Journal of Computer Aided Chemistry, Vol.2, 87-92(2001) **ISSN 1345-8647**

Applying the Apriori-based Graph Mining Method to Mutagenesis Data Analysis

Akihiro Inokuchi[a], Takashi Washio[a], Takashi Okada[b], and Hiroshi Motoda[a *]

[a] Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
[b] Center for Information & Media Studies, Kwansei Gakuin University,
1-1-155, Uegahara, Nishinomiya, Hyogo, 662-8501 Japan

(Received September 29, 2001; Accepted October 29, 2001)

The Apriori-based graph mining method is an extension of the Apriori algorithm for association rule mining. It constructs a lattice of graph nodes, in which a node at the *k*-th level of the lattice has *k* vertices and the number of supporting instances exceeds a user-specified minimum support. The method can devise a rule "IF subgraph G_a is in transaction G, the union of subgraphs $G_a \cup G_b$ is also contained in G with a ceratin confidence level". When we give a transaction consisting of a chemical graph and virtual vertices expressing molecular properties, we can obtain rules representing structure activity relationships. The method was used to analyze mutagenicity data for 230 aromatic nitro compounds. Several interesting substructures were found to affect the mutagenicity.

Key Words: Apriori-based graph mining, association rule, mutagenesis, aromatic nitro compounds, SAR

1. Introduction

Data with a graph structure appear in many practical fields, such as the molecular structure of chemical compounds and information flow patterns on the Internet. We investigated algorithms for mining frequently occurring subgraph patterns from graph-structured data. Our recent study introduced algebraic graph theory to the framework of Basket Analysis. The method is called Apriori-based graph mining (AGM).[1] It extends the conventional Apriori algorithm,[2] and can efficiently mine a complete set of frequent subgraphs from a general class of graph structures. Graphs can be either

* *motoda@ar.sanken.oosaka-u.ac.jp*

directed or undirected, and can have loops including self-loops; the vertices and edges can have labels, e.g., C (carbon) and N (nitrogen), or single and aromatic bonds in chemical compounds. Furthermore, it can mine unconnected subgraph patterns.

The KDD Challenge 2000 Workshop was held to bring together researchers and practitioners interested in discovering knowledge from real-world databases.[3] One of the target datasets was the mutagenesis activity of 230 aromatic and heteroaromatic nitro compounds, compiled by Debnath et al.[4] Our algorithm was applied to this dataset in order to obtain rules of value to the investigation of the mutagenicity of chemical compounds. Many association rules with meaningful confidence were discovered, identifying characteristic substructures with either higher or lower mutagenesis activity.[5] We inspect the resulting rules referring the structures of supporting compounds in this paper.

2. Apriori-based Graph Mining

The method is intended to find all the association rules from a database of graphs, satisfying user-specified *minimum support* and *minimum confidence* thresholds. A rule takes the following form:

$$G_a \Rightarrow G_b \quad .$$

Here, G_a and G_b represent a graph. The rule means, "If transaction graph G contains G_a as a subgraph, G also contains G_b as a subgraph". The occurrence of the union of the graphs, $G_a \cup G_b$, in the database is called the support of the rule. The ratio of the occurrences of $G_a \cup G_b$ to G_a is called the confidence of the rule. This method mines all rules whose support and confidence values exceed threshold values.

This method constructs a lattice of frequent graphs to obtain association rules. The details of the algorithm are in our original paper.[1] Here, we briefly introduce the method.

2.1 Representing graph-structured data

In the framework of this paper, one graph in a database constitutes one transaction. We employ an adjacency matrix representation of a graph. The vertex that corresponds to the i-th row (the i-th column) is called the i-th vertex v_i, and the number of vertices contained in a graph is its *size*. Let an adjacency matrix of a graph whose size is k be X_k, the ij-element of X_k, x_{ij} and its graph, $G(X_k)$. The vertex labels are defined as N_p (p = 1,…, α) and the edge labels as L_q (q = 1,…, β). Vertex and edge labels are indexed using natural numbers for computational efficiency.

Let the set of vertices of G be $V(G)$, and the set of edges of G be $E(G)$. An induced subgraph G' of G is defined as follows.

$$V(G') \subset V(G) \quad ,$$
$$E(G') \subset E(G) \quad ,$$
$$\forall u,v \in V(G') \ \{u,v\} \in E\{G\} \Leftrightarrow \{u,v\} \in E\{G'\} \quad ,$$

where $\{u, v\}$ represents an edge connecting the vertices u and v.

Based on this definition, the *support* of an induced subgraph $G_a \cup G_b$ in the database and the *confidence* of an association rule $G_a \Rightarrow G_b$ are defined as follows.

$$sup(G_a \cup G_b) = \frac{cnt(G_a \cup G_b)}{cnt} \quad ,$$
$$conf(G_a \Rightarrow G_b) = \frac{sup(G_a \cup G_b)}{sup(G_a)} \quad ,$$

where cnt is the total number of transaction graphs and $cnt(G_a \cup G_b)$ is that including $G_a \cup G_b$ as an induced subgraph.

Our algorithm generates association rules with *support* and *confidence* exceeding user-specified *minimum support* and *minimum confidence*, respectively. A graph whose frequency exceeds the *minimum support* is called a "*frequent graph*".

2.2 Candidate generation of frequent graphs

The adjacency matrix of a graph is defined as follows:

$$X_k = \begin{pmatrix} 0 & x_{1,2} & x_{1,3} & \cdots & x_{1,k} \\ x_{2,1} & 0 & x_{2,3} & \cdots & x_{2,k} \\ x_{3,1} & x_{3,2} & 0 & \cdots & x_{3,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k,1} & x_{k,2} & x_{k,3} & \cdots & 0 \end{pmatrix} \quad .$$

where an element x_{ij} denotes the edge label between vertices i and j.

Two frequent graphs of size k are joined in order to generate a candidate frequent graph of size k+1. Let X_k and Y_k be adjacency matrices of two frequent graphs $G(X_k)$ and $G(Y_k)$ of size k. If $G(X_k)$ and $G(Y_k)$ share equal matrix elements except for the elements in the k-th row and the k-th column, then they are joined to generate a candidate graph Z_{k+1}.

$$X_k = \begin{pmatrix} X_{k-1} & x_1 \\ x_2^T & 0 \end{pmatrix} ,$$

$$Y_k = \begin{pmatrix} X_{k-1} & y_1 \\ y_2^T & 0 \end{pmatrix} ,$$

$$Z_{k+1} = \begin{pmatrix} X_{k-1} & x_1 & y_1 \\ x_2^T & 0 & z_{k,k+1} \\ y_2^T & z_{k+1,k} & 0 \end{pmatrix} ,$$

where X_{k-1} is the adjacency matrix representing the graph of size k-1; x_i and y_i (i = 1,2) are (k-1)×1 column vectors.

Here, the (k, k+1) and the (k+1, k) elements of the adjacency matrix Z_{k+1} are not determined by X_k and Y_k. For an undirected graph, two possible cases are considered: 1) there is an edge labeled L_q between the k-th vertex and the k+1-th vertex of $G(Z_{k+1})$ or 2) there is no edge between them. Accordingly, we must generate β+1 adjacency matrices whose (k, k+1)-element and (k+1, k)-element are one of "0" and "L_q"s. In case of undirected graphs, the number of necessary Z_{k+1}'s is $(\beta+1)^2$.

Graph G of size k+1 can be a candidate frequent graph only when the adjacency matrices of all the induced subgraphs of size k are confirmed to be frequent graphs. Conversely, if one of the induced subgraphs of $G(Z_{k+1})$ is not a frequent graph, Z_{k+1} cannot be a candidate frequent graph. This is because any induced subgraph of a frequent graph must be a frequent graph, due to the monotonicity of the support values in the lattice.

2.3 Lattice construction

Construction of the lattice follows the Apriori algorithm in association rule mining. It starts from nodes with single vertex graphs at the top level of the lattice. The frequencies of the supporting graphs in the database are counted, and candidate graph nodes at the next level of the lattice are generated from the frequent graphs. The procedure is repeated until no new candidate graphs appear.

3. Results and Discussion

3.1 Computation procedure

Debnath[4] originally compiled the mutagenesis data used in this work. It contains 230 aromatic nitro compounds. An SDF format file for these compounds was provided for KDD Challenge 2000 and it is attached in a separate paper.[6] A record contains a chemical graph as well as activity, LogP, and LUMO values.

Association rule mining cannot handle numerical attributes. Activity value is categorized as *inactive*, *low*, *medium*, and *high* using the threshold values (-90.0, 0.0, 3.0) suggested on the KDD Challenge 2001 web page. The respective percentage of transactions with *high*, *medium*, *low*, and *inactive* classes is 15.2, 45.7, 29.5, and 9.6%.

LogP and LUMO are categorized by an AIC (Akaike Information Criterion)-based method that we proposed.[7] The method discretizes the numeric features so that the *AIC* of the following equation is minimized in a greedy manner.

$$AIC = 2\sum_r n(r)Ent(r) + 2m \quad ,$$

where $n(r)$ is the number of transactions located in discretized region r of the feature space, $Ent(r)$ is the information entropy of the data in r, and m is the total number of cut points. This method produces two threshold values: LogP = 3.3 and LUMO = -1.834.

These categories for a chemical compound are added to the transaction in the form of isolated vertices, as shown in Fig. 1. Furthermore, we add artificial edges between a pair of vertices when the number of intervening edges between the vertices is between 2 and 6. These artificial edges are introduced because they decrease the computation time, and they are also useful in recognizing subgraphs that include unspecified vertices and edges.

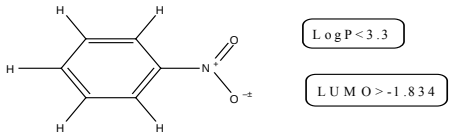


Fig. 1 A sample transaction.

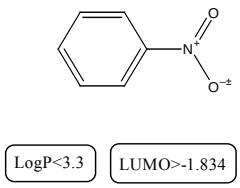
The algorithm explained above was implemented, and association rules were derived from the mutagenesis dataset. In this application, there were 64,973 frequent graphs derived for a *minimum support* of 20%.

We selected rules, $G_a \Rightarrow G_b$, that contained the activity feature in G_b. Some of the rules contained trivial graphs that consisted of isolated vertices. Others showed no meaningful changes in the activity distribution. We selected rules indicating lower and higher activity with meaningful substructures. They are discussed below.

3.2 Lower activity patterns

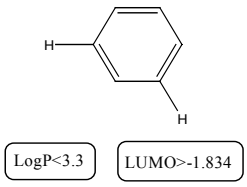
Two patterns leading to lower activity are depicted in Fig. 2. In Fig. 2(a), the support of a frequent subgraph consisting of a nitrobenzene substructure, *low-LogP*, and

high-LUMO is 33.5%, and it is employed as G_a. Adding the *low-activity* vertex to this graph, we get another frequent subgraph, $G_a \cup G_b$, with 20.0% support. These two frequent subgraphs give us an association rule with 59.7% confidence. Since it is interesting to know the entire activity distribution for G_a, we counted the number of supporting compounds for the graphs with *high*, *medium*, *low*, and *inactive* activity. The results are shown in the table to the right in Fig. 2 (a).



activity	support	confidence
High	0.0%	0.0%
Medium	5.7%	16.9%
Low	20.0%	59.7%
Inactive	7.8%	23.4%
Sum	33.5%	100.0%

(a)



activity	support	confidence
High	0.0%	0.0%
Medium	6.1%	17.9%
Low	20.0%	59.0%
Inactive	7.8%	23.1%
Sum	33.9%	100.0%

(b)

Fig. 2 Lower activity patterns.

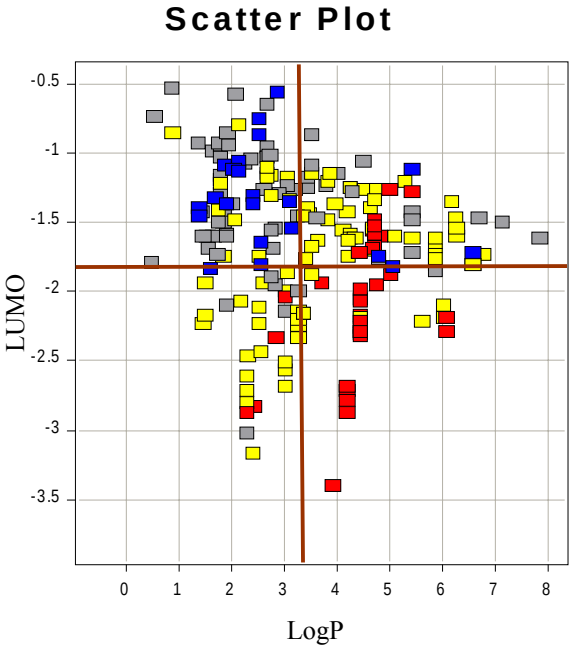
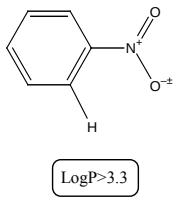


Fig. 3 Scatterplot of activity for nitrobenzene derivatives using LUMO and LogP. : inactive, : low, : medium, : high. Brown lines show categorization thresholds.

The distribution indicated by the confidence values clearly shifts to lower activity than for the 230 compounds. The substructure and features shown in Fig. 2 (b) indicate a similar tendency. The features in Fig. 2 (a) and (b) specify 77 and 78 compounds, respectively, of which 76 are shared. Eighty-five compounds have a *low LogP* and *high LUMO*, and the substructural features excluded molecules consisting of 5-membered rings. Therefore, the results in Fig. 2 indicate the lower mutagenicity of nitrobenzene derivatives, when they have *low LogP* and *high LUMO*. Figure 3 is a scatterplot of the activity levels for 213 nitrobenzene derivatives, using LogP and LUMO as the x- and y-axes, respectively. We can see that the categorization using the AIC-based method is quite functional.

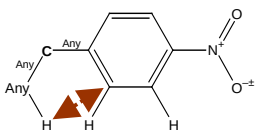
3.3 Higher activity patterns

Two frequent graphs shown in Fig. 4 indicate substructures leading to higher activities. The symbol "Any" for an atom or a bond shows that its label is arbitrary. The meaning of Fig. 4 (a) is well understood referencing Fig. 2 (a). We can see that the *high-LogP* feature leads to higher mutagenicity, when we look at the scattergram in Fig. 3. The LUMO feature is absent from this rule because there are too few supporting compounds in the lower right area of the scattergram. Another interesting difference is the presence of an *ortho* hydrogen atom. There are 94 supporting compounds in this pattern, but the number increases to 104 if we omit the *ortho* hydrogen condition. The 10 compounds without an *ortho*-H had differing activity: 3 inactive, 4 low, and 3 medium. Two typical structures are illustrated in Fig. 5, and we hypothesize that the steric hindrance to the coplanarity of the benzene ring and NO_2 may decrease the mutagenicity of a molecule. This hypothesis



activity	support	confidence
High	13.5%	33.0%
Medium	20.4%	54.3%
Low	4.8%	11.7%
Inactive	0.4%	1.0%
Sum	40.9%	100.0%

(a)



activity	support	confidence
high	7.8%	23.1%
Medium	20.4%	60.2%
Low	5.7%	16.7%
Inactive	0.0%	0.0%
Sum	33.9%	100.0%

(b)

Fig. 4 Higher activity patterns.

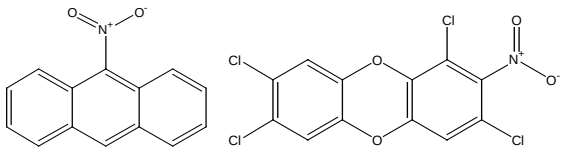


Fig. 5 Typical structures without *ortho* hydrogens.

is the same as that derived in a separate paper by one of the authors.[6]

Figure 4 (b) shows another substructure leading to higher activities. The derived graph contains an artificial edge of length 4 illustrated by an arrow. It is interesting that no isolated vertex for LogP and LUMO appears. However, the supporting compounds consist of various skeletons, and we could not reach any working hypothesis from this pattern.

4. Related work

Propositional classification techniques, e.g., C4.5, and inductive logic programming (ILP) techniques have been applied to carcinogenesis predictions of chemical compounds.[8, 9] However, these approaches can only discover limited characteristic substructures, because the graph structures must be predefined by some specific features or grounded instances of predicates, such as that a benzene ring is involved in the compound. This data preprocessing is inevitably necessary for propositional classification techniques, since they can handle only feature tables. This preprocessing is also necessary for ILP techniques, to reduce the computation time in the mining process. However, our algorithm can directly handle graph structures in general.

Recently, Dehaspe et al. [10] proposed a technique to mine the frequent substructures characterizing the carcinogenic activity of chemical compounds that does not require conversion of substructures to specific features. They used the framework of the ILP combining level-wise search to minimize the access frequency to the database. Since the efficiency achieved by this approach is better than that of previous ILP approaches, the discovery of some substructures characterizing carcinogenesis was expected. However, the full search space was still so large that the search had to be limited to the sixth level, where the substructures consist of a few atoms at maximum, and they reported that no significant substructures were obtained within the search level.

Another analysis of the same data set was done in a separate paper.[6] Both methods employ level-wise search of the lattice. The resolving power is better when we use the cascade model and linear substructure patterns. On the other hand, the readability of a common substructure is better in the current approach.

5. Conclusions

By applying the Apriori-based graph mining method, we obtained many association rules with the subgraph representation. Some had meaningful confidence and led to deeper understanding of the mutagenesis data. The limitation of the current method is that it prevents us from using a low support level for frequent graphs. We will be able to find many reasonable hypotheses in the SAR region if we can obtain frequent graphs with weaker support. Developing the system in this direction is in progress.

Part of this research is supported by Grant-in-Aid for Scientific Research on Priority Areas (B) 13131206, 13131210 and Grant-in-Aid for Scientific Research (B) 12480088.

References and Notes

[1] A. Inokuchi, T. Washio, & H. Motoda, *Principles of Data Mining and Knowledge Discovery* (*PKDD* 2000), 13-23, LNAI 1910, Springer-Verlag, (2000).

[2] R. Agrawal & R. Srikant, *Proc. of the 20th VLDB Conference*, 487-499 (1994).

[3] E. Suzuki, KDD challenge 2000: (URL= http://www.slab.dnj.ynu.ac.jp/challenge2000/).

[4] A.K. Debnath et al., *J. Med. Chem*. **34**, 786-797 (1991).

[5] A. Inokuchi, T. Washio, T. Okada & H. Motoda, *Proc. Int. Workshop KDD Challenge on Real-world Data*, 41-46, PAKDD-2000 (2000).

[6] T. Okada, submitted to *J. Computer Aided Chemistry*, **2**, 79-86 (2001).

[7] A. Inokuchi, T. Washio, & H. Motoda, *Proc. of the 12th Annual Conference of JSAI*, 74-76, (1998) *in Japanese*.

[8] S. Kramer, B. Pfahringer, & C. Helma, *Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining*, 223-226 (1997).

[9] R. King, S. Muggleton, A. Srinivasan, & M. Sternberg, *Proc. Nat. Acad. Sci*., **93**, 438-442 (1996).

[10] L. Dehaspe, H. Toivonen, & R.D. King, *Proc. of the 4th International Conference on Knowledge Discovery and Data Mining*, 30-36, (1998).

アプリオリ型グラフマイニング法による変異原性化合物の解析

猪口明博[a], 鷲尾隆[a], 岡田孝[b], 元田浩[a*]

a 大阪大学産業科学研究所, 〒567-0047 大阪府茨木市美穂ヶ丘8-1
[b] 関西学院大学情報メディア教育センター, 〒662-8501 兵庫県西宮市上ヶ原1-1-155

アプリオリ型グラフマイニング法とは相関ルール探索におけるアプリオリアルゴリズムをグラフ探索に適用するように拡張したものである。その際、k次のラティスレベルには、k個の頂点を有するグラフの中から、与えられた最小サポート値以上のものが置かれる。本方法により、"IF 事例中にサブグラフG_a が存在するならば、THEN サブグラフ$G_a \cup G_b$ もその事例中にある確信度で存在する"というルールを導くことができる。本研究においては、事例を構成するグラフとして、化学構造式および分子の各種性質を表現する仮想的な孤立グラフ頂点を与えることにより、構造活性相関関係を表現するルールを得ることができる。本方法を230種の芳香族ニトロ化合物群における変異原性の解析に適用した結果、有効な作業仮説を得ることができた。本方法は、一般的な構造活性相関研究の方法論として採用できるものである。方法の原理と応用結果について述べる。

キーワード: アプリオリ型グラフマイニング, 相関ルール, 変異原性, 芳香族ニトロ化合物, 構造活性相関

[*] *motoda@ar.sanken.oosaka-u.ac.jp*