}$, $x_i \in \chi$ is a graph and $y_i \in\{+1,-1\}$, the graph classification problem is to induce a mapping f: $\chi \longrightarrow \{+1,-1\}$.

A general approach for graph classification problem is to first find the frequent sub graphs from graph database, then find interesting measure of sub graphs, whether it is useful for classification or not with suitable threshold. Finally, use those subgraphs for graph classification to evaluate them for different classifiers [2],[3]. As it is typical task, to extract informative sub graph features from a set of graph data. Mostly it uses some filtering criteria by considering its label information. It is of 2 approaches: Supervised and Unsupervised.

Graph classification is of two types. In the supervised approach, it collects a set of features from graphs, and this feature vector may be applied for classification with any of the existing classification methods. Whereas in Unsupervised approach is to implicitly collect a set of features from graphs rather than computing the features, it computes the similarity of graphs.

B. Frequent Sub graph discovery

Pattern discovery is to select the patterns in a Graph Database (GDb) termed as frequent pattern or frequent subgraph. Given a graph G, its support value in a GDb is the number of times the graph occurs in the DB. If this number is sufficiently large, as compared to specific threshold, the graph G is called as frequent sub graph.

The algorithms to find the frequent patterns are divided into 3 categories.

Apriori based graph mining (AGM) and Frequent sub graph discovery (FSG) uses level-wise strategy. The algorithms in this strategy mine the frequent sub graph in graph database, starting from selecting frequent single node [4],[5]. From that node, all the candidate frequent single edge is proposed and their frequency is determined by linear scan of related GDB. The second strategy uses depth first search, include gSpan and FFSM utilizes backtrack algorithm to mine the frequent sub graphs. Starting from one frequent of G, select frequent ones and performs the search recursively. Finally, the third category of frequent sub graph mining algorithms, work on graph space to identify frequent ones. Algorithms of this strategy first project a graph space to another space such that of trees, then identify frequent patterns in the projected space and finally reconstruct all frequent patterns in the graph space. This strategy is called as progressive mining.

2. APPROACHES TO GRAPH CLASSIFICATION

A. SUBDUE

Subdue is an unsupervised graph based relational learning system. The objective of this algorithm it uses heuristic search for sub graphs present in positive examples and absent in negative examples. The hypothesis space of Subdue consists of all connected sub graphs labeled positive. Subdue discovers the substructure or sub graph of input graph that best compresses input graph dataset according to Minimum Description Length(MDL)principle. Subdue algorithm performs computationally–constrained beam search which begins from substructures consisting of all vertices with unique labels. Extending this substructure by a single edge and vertex in all possible ways to generate candidate substructures. All the instances of substructures are maintained as queue based on their compression values i.e., using MDL or classification accuracy. It uses graph isomorphism to determine the instance of candidate substructure in the input graph. The length of beam is determined by the number of substructures retained for further expansion. Only the top beam substructure remains on the queue for expansion for the next pass through main discovery loop. The search terminates upon reaching the limit and removes the list of best substructures. Thus, the procedure ends until all positive examples covered by best substructure [6],[7]. This model thus consists of a decision list each member of connected graph, to classify unseen example by testing its sub graph isomorphism. If unseen example is present in the decision list then it is predicted as positive otherwise, negative. Thus, subdue not only acts as unsupervised discovery system but also acts as supervised for relational graph based concept learning so it is referred to as SubdueCL.

B. MIGDAC:

A novel graph mining algorithm Mining Graph Data for Classification (MIGDAC), is a supervised learning algorithm. MIGDAC applies graph theory and an interestingness measure to discover sub graphs, which can be both characterized and easily distinguished from other classes. The key aspect of this algorithm is first transforming the input graph into hierarchical graphs. MAGMA (Multi-Level Attributed Graph Mining algorithm) is the basis for constructing of hierarchical graphs. MAGMA is to group components of attributed graph into different levels according to their attributed structural relations. MIGDAC extracts a set of class specific patterns defined in terms of an interestingness threshold and measure with residue analysis. The next step is to use weight of evidence to identify class-specific pattern and thus will be positively or negatively characterized for a class.

The problem definition of MIGDAC is as follows for given class of graph data samples Gc ,each consists of set of sub graphs M_j =(m_{ji}, m_{j2} ,m_{j3},...... ,m_{jn}), first the algorithm gets a set of extracted sub graphs and determines the frequency count of m_{j1} to m_{jn} occurrences in each class. Then calculate interestingness measure d for each m_{jk} in M_j as D_j =(d_{ji},d_{j2},d_{j3}...d_{jn}).So the subgraph, which qualifies for characterization will be considered as class specific pattern and remaining will be filtered out [8]. MIGDAC algorithm offers four benefits than the existing algorithms like firstly, construction of hierarchical graphs to represent the graph samples. Second it discovers graph patterns that are class specific resulting in higher classification accuracy. Third, usage of class specific patterns reduces the number of potential interesting patterns that speeds up the graph classification problem. Fourth, to identify patterns that is distinguishable between classes by using weight of evidence rather than frequency. With, this it is widely applicable to variety of datasets.

C. gSSC:

Semi supervised Feature Selection For Graph classification (gSSC) is a Semi supervised learning algorithm (gSSC Algorithm) .In most of the applications ,the labelled graph data is very expensive or difficult to obtain, as there are abundant amounts of unlabelled graph data is available. A major difficulty in graph classification lies in the complex structure of graphs and lack of vector representation. The key aspect of gSSC is to find the the most informative subgraph patterns from a number of labelled graphs and a large number of unlabelled graphs. In order to find the most optimal subgraph, the need to select the proper set of features. For feature evaluation criteria, gSemi is used to estimate the usefulness of subgraph features based upon labeled and unlabeled graphs. Then application of branch and bound algorithm to efficiently search for optimal subgraph features by prudently pruning the subgraph search space.

The problem formulation of gSSC is as follows D={G_1,G_2,.....G_n} denotes graph dataset as both labeled and unlabelled, having n graph objects as connected graphs [9]. Assume that the first l graphs within D are labelled graph dataset D_1={ G_1,G_2... G_l} and unlabelled as D_u={ G_{l+1}...G_n}, D= D_i $\cup$ D_u. The idea of subgraph-based graph classification approach is to assume that each graph object G_i is represented as feature vector x_i=$[x_i^1 ..... x_i^m]_T$ corresponding to set of subgraphs patterns {g_1...g_m}. Denote x_i^k as the binary feature associated with subgraph g_k. x_i^k =1 iff g_k is a subgraph of G_i($g_k \sqsubseteq G_i$), otherwise x_i^k =0.

Semi-Supervised feature selection approach for graph classification outperforms supervised and unsupervised approaches and is very efficient by pruning the subgraph search space using both labelled and unlabelled graphs.

D. gSpan:

Graph based substructure Pattern Mining (gSpan) is also a popular graph–mining algorithm that has been used for graph classification. The key aspect of gSpan is searching for frequent subgraph on canonical forms using a depth-first search (DFS) strategy.gSpan uses two techniques, DFS lexographic order and minimum DFS code. Firstly, it chooses a random vertex, then visiting and marking of vertices continues to which the chosen vertex is connected to .This process is repeated processing until full DFS tree is constructed. So for each graph searched, graph it is possible to have more than one tree with DFS depending on the order in which the vertices are visited. The gSpan eliminates candidate generation and prunes all false positives of discovered frequent subgraph [10].

DFS code: Given a DFS tree T for a graph G, an edge sequence (e_i) can be constructed on $\prec$T. Such that $e_i \prec T\ e_{i+1}$, where i=0,....|E| -1. (ei) is called a DFS code, denoted as code(G,T).

DFS Lexicographic Order: Suppose Z={ code(G,T) | T is a DFS of G}, i.e., Z is a set containing all DFS code for all the connected labelled graphs. Suppose there is a linear order ($\prec$L) in the label set (L), and then the lexicographic combination of $\prec$T and $\prec$L is a linear order ($\prec$e) on the set $ET \times L \times L \times L$.

DFS Lexicographic order is defined as linear order. If $\alpha=code(G_\alpha,T_\alpha)=(a_0,a_1,....a_m)$ and $\beta=code(G_\beta,T_\beta)= (b_0,b_1,....b_n),\alpha,\beta \in Z$ then $\alpha \leq \beta$ iff either of the following is true.

1) $\exists t, 0 \leq t \leq \min (m,n), a_k=b_k$ for $k< t$, $a_t \prec e\ b_t$.

2) $a_k=b_k$ for $0\leq k\leq m$, and $n\geq m$.

Minimum DFS code: Given a graph G, Z(G)= {code(G, T) | T is a DFS tree} based on , DFS Lexicographic order, the minimum one, min(Z(G)), is called minimum DFS code of G or canonical label of G. Thus, the problem of mining frequent connected subgraphs is equivalent to their corresponding minimum DFS codes [11]. Thus, gSpan is capable to mine large frequent subgraphs in a bigger graph set with lower minimum supports.

E. MISMOC:

Mining Interesting Substructures in MOlecular Data for Classification (MISMOC) is supervised learning algorithm. The key idea of MISMOC is to discover the interesting structural patterns in data such that “unseen” molecules that are not originally in the dataset can be classified accurately. Algorithm is to mine the interesting substructures in a molecular database for classification and can discover the interesting frequent sub-graphs for characterization of a class and discriminate it from the other classes [12]. MISMOC algorithm first finds the frequent sub-graphs of the molecule by using FSG or gSpan algorithm, with a specified threshold, then the interestingness for each of these sub-graphs for each class is calculated [13]. Choice of using threshold doesn’t allow any unique frequent subgraph to be discovered for each class. If same subgraph occurs in 1 or 2 classes then information theoretic measure of each subgraph is used in each class. This is useful for classification and all the uninteresting ones should be screened out. Based on the weight of interestingness measure, the unknown molecule is classified. Degree of interestingness is based on the use of information theoretic measure called as weight of evidence. This is combined to form overall total interestingness for the purpose of classifying an unseen graph [14] .Thus Algorithm filters out subgraphs, which do not occur frequently and are irrelevant.

MISMOC problem formulation is stated as follows:

Given a set of molecular data G, containing n molecules pre-classified into p classes, the molecular graphs G_1, G_2, G_3............. G_n, where $G_i= G_i(V_i,E_i)$, $i\in\{1,.....n\}$ is a labeled graph with vertices representing atoms and edges representing bonds between atoms. The p classes that the n molecules and their corresponding molecular graphs are classified into can be represented as $C_{(1)}$,.....$C_{(p)}$,where $C_{(i)}=\{G_1^{(i)},.......G_{C_i}^{(i)}\}\subseteq G$, i=1...p.The algorithm will provide better performance than the existing algorithms for frequent sub-structure mining for molecular classification.

3. CONCLUSION

The objective of this paper is to focus on the various approaches of Graph Classification mining frequent subgraphs from various graph datasets observes some special features of Subdue algorithm and discovers frequent subgraphs which are fewer in number and higher of interest. It maintains the best set of substructures to compress the graph dataset using heuristic principle called MDL. Subdue is advantageous in approximations of input graph data and improves the computational efficiency. Subdue acts as both supervised and unsupervised learning algorithm. The second category algorithm takes depth first search strategy called gSpan is useful because of its backtrack technique to enumerate the super graph recursively and to build dfs tree in lexicographic order. The minimum dfs code subgraph is used as its canonical labeling. It avoids candidate generation like Subdue and improvises in memory utilization. gSpan and Subdue both are used for finding frequent subgraphs, which are interesting for classification and discrimination from other classes. gSSC is a semi supervised feature selection algorithm .It uses gSemi to evaluate the subgraph features and derive upper-bound for gsemi to prune the subgraph search space. On application of branch and bound algorithm it finds the set of sub graph features which are useful for graph classification. gSSC efficiently works for both labelled and unlabelled graphs. MIGDAC and MISMOC algorithms make use of Subdue and gSpan to find the frequent subgraphs. MIGDAC is advantageous in transforming in to a set of hierarchical graphs so, that simplifies the graph complex structure into more informative than that of traditional formats. Use of interestingness measure to extract class specific pattern and weight of evidence for positive or negative characterization. MISMOC aims at discovering "interesting" subgraphs that do not just occur frequently but can also allow graph class to be better discriminated from another. Such that an "unseen" molecule can be easily classified. MISMOC can better handle big data size and number of graphs with higher accuracy. This paper gives an idea of overall comparative study of approaches for graph classification by projecting the strengths and weakness. The next version of the paper aims at developing classifiers for supervised, unsupervised and semi-supervised approaches for graph classification of an uncertain data with an experimental comparison of both real and artificial datasets.

REFERENCE

1. D.J. Cook and L.B. Holder, Mining Graph Data. Wiley, 2006.
2. Nikhil S.Ketkar, Lawrence B.Holder, Diane J.Cook, "Empirical Comparison on Graph Classification Algorithms," IEEE Symposium on Computational Intelligence and Data Mining, pp. 259-266, 2009.
3. Aaron Smalter, jun Huan,Jia Yi, and Gerald Lushngton, "GPD: A Graph Pattern Diffusion Kernel For Accurate Graph Classification with Applications in Cheminformatics,"IEEE Transactions on Computational Biology and Bioinformatics, vol.7,no.2 ,pp.197-207,2010.
4. A. Inokuchi, T. Washio, and H. Motoda, "An apriori-based algorithm for mining frequent substructures from graph data," in Proc. 4th Eur. Conf. Principles Pract. Knowl. Discov. Databases (PKDD), 2000, pp. 13–23.
5. R. Chittimoori, L. B. Holder, and D. J. Cook, "Applying the subdue substructure discovery system to the chemical toxicity domain," presented at the AAAI Spring Symp. Predictive Toxicol. Chem.: Exp. Impact AI Tools, Menlo Park, CA, 1999.
6. D. Cook and L. Holder, "Substructure Discovery Using Minimum Description Length and Background Knowledge," Journal of Artificial Intelligence Research, vol. 1, pp. 231–255, 1994
7. N. Ketkar, L. Holder, and D. Cook, "Subdue: compression-based frequent pattern discovery in graph data," Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations, pp. 71–76, 2005.
8. M. Deshpande, M.Kuramochi, N. Wale, and G. Karypis, "Frequent substructure-based approaches for classifying chemical compounds," IEEE Trans. Knowl. Data Eng., vol. 17, no. 8, pp. 1036–1050, Aug. 2005.
9. Xiangnan Kong, and Philip S.Yu "Semi-Supervised Feature Selection For Graph Classification," Proc.Int' l Conf Of KDD'10 ,pp. 793-802, 2010.
10. X. Yan and J. Han, "gSpan: Graph-based substructure pattern mining," in Proc. IEEE Int. Conf. Data Mining, 2002, pp. 721–724.
11. S. Ranu and A.K. Singh, "GraphSig: A Scalable Approach to Mining Significant Subgraphs in Large Graph Databases," Proc. Int'l Conf. Data Eng., 2009.
12. C. Borgelt and M. R. Berthold, "Mining molecular fragments: Finding relevant substructures of molecules," in Proc. 2nd IEEE Int. Conf. Data Mining (ICDM), 2002, pp. 51–58.
13. Winnie W. M. Lam, Kieth C.C.Chan "MISMOC: Discovering Interesting Molecular Substructures for Molecular Classification," IEEE Transactions on Nano Bioscience, vol.9, no.2, pp. 77-89, 2010.
14. M. Deshpande and G. Karypis, "Automated approaches for classifying structure," in Proc. 2nd ACM SIGKDD Workshop Data Mining Bioinf., 2002, pp. 11–18.