A Review on Graph-based Image Classification

Mrs. Snehal N. Amrutkar , Prof.J.V.Shinde
1,2 Department of Computer Engineering
University of Pune
Late G.N.Sapkal College of Engineering, Anjaneri, Nashik
INDIA

Abstract: Graphs are increasingly important in modelling complicated structures, such as circuits, images, chemical compounds, protein structures, biological networks, social networks and the web. Graph-based data representations are an important research topic due to the suitability of this kind of data structure to model entities and the complex relations among them. We are proposing to combine a powerful graph-based image representation and frequent approximate sub-graph (FAS) mining algorithms in order to classify images. Among the many approaches in performing image classification, graph based approach is gaining popularity primarily due to its ability in reflecting global image properties. This paper critically reviews existing important frequent subgraph mining algorithms. This review not only reveals the pros in each method and category but also explores its limitations.

Keywords: Frequent approximate sub-graphs, Graph-based image representation, and Image classification

I. Introduction

The need to convert large volumes of data into useful information has been increased. In many datasets, the objects may be represented as graphs. In Fig. 1, an example of graphs for representing images is shown.

As a result of this need, several authors have developed techniques and methods to process these datasets [64,65], e.g. Frequent pattern discovery [66,67,68]. In many research fields, graphs have been largely used to model data due to their representation expressiveness and their suitability for applications where some kind of entities and their relationships must been coded within some data structure.

The discovery of frequent patterns, mainly the detection of frequent sub-graphs in graph collections is an important problem in graph mining tasks [5,69,70]. In frequent sub-graph mining, there are two approaches for evaluating the similarity of graphs, known as graph matching: exact matching and approximate matching.

The exact matching determining whether the labels and the structure of two graphs are identical. This matching has been successfully used in many applications [71,72,73]; however, there are some problems where exact matching could not be applicable with positive outcome. Sometimes, the interesting sub-graphs show slight differences throughout the data. For example these differences can be seen in protein analysis. Other examples can be seen on image processing, where these differences may be due to noise and distortion, or may just illustrate slight spatial differences between instances of the same objects. This means that we should tolerate certain level of geometric distortion, slight semantic variations or vertices and/or edges mismatch in frequent sub-graph pattern search. For this reason, it has become necessary to evaluate the similarity between graphs allowing some structural differences, i.e. approximate matching techniques. The approximate matching consists in finding the match between vertices and/or edges of two graphs in order to determine their similarity, allowing structure and/or label differences. On this basis, the need to perform frequent sub-graph mining using approximate graph matching has been raised [74,75,76,77]. Several authors have expressed the necessity to use approximate graph matching for frequent sub-graph mining on graph collection. These authors defend the idea that frequent and more interesting sub-graphs for applications and users could be found.

Frequent subgraph mining is the processor of extracting all frequent subgraphs from graph dataset who have occurrence count greater than or equal to the specified threshold. Following figure gives example of frequent subgraph mining, where figure 1(a) and 1(b) represents any two graphs and figure 1(c) gives frequent subgraph which is present in both graphs.

These graph mining methods might be grouped as:
A. Apriori-Based Approach: Apriori based frequent substructure mining algorithm share similar characteristics with Apriori-based frequent item set mining algorithms. The search for frequent groups starts with graphs a small “size” and proceeds in a bottom-up manner by generating candidate having an extra vertex, edge or path. The definition of graph size depends on algorithm used.

B. Pattern-Growth Approach: The Apriori-based approach has to use the breadth-first search (BFS) strategy because of its level-wise candidate generation.

II. Survey of algorithms

Various algorithms on graph mining were developed by many researchers. Some algorithms give approximate results while others provide exact results. Some of them are reviewed in this section.


Holder, Cook & Bunke in 1987 presented SUBDUE[2] algorithm which is based on graph edit distance. Using fuzzy graph match, SUBDUE can identify slightly different instances of substructures which are replaced by pointer.

Blockeel and Raedt [3] in 1998 introduced a first-order framework for top-down induction of logical decision tree. Top-down induction of decision trees is the best known and most successful machine learning technique. It has been used to solve numerous practical problems. It employs a divide-and-conquer strategy, and in this it differs from its rule-based competitors which are based on covering strategies. Chakrabarti, Dom and Indyk [4] in 1998 developed a new method for automatically classifying hypertext into a given topic hierarchy, using an iterative relaxation algorithm. After bootstrapping off a text-based classifier, they used both local texts in a document as well as the distribution of the estimated classes of other documents in its neighborhood, to refine the class distribution of document being classified. They discussed three area of research: text and hypertext information retrieval, machine learning in context other text or hypertext, and computer vision and pattern recognition.

Gago-Alonso, Carrasco-Ochoa, Medina-Pagola in 2000 presented gdFil[5] Algorithm which is based on pattern growth. gdFil allow removing all duplicate candidates before support calculation. But, Support calculation requires expensive sub-graph isomorphism tests.

In FSG[6] proposed by Kuramochi in 2001, candidates are generated by increasing one edge at a time. Size of the graph in FSG is proportional to the number of edges. To check unique subgraph, FSG uses canonical labeling technique in which each graph have unique canonical label. Canonical labels of two graphs are same only when both graphs have same topological structure and same edges and vertices labels. FSG also have some drawbacks. It also generates multiple candidates. FSG uses isomorphism testing technique, is very costly.

---

**Figure 2(a): PPI Network**

**Figure 2(b): Inexact Graph Matching for CSMiner[1] Algorithm**

**Figure 2(c): PPI Network**

**Figure 2(d): Inexact Graph Matching for CSMiner[1] Algorithm**

**Figure 3(a): Cortisone**

**Figure 3(b): Substructures found in Cortisone compound for SUBDUE algorithm[2]**

**Figure 4: Multiple Candidate Generation for FSG[6]**
Calders and Wisen [7] in 2001 presented on monotone data mining layer a simple data-mining logic (DML) that can express common data mining tasks like “find Boolean association rules” or “find inclusion dependencies”. Kramer, Raedt, and Helma [8] in 2001 presented the application of feature mining techniques to the developmental therapeutics program’s AIDS antiviral screen database. Kuramochi and Karypis [9] in 2001 presented a computationally efficient algorithm for finding all frequent subgraphs in large graph databases. They evaluated the performance of the algorithm by experiments with synthetic datasets as well as a chemical compound dataset. Pei, Han, Mortazavi-Asl, and Pinto [9] in 2001 proposed a novel sequential pattern mining method called PrefixSpan that is prefix-projected sequential pattern mining. Asai, Abe, and Kawasoe [10] in 2002 discovered the efficient substructure from large semi structure data and patterns by labeled ordered trees and studied the problem of discovering all frequent tree-like patterns that have at least a minimum support in a given collection of semi-structured data. They presented an efficient pattern mining algorithm FREQT for discovering all frequent tree patterns from a large collection of labeled ordered tree. Borgelt and Berthold [11] in 2002 presented an algorithm to find fragments in a set of molecules that help to discriminate between different classes for instance, activity in a drug discovery context. Zaki [12] in 2002 presented TREE_MINER algorithm to discover all frequent subtrees in a forest, using a new data structure called scope-list. Deshpande, Kuramochi, and Karypis [13] in 2002 proposed the technique for classifying chemical compounds. These techniques can be broadly categorized into two groups. The first group consists of techniques that rely mainly on various global properties of the chemical compounds, such as molecular weight, ionization potential, inter-atomic distance etc. for capturing the structural properties of the compounds. Since this information is not relational, existing classification techniques can be easily used on these datasets. However the absence of actual structural information limits the accuracy of such classified. The second group of techniques directly analyzes the structure of the chemical compounds to identify patterns that can be used for classification [14; 16; 15; 8]. One of the earliest studies in discovering substructures was carried out by Dehaspe et al. [8] in 1998 which Inductive Logic Programming (ILP) techniques were used though this approach is quite powerful it is not designed to scale to large graph databases hence may not be able to handle large chemical compound databases. A number of recent approaches focused on analyzing the graph representation of the chemical compounds, to identify frequently occurring patterns, and use these patterns to aid in the classification. Wang et al. [16] in 1997 developed an algorithm to find frequently occurring blocks in the geometric representation of protein molecules and showed that these blocks can be used for classification. Inukuchi et al. [15] developed an algorithm to find all frequently occurring induced sub graphs and presented some evidence that such subgraph can be used to features for future classification. Cooper and Frieze [17] in 2003 described a general model of a random graph process whose proportional degree sequence obeys a power law. These law recently observed in graph associated with www. Dzeroski [18] in 2003 introduced the Multi-Relational Data Mining. Geotool [19] in 2003 studied on link mining. Link among the objects may demonstrated certain patterns which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical models. Link mining is promising new area where relational learning meets statistical modeling. Hu, Wang, and Prince [20] in 2003 proposed a novel subgraph mining algorithm: FSM, which employs a vertical search scheme within an algebraic graph framework. They have developed to reduce the number of redundant candidates proposed. Their studied on synthetic and real datasets demonstrates that FSM achieves a substantial performance gain over the current start-of-the-art subgraph mining algorithm SPAN. Washio and Motoda [21] in 2003 introduced the theoretical basis of graph based data mining and surveys the state of the art of graph-based data mining. Yan, Han, and Afshar [22] in 2003 proposed an alternative but equally powerful solution which is based on pattern growth: instead of mining the complete set of frequent subsequences, they mined frequent closed subsequences only that are those contained no super-sequence with the same support. They also introduced an efficient algorithm, called CloSpan. (CloSpan is stand for closed sequential pattern mining.) This outperforms the previous work by one order of magnitude. Moreover a deep a deep understanding of efficient sequential pattern mining methods may also have strong implications on the development of efficient methods for mining frequent subtrees, lattices, subgraphs, and other structured patterns in large databases. The sequential pattern mining algorithms developed so far have good performance in databases consisting of short frequent sequences. Unlike Apriori algorithm, gSpan does not generate candidate patterns & tests for false positive pruning which is time & space efficient but, it has high computational complexity.

Yan and Han [23] in 2003 proposed to mine close frequent graph patterns. A graph g is closed in a database if there exists no proper subgraph of g that has the same support as g. A closed graph pattern mining algorithm, CloseGraph, is developed by exploring several interesting looping methods. Their performance studied shown that CloseGraph not only dramatically reduces unnecessary subgraphs to be generated but also substantially increases the efficiency of mining, especially in the presence of large graph patterns. Yin and Han [24] in 2003 developed a new classification approach is called CPAR (CPAR is classification based on predictive Association Rules). Based on their study performance study, CPAR achieved high accuracy and efficiency, which have many useful features. CPAR represents a new approach towards efficient and high quality...
classification. It is interesting to further enhance the efficiency and scalability of this approach and compare it with other well-established classification schemes. Moreover, the strength of the derived predictive rules also motivates us to perform an in-depth study on alternative approaches towards effective association rule mining.

Huan, Wang, Prins and Yang [25] in 2004 developed a new algorithm that mines only maximal frequent subgraphs, that is subgraph that are not a part of any other frequent subgraphs. Their algorithm can achieve a five-fold speed up over the current state-of-the-art subgraph mining algorithms. Their mining method is based on a novel graph mining framework in which they first mine all frequent tree patterns from a graph database and then construct maximal frequent subgraphs from trees. Huan, Wang Bandyopadhyay Snoeyink, and Prins [26] in 2004 applied a novel subgraph mining algorithm mining algorithm to three related graph representation of the sequence and proximity characteristics of a proteins amino acid residues. The subgraph mining algorithm is used to discover spatial motifs that can be used to discriminate among protein in different families found in the SCOP database. Koyuturk, Grama, Szpankowski, [27] in 2004 presented an innovative new algorithm for detecting frequently occurring patterns and modules in biological network. They show experimentally that their algorithm can extracted from the KEGG database within seconds. The proposed model and algorithm are applicable to a variety of biological networks either directly or with minor modification. Meinl, Borgelt and Berthold [28] in 2004 shown that is possible to mine meaningful, discriminative molecular fragments from large databases. Using an existing algorithm that employs a depth-first strategy and a sophisticated ordering scheme allows avoiding costly re-embeddings throughout the candidate growth process, which in turn enables us to find also larger fragments. Yin, Han, Yang and Yu [29] in 2004 developed a CrossMine, an efficient and scalable approach for multi-relational classification. Several novel methods are developed in CrossMine, including 1. tuple ID propagation, which performs semantics-preserving virtual join to achieve high efficiency on databases with complex schemas. 2. a selective sampling method which makes it highly scalable with respect to the number of tuples in the databases. Both theoretical backgrounds and implementation techniques of CrossMine are introduced. Yan, Yu and Han [30] in 2004 discovered the issues of indexing graphs and proposed a novel solution by applying a graph mining technique. Different from the existing path-based methods, our approach, called gIndex, makes use of frequent substructure as the basic indexing feature. Frequent substructures are ideal candidates since they explore the intrinsic characteristics of the data and are relatively stable to database updates. gIndex has 10 times smaller index size, but achieves 3-10 times better performance in comparison with a typical path-based method.

Yongling Song, Su-Shing Chen in 2005 presented RNGV[31] Algorithm which is based on graph edit distance. RNGV algorithm uses modified Apriori algorithm, which treat edges as items. Perform edit operations that includes vertex & edge insertion, deletion & relabeling.RNGV has a limitation that it Explore possible edit paths keeping one expected as a candidate and this do not claim completeness. Hu, Yan, Huang, Han and Zhou [32] in 2005 developed a novel algorithm, CODENSE, to efficiently mine frequent coherent dense subgraphs across large number of massive graph on biological networks for function discovery. Li and Tan [33] in 2005 proposed a novel graph mining algorithm to detect the dense neighborhoods in an interaction graph which may correspond to protein complexes. Their algorithm first located local cliques for each graph vertex and then merge the detected local cliques according to their affinity to form maximal dense regions. Yan, Yu and Han [34] in 2005 investigated the issues of substructure similarity search using indexed features in graph databases. By transforming the edge relaxation ratio of a query graph into the maximum allowed missing features, their structural filtering algorithm called Grafil, can filter many graphs without performing pairwise similarity computations. Yan, Han and Yu [35] in 2005 proposed a new approach, called CROSSCLUS, which performs cross-relational clustering with user’s guidance.

Kuramochi and Karypis in 2006 presented GREW[36] Algorithm which is based on sub-isomorphism employing ideas of edge contraction & graph rewriting . They discovers frequent sub-graphs by repeatedly merging embeddings of existing freq. sub-graphs. Chakrabarti and Faloutsos [37] in 2006 discussed the problem of detecting abnormalities in a given graph and of generating synthetic but realistic graphs have received considerable attention recently. Both are tightly coupled to the problem of finding the distinguished characteristics of real-world graphs i.e. the patterns that show up frequently in such graphs and can be considered as marks of realism. Karunaratne and Bostrom [38] in 2006 developed a method for learning from structured data are limited with respect to handling large isolated substructures and also imposed constraints on search depth and induced structure length. An approach to learning from structured data using a graph based canonical representation method of structured called finger printing. Krasky, Rohwer, Schroeder and Selzen [39] in 2006 discussed on a combined bioinformatics and chemoinformatics approach for the development of new ant parasitic drugs. Meinl, [40] in 2006 solved the problem Parallel molecular mining Worlein, Urzova, Fischer, and Philippsen which are used in chemoinformatics to find common molecular fragments a database of molecules represented as two-dimensional graphs. In ParMol package they have implemented four of the most popular frequent subgraph miners using a common infrastructure: MoFa, gspan, FFSM and Gaston. They also added additional functionality to some of the algorithms like parallel search, mining directed graphs and mining in one big graph instead of a graph database. Meinl, Worlein, Fischer, and Philippsen [41] in 2006 presented the
thread-based parallel versions of MoFa and gSpan that achieve speedup up to 11 on a shared memory SMP system using 12 processors. They discussed the design space of the parallelization, the results, and the obstacles, that are caused by the irregular search space and by the current state of Java technology. Tsuda and Kudo [42] in 2006 proposed a graph clustering method that selects informative patterns at the same time. Their task is analogous to feature selection for vectors however the difference is that the features are not explicitly listed. This method is fully probabilistic adopting a binomial mixer model defined on a very high dimensional vector indicating the presence or absence of all possible patterns. Wegner, Frohlich, Mielenz and Zell [43] in 2006 presented a classification method which is based on a coordinate-free chemical space. Thus it does not depend on descriptor values commonly used coordinated-based chemical space methods.

Figure 5: Sequence of edge collapsing operations performed by GREW [36]

Chen, Yan, Zhu, Han in 2007 presented gApprox [44] algorithm, which is based on substitution probabilities. Due to noise & distortion, it considers approximation. So interesting patterns can be captured. gApprox algorithm finds patterns for a single network not for a set of graphs. Shijie Zhang & Jiong Yong in 2007 presented Monkey [45] Algorithm based on β-edge sub-isomorphism It is capable of finding frequent patterns that were missed by exact matching algorithms in presence of edge distortions. They handles only missing edge & edge label mismatch. Merkwirth and Ogorzalek [46] in 2007 described a method for construction of specific types of neural networks composed structures directly linked to the structure of the molecule under consideration. Each molecule can be represented by a unique neural connectivity problem (graph) which can be programmed on to a cellular neural network. A composite network can further successfully perform classification and regression on real-world chemical data-set. Dong, Gilbert, Guha, Heiland, Kim, Pierce, Fox, and Wild [47] in 2007 developed an infrastructure of chemoinformatics web service that simplifies that access to this information and the computational techniques that can be applied to it. They described this infrastructure and give some examples of its uses and then discuss their plans to use it as a platform for chemoinformatics application development in the future. Rhodes, Boyer, Kreulex, Chen, and Ordonez [48] in 2007 developed a system that allow user to explore the US patent corps using Molecular information. Their system contains three main technologies. In this system a user may go to a web page, draw a molecule search for related Intellectual property (IP) and analyzed the results.

Figure 6: Exploring the Pattern Space in [44]: (a) the original graph, (b) decomposition of the problem and stage 1, (c) stage 2, (d) stage 3, (e) stage 4, where each stage is enclosed respectively within a loop. We darken marked vertices along the way, while edges are correspondingly changed from dotted to thickened as new vertices are absorbed

Figure 7 β-edge isomorphism (β>=2) for Monkey algorithm [45]
Bogdanov [49] in 2008 studied on Graph searching, indexing, mining and modeling for Bioinformatics, chemoinformatics and Social network. Fahim et al. [50] in 2008 proposed a method which is based on shifting the center of the large cluster toward the small cluster and recompiling the membership of small cluster points, the experimental results reveal that the proposed algorithm produce satisfactory results. Godeck and Lewis [51] in 2008 stated that QSAR models can play a vital role in both the opening phase and the endgame of lead optimization. In the opening phase there is often a large quantity of data from high throughput screening (HTS) and potential leads need to be selected from several distinct chemotypes. Guha and Schurer [52] in 2008 investigated various aspects of developing computational models to predict cell toxicity based on cell proliferation screening data generated in the MLSCN. By capturing feature-based information in that data set, such predictive models would be useful in evaluating cell-based screening results in general and could be used as an aid to identify and eliminate potentially undesired compounds. Hubler, Kriegel, Borgwardt, Gahramani and Metropolis [53] in 2008 presented metropolis algorithm for sampling a representative small subgraph from the original large graph with representative describing the requirement that the sample shall preserve crucial graph properties of the original graph. Karunaratne and Bostrom [54] in 2008 presented a case study in chemoinformatics in which various types of background knowledge are encoded in graphs that are given as input to a graph learner. In this paper shown that the type of background knowledge encoded indeed has an effect on the predictive performance. Lam and Chan [55] in 2008 studied on graph data mining algorithm are increasingly applied to biological graph data set. In this paper they proposed graph mining algorithm MIGDAC (Mining graph data for classification) that applies on graph theory and an interestingness measure to discover interesting sub graphs which can be both characterized and easily distinguished from other classes. Maji, and Mehta [56] in 2008 proposed a novel a measure of similarity between labeled graphs which has applications to structured data analysis for example chemical informatics, web document clustering etc. their metric on graphs exploits vertex context similarity and computes an overall matching score in polynomial time in the size of the graphs using a network flow formulation of the problem. Tsuda and Kurihara [57] in 2008 proposed a nonparametric Bayesian method for clustering graph and selecting salient patterns at the same time. Variation inference is adopted here because sampling is not applicable due to extremely high dimensionality. Schietget, Costa, Ramon, and Raedt [58] in 2009 proposed a direct efficient and simple approach for generation of interesting graph pattern. They computed maximum common subgraph from randomly selected pairs of examples and directly use them as features. Zou, Li, Gao, and Zhang in 2010 developed MUSE[59] Algorithm which is based on sub-isomorphism on uncertain graphs. MUSE algorithm find set of FAS patterns by allowing error tolerance on expected support of discovered sub-graph patterns. Jiang, Coenen and Zito [60] in 2010 examined a number of edge weighting schemes; and suggested three strategies for controlling candidate set generation. Spjuh, Willighagen, Guha, Eklund and Wikberg [59] in 2010 studied on toward interoperable and reproducible QSAR analysis: Exchange of data sets. QSAR is widely used method to relate chemical structures to responses or properties based experimental observations. Much effort has been made to evaluate and validate the statistical modeling QSAR, but these analyses treat the dataset as fixed. An overlooked but highly important issue is the validation of the setup of the dataset, which comprises addition of chemical structure as well as selection of descriptors and software implementations prior to calculations. This process is hampered by the lack of standard and exchange formats in the field, making it virtually impossible to reduce and validate analyses and drastically constrain collaboration and re-use of data. Yang, Parthasarathy and Sadayappan[60] in 2010 presented a novel approach to data representation for computing this kernel, particularly targeting sparse matrices representing power-law graphs. They shown their representation scheme, coupled with a novel tiling algorithm, can yield significant benefits over the current state of the art GPU and CPU efforts on a number of core data mining algorithms such as PageRank, HITS and Random Walk with Restart. A graph transaction is represented by adjacency matrices and the frequent patterns appearing in matrices are mined through the extended algorithm. These are modeled as attribute graph in which each vertex represents an atom and each edge a bond between atoms. Each vertex carries attribute that indicates the atom type. Papapetrou,Ioannou in 2011 developed UGRAP[61] algorithm which is based on sub-isomorphism on uncertain graphs. This algorithm required for applications which has uncertainty in their data. UGRAP algorithm uses an Index (containing graph edges & their probabilities) which reduces number of comparisons required for computing expected support of each candidate pattern. And improve efficiency by optimization for early termination & scheduling of graph comparisons. ChenJia, Zhang, Huan in 2007 presented APGM [62] algorithm and Acosta-Mendoza, Gago-Alonso, Medina-Pagola in 2012 presented VEAM[63] algorithm which are based on substitution probabilities. Due to noise & distortion, it considers approximation. So interesting patterns can be captured. APGM and VEAM Algorithms specify which vertices, edges or labels can replace others to perform frequent sub-graph mining. These methods perform feature (sub-graph) selection taking into account semantic distortions.
APGM algorithm allows these distortions only between vertices of graphs, while VEAM algorithm allows distortions between vertices and between edges.

**III. Conclusion**

Recently, there has been increasing interest in using graph based methods as a powerful tool for image classification. This review has discussed some of the major graph mining methods and highlighted their strengths as well as limitations. Some difficulties of these methods have brought down their use in practical applications. In case of exact matching, sometimes, the interesting sub-graphs show slight differences throughout the data. The current research in graph mining methods orients towards producing approximate solution (sub-optimal solution) for such graph matching problem to find frequent and more interesting sub-graphs for applications. In fact, many of the graph mining methods have their own disadvantages. This survey demonstrates that most of the methods fail to find use in many real time applications. This demonstrates the need for further research to refine the graph mining techniques to be applied for real time scenarios.

**IV. References**


L. Dehaspe, H. Toivonen, R. D. King, “Finding frequent substructures in chemical compounds”. In 4th Intl. Conf. on knowledge Discovery and Data mining, 1998.


P. Bogdanov, “Graph searching, indexing, mining and modeling for Bioinformatics, chemoinformatics and Social network.” 2008.


Z. Zou, J.L, H.Gao, Z.Shang, Mining frequent subgraph patterns from uncertain graph data, IEEE Transactions on Knowledge and Data Engineering 22 (9)(2010)1203–1218.


X. Yan, J. Huan, gSpan: Graph-based substructure pattern mining, in: Proceedings International Conference on Data Mining, Maebashi, Japan, 2002, pp. 721–724.


