A Graph Mining Algorithm for Classifying Chemical Compounds

Winnie W.M. Lam, Keith C. C. Chan
Department of Computing, The Hong Kong Polytechnic University
cswinnie@comp.polyu.edu.hk, cskcchan@comp.polyu.edu.hk

Abstract

Graph data mining algorithms are increasingly applied to biological graph dataset. However, while existing graph mining algorithms can identify frequently occurring sub-graphs, these do not necessarily represent useful patterns. In this paper, we propose a novel graph mining algorithm, MIGDAC (Mining Graph DAta for Classification), that applies graph theory and an interestingness measure to discover interesting sub-graphs which can be both characterized and easily distinguished from other classes. Applying MIGDAC to the discovery of specific patterns of chemical compounds, we first represent each chemical compound as a graph and transform it into a set of hierarchical graphs. This not only represents more information that traditional formats, it also simplifies the complex graph structures. We then apply MIGDAC to extract 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 estimate whether the identified class-specific pattern will positively or negatively characterize a class of drug. Experiments on a drug dataset from the KEGG ligand database show that MIGDAC using hierarchical graph representation greatly improves the accuracy of the traditional frequent graph mining algorithms.

1. Introduction

In recent years, graph mining has increasingly been applied in the area of bioinformatics. In part this is because of the greater availability of biological graph datasets. Graphs are used to represent the complicated structures of chemical compounds, genes interactions and metabolic pathways by using vertices and multiple directed or undirected edges. These structures form patterns, typically frequently-occurring sub-graphs that can be discovered using suitable algorithms and then used in graph classification. Many such graph data mining algorithms have now been developed [1, 2, 3].

Dehaspe et al. proposed an ILP-based mining algorithm called WARMR [2] to search for frequent sub-graphs in databases and used first order predicate logic to represent the input data, but it is not robust enough against noisy or unseen data in real world domains with large databases. Some frequent sub-graphs mining algorithms like FSG [4] and gSpan [5] have been developed to overcome the drawbacks of high complexity. FSG, proposed by Kuramochi et al., adopts an edge-based candidate generation strategy that expands the sub-graph by using the same level-by-level expansion as in the Apriori algorithm [6]. gSpan, proposed by Han et al., discovers frequent sub-graphs based on canonical forms of graphs, and explores frequent patterns by depth-first search (DFS) [7] and visits vertices and marks them with their status.

However, they share the same drawback that they both identify frequent patterns against a given support threshold, which has two disadvantages. First, if the threshold is set too low, many of the patterns that are discovered will not be meaningful and, similarly, setting the threshold too high will lead to meaningful patterns being overlooked. Second, the simple fact that a pattern occurs frequently does not mean that it is sufficiently unique to characterize a class. The discovery of potentially useful patterns thus requires us to consider not just the frequency of subgraphs but also their ability to characterize, or what we might call their degree of uniqueness.

In this paper we propose MIGDAC (Mining Graph Data for Classification), an algorithm for discovering and classifying sets of interesting graph patterns. We first represent each compound in a chemical compound database as an attributed graph and transform it into a set of hierarchical graphs. We then calculate an interestingness measure for each discovered frequent sub-graph and use an interestingness threshold to distinguish between the interesting and the less interesting sub-graphs. The interesting sub-graphs consist of patterns that can uniquely characterize a class. We further define these as class-specific patterns according to their ability to characterize a class and to distinguish a graph sample across multiple classes. These class-specific patterns are then compared with an unseen drug sample by graph matching and finally, after a calculation of the weight of evidence, the unseen sample is classified into a class. The experimental results show that the addition of MIGDAC works well with large biological datasets and greatly increases the classification accuracy of both FSG and gSpan.

The rest of the paper is organized as follows. Section 2 describes our basis for using hierarchical graphs to represent the chemical compound. Section 3

978-0-7695-3452-7/08 $25.00 © 2008 IEEE
DOI 10.1109/BIBM.2008.36

states the graph classification problem which this paper addresses and describes the details of interestingness calculation and weight of evidence. Section 4 describes the results of our experiments classifying a multi-class drug data. Finally, Section 5 summarizes the work and describes possible future work.

2. Using Hierarchical Graphs to Represent Chemical Compounds

In this work, the proposed MIGDAC algorithm is applied to a chemical compound database. Each compound is represented as an attributed graph and then transformed into a set of hierarchical graphs. To build these hierarchical graphs, we use MAGMA [8], Multi-Level Attributed Graph Mining Algorithm, to group components of the attributed graph into different levels according to their attributed structural relations. An attributed graph is the basis of hierarchical graph. We define an Attributed Graph (AG) as an ordered pair $G_a = (V_a, A_a)$ where $V_a = \{v_1, \ldots, v_p, \ldots, v_q, \ldots, v_m\}$ is a set of attributed vertices and $A_a = \{\ldots, a_{pq}, \ldots\}$ is a set of attributed arcs. It provides a means to group components (sub-graphs) of the attributed graph in different levels according to whatever relation has induced the attributed vertex and arc sets. A MAG is defined as an ordered pair $M_a=(X_a, E_a)$ such that $X_a=\{x_1, \ldots, x_p, \ldots, x_q, \ldots, x_m\}$ is a set of attributed vertices with attribute values which are attributed graphs at a lower level; and $E_a=\{\ldots, e_{pq}, \ldots\}$ is the set of arcs connecting those vertices.

3. Classification Using Interestingness and Weight of Evidence

MIGDAC operates in two steps. First it uses an interestingness measure and threshold to discover a set of class-specific patterns and then uses them to classify unseen samples by calculating and comparing a weight of evidence measure as we now describe.

The graph classification problem which this paper addresses is as follows. Given a chemical compound database consisting of n drug samples represented as graphs, $G_1(V_1, E_1)$, $G_2(V_2, E_2)$, …, $G_n(V_n, E_n)$, where each graph, $G_i(V_i, E_i)$, $i \in \{1, \ldots, n\}$ is an attributed relational graph with a vertex set, $V_i = \{v_{i1}, v_{i2}, \ldots, v_{in}\}$ and edge set $E_i = \{e_{i1}, e_{i2}, \ldots, e_{im}\}$ where v_{ij}, $j = 1, \ldots, n_i$, are values in domain(A_{ij}), $j = 1, \ldots, n_i$, of attributes A_{ij}, $j = 1, \ldots, n_i$, respectively and e_{ij}, $j = 1, \ldots, m_i$, which connects two vertices $v_{ik}, v_{jl} \in V_i$ and $v_{ik} \neq v_{jl}$, represents the relationship, $R_{ij}(v_{ik},v_{jl})$, between two attribute values v_{ik} and v_{jl}. Given also that these n graphs are pre-classified into p classes, the graph classification problem is concerned with the discovery of graph patterns to allow graphs that are not originally among $G_1, G_2, \ldots, G_n$ to be correctly classified into one of the p classes.

MIGDAC addresses this problem by using residual analysis [9] to improve the accuracy of graph classification. Residual analysis can identify interesting associations between attributes by using an objective interestingness threshold and measure. It makes use of a contingency table of events and of statistically significant events in the dataset. Patterns that do not provide significant assistance in differentiating between samples in different classes are treated as outliers and, to speed up the mining process, can be omitted.

Given a class of graph data samples, G_c, each of which consists of a set of sub-graphs $M_j = (m_{j1}, m_{j2}, \ldots m_{jn})$, MIGDAC first gets a set of extracted sub-graphs and determines the frequency count of m_{j1} to m_{jn} occurrences in each class. Equation 2 is then used to calculate the interestingness measure $D_j = (d_{j1}, d_{j2}, \ldots d_{jn})$ for each m_{jk} in M_j. As not all sub-graphs qualify as class-specific patterns, that is, a pattern that is interesting enough to characterize as a class, MIGDAC filters out less interesting sub-graphs. MIGDAC determines whether a sub-graph m_{jk} is a class-specific pattern corresponding to a specified class against an interestingness threshold T. If the maximum value of d_{jk} in all classes is higher than T, m_{jk} qualifies as a class-specific pattern that has a degree of interestingness characterized by the interestingness measure d. The value of T is calculated based on the confidence interval, in this paper, we set the confidence probability at 95% (i.e. T = 1.96).

In the presence of uncertainty, a class-specific pattern can be regarded as providing useful information for determining whether the graph sample it characterizes should be assigned to a class, C_p, if Pr(graph sample is in $C_p \mid M_j = m_{jk}$) is significantly different from Pr(graph sample is in C_p). We regard m_{jk} as a class-specific pattern with association between C_p.

The interestingness measure d can be objectively evaluated and is defined in Equation (2):

$$d_{C_p m_{jk}} = \frac{z_{C_p m_{jk}}}{\sqrt{\gamma_{C_p m_{jk}}}}$$

where $z_{C_p M_{jk}}$ is a standardized difference given by Equation (3):

$$z_{C_p m_{jk}} = \frac{count_{C_p m_{jk}} - e_{C_p m_{jk}}}{\sqrt{e_{C_p m_{jk}}}}$$

where $e_{C_p M_{jk}}$ is the number of graph samples expected to contain C_p and M_{jk} calculated by Equation (4):

$$e_{C_p m_{jk}} = \frac{\sum_{i=1}^{s_k} count_{C_p m_{ji}} \sum_{i=1}^{s_p} count_{C_i m_{jk}}}{T}$$

where $T = \sum_{p=1}^{S_p} \sum_{i=1}^{S_i} count_{C_p M_{ji}}$ and $\gamma_{C_p m_{jk}}$ is the maximum likelihood estimate of the variance of $z_{C_p m_{jk}}$ and is given by Equation (5):

$$\gamma_{C_p m_{jk}} = (1 - \frac{\sum_{i=1}^{s_j} count_{C_p m_{ji}}}{T})(1 - \frac{\sum_{i=1}^{s_p} count_{C_i m_{jk}}}{T})$$

If $d_{C_p m_{jk}} > T$, we can conclude that the discrepancy between Pr(class = C_p | $\underline{M}_j$ = $\underline{m}_{jk}$) and Pr(class = C_p) is significant and therefore the association between $\underline{m}_{jk}$ and C_p is interesting and useful for classification. If $d_{C_p m_{jk}} > +T$, it implies that the presence of $\underline{m}_{jk}$ in C_p is significant or, in other words, that the sub-graph m_{jk} is the class-specific pattern of class C_p. If $d_{C_p m_{jk}} < -T$, it implies that the absence of $\underline{m}_{jk}$ in C_p is significant, and we can say that $\underline{m}_{jk}$ is negatively associated with C_p.

When we apply this concept to graph classification, $\underline{m}_{jk}$ refers to an extracted M that corresponds to a class-specific pattern in a class and C_p refers to the class that it belongs to. If the value of $d_{C_p m_{jk}} > +T$, we can conclude that the m_{jk} is a positive class-specific pattern that is useful in characterizing its class (class label is C_p) as highly unique. If the value of $d_{C_p m_{jk}} < -T$, it means the m_{jk} is a negative class-specific pattern that is useful in characterizing its class (class label is C_p). If the value of $d_{C_p m_{jk}}$ is 0, we will regard m_{jk} as having no discriminative power at all. In some cases, the same class-specific pattern may occur in different classes at the same time. The interestingness measure acts as a weight to show the level of importance of the class-specific pattern in different classes.

After discovering a set of class-specific patterns, we can use them to classify unseen samples by calculating and comparing the weight of evidence measure W. The weight of evidence provided by m_{jk} for or against the assignment of the unseen sample characterized by m_{jk} into class c_p can be defined as in Equation (7):

$$W(Class = c_p / Class \neq c_p \mid m_{jk}) = I(Class = c_p : m_{jk}) - I(Class \neq c_p : m_{jk})$$

The class-specific patterns, extracted by MIGDAC, are then matched against the unseen graph by graph matching. This also determines the value of W of the significant sub-graphs. The class producing the greatest W when the graph is assigned is the class to which it should be assigned. W can be interpreted as a measure of the difference in the gain in information when the sample containing m_{jk} is assigned to c_p compared with when it is assigned to other classes. By comparing W_p of each class c_p, the unseen sample is predicted as belonging to class c_p if the value of W_p is the largest.

4. Experiments and Results

We applied our classification algorithm on two benchmark graph mining algorithms: FSG and gSpan. Their executable files can be obtained from [10, 11] respectively and LIBSVM [12] is chosen as the classification model. We selected three classes of drug: Benzodiazepins, Phenothiazines, and Antivirals from the KEGG ligand database.

A chemical compound is a collection of atoms connected by covalent bonds. The atoms and bonds can be represented using a labeled graph in which all atoms are represented by attributed vertices and all bonds are represented by attributed edges. The same atoms in chemical compounds are distinguished by different labels as they represent different physiochemical properties in accordance with their spatial and chemical situations. Each atom in the compound is represented by a vertex, and each bond is represented by an edge. We use FSG and gSpan to extract sub-graphs occurring with a frequency above a given support threshold (σ). If the classification accuracy is low, the value of σ is decreased in decrements of 10% until the maximum accuracy is achieved.

Table 1 shows the classification accuracy of FSG with and without MIGDAC at σ values of 80%, 70% and 60%. At σ = 80%, the accuracy of FSG and gSpan are below 50% because the discovered frequent sub-graphs are not useful in characterization. At σ = 70%, the classification accuracies are nearly the same, so we further decrease the threshold by 10%. At σ = 60%, the improvements in accuracy associated with the MIGDAC-supplemented algorithms are very great, whereas FSG and gSpan have only improved a little. This is because MIGDAC can filter out noisy patterns from the frequent sub-graphs discovered by FSG and gSpan. Although more sub-graphs can be discovered at a lower support threshold, this will capture meaningless as well as useful patterns. MIGDAC uses residue analysis to extract the class-specific patterns and outperform FSG and gSpan by over 55%.

Table 1 Classification accuracies

	Average accuracy (5-fold validation)		
Support σ	σ=80%	σ=70%	σ=60%
FSG	42%	41%	50%
gSpan	40%	42%	48%
FSG+MIGDAC	40%	40%	78%
gSpan+MIGDAC	42%	42%	75%

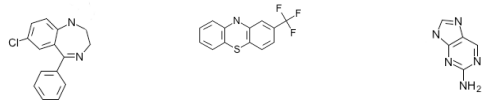
To speed up the classification process, we further introduced the hierarchical graph representation to group related atoms and bonds into a set of components. For example, the six carbon atoms in a benzene ring are represented as a cycle-6 component, and a level-1 hierarchical graph, MAG_1.

In the first stage, the components are represented by the degree of connection. The extracted components are cycle, star and linkage with its degree of connection, for example, a benzene ring is belonging to Cycle-6. The interestingness measure of each component in each class is then calculated.

Table 2 shows interestingness measures for components of three classes of drug. Some components in a class occur and some do not. For example, Cycle-7 is a positive class-specific component in class 1, and is not likely to occur in the other classes, especially not in Class 3, seeing as d_1 of Cycle-7 is greater than $+T$ and d_3 of Cycle-7 is less than $-T$. It is possible to form larger sub-graphs by combining class-specific components with other adjacency components. An interesting pattern is selected from each class and they are shown in Figure 1. After applying hierarchical graph representation, the average accuracy of MIGDAC with hierarchical graph representation is 77% at σ = 60%. This shows that applying MIGDAC with hierarchical graph representation can simplify the sub-graph discovery process and at the same time retain high classification accuracy.

Table 2. Interestingness measure

d	Class 1	Class 2	Class 3
Cycle-5 C_2N_3	0.98	**-3.34**	**2.20**
Cycle-6 C_6	1.22	0.46	-1.68
Cycle-7 C_5N_2	**3.96**	-1.13	**-2.96**
Star-3 $C(C_3)$	**-2.66**	-0.92	**3.59**
Star-4 $C(CF_3)$	**-3.81**	**2.71**	1.24



(a) Class 1 Benzodiazepins (b) Class 2 Phenothiazines (c) Class 3 Antiviral

Figure 1 The interesting patterns

5. Conclusion

In this paper, we have introduced MIGDAC, a novel graph mining algorithm that supports the discovery of useful patterns in graph databases. We applied MIGDAC on chemical compound data to discover patterns that could be used to characterize different kinds of drug. We applied MIGDAC on the existing graph mining algorithms (FSG and gSpan). The experimental results show that MIGDAC improves the classification accuracies by over 55%. MIGDAC discovers sets of class-specific patterns which are statistically frequent enough to uniquely represent a class. Our algorithm offers four benefits over the other graph mining algorithms. First, it uses hierarchical graphs to represent the graph samples. This allows the representation of more structural information in a way that is nonetheless simple. Second, the graph patterns that it discovers are class-specific, resulting in higher classification accuracy. Third, the use of class-specific patterns reduces the number of potential interesting patterns and so speeds up the graph classification process. Fourth, to identify patterns that are distinguishable between classes it uses weight of evidence rather than frequency. This obviates the difficulty where frequent sub-graphs may characterize a class but nonetheless be of no value in distinguishing between various classes of a graph sample. In future work, we would like to test the adaptability of MIGDAC with hierarchical graph representation by applying it to a wider variety of datasets.

6. References

[1] Y. Yoshida, Y. Ohta, K. Kobayashi, N. Yugami, "Mining Interesting Patterns Using Estimated Frequencies from Subpatterns and Superpatterns", Lecture Notes in Computer Science, Vol 2843, 2003, pp. 494-501.

[2] R. D. King, A. Srinivasan, and L. Dehaspe, "Warmr: a data mining tool for chemical data", Journal of Computer-Aided Molecular Design, 2001, 15(2), pp. 173-181.

[3] Christian Borgelt, Michael R. Berthold, "Mining Molecular Fragments: Finding Relevant Substructures of Molecules", Second IEEE International Conference on Data Mining ICDM, 2002, pp. 51.

[4] Michihiro Kuramochi, George Karypis, "Frequent Sub-graph Discovery", icdm, First IEEE International Conference on Data Mining (ICDM'01), 2001, pp. 313.

[5] Xifeng Yan, Jiawei Han, "gSpan: Graph-based substructure pattern mining", Proceedings of IEEE International Conference on Data Mining ICDM, 2002, pp. 721-724.

[6] A. Inokuchi, T. Washio, H. Motoda, An Apriori-based Algorithm for Mining Frequent Substructures from Graph Data, Proc. of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2000.

[7] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, "Depth-first search", Introduction to Algorithms, Second Edition, MIT Press and McGraw-Hill, 2001, pp. 540-549.

[8] Winnie W. M. Lam, Keith C. C. Chan, David K. Y. Chiu, Andrew K. C. Wong, "MAGMA: An Algorithm for Mining Multi-level Patterns in Genomic Data", Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, 2007, pp. 89-94.

[9] K.C.C. Chan and A.K.C. Wong, "A Statistical Technique for Extracting Classificatory Knowledge from Databases," Knowledge Discovery in Databases, G. Piatetsky-Shapiro and W.J. Frawley, eds., Cambridge, Mass.: AAAI/MIT Press, 1991, pp. 107-123.

[10] FSG, http://www-users.cs.umn.edu/~karypis/pafi/

[11] gSpan, http://illimine.cs.uiuc.edu/download/index.php

[12] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines", 2001.