Attribute Selection in Chemical Graph Mining Using Correlations among Linear Fragments

Takashi OKADA

Department of Informatics, Kwansei Gakuin University, 2-1 Gakuen, Sanda-shi, Hyogo, 669-1337 Japan
E-mail: okada-office@ksc.kwansei.ac.jp,

Abstract. Data mining often encounters a problem with a huge number of descriptive features. Authors have analyzed structure activity data of dopamine antagonists, where we have to select useful features out of numerous fragments extracted from chemical structures. Correlation coefficients among categorical variables are used to select attributes. Rules obtained by the cascade model were evaluated from chemists' point of view, and the importance of attribute selection was confirmed.

1 Introduction

One of the challenging problems in data mining is to cope with vast amount of attributes. A typical example is to find important genes from millions of single nucleotide polymorphisms (SNPs) that explain the cause of some disease. Authors have analyzed the structure-activity relationships using linear fragments derived from chemical graphs. The number of meaningful fragments was about 2000-3000. Its number is much fewer than that of SNPs problem, but we could not reach valuable knowledge unless we overcame this problem.

Association rule mining is a method that succeeded to solve the numerous attributes problem [1]. It can detect frequent itemsets in customer's baskets selected from thousands of items sold in a supermarket. However, its success depends on the sparseness of the data. That is, the method think of a few items in a basket, and it does not take into account the items that do not appear in the basket. When we treat a dense dataset, there appears a huge number of itemsets resulting in the combinatorial explosion of the itemset lattice. The cascade model developed by the author constructs the itemset lattice, too [2, 3]. It can handle a dense dataset, as it detects a useful rule from a single link located at the shallow level of the lattice. But, the number of attributes is limited to 100-150, and some improvements were necessary in order to treat a dataset with numerous attributes.

In the regression analysis we usually employ attribute selection procedures in order to avoid the overfitting and the instability of the model arising from the collinearity among explanation variables. Attribute selection was also found useful in the decision tree approach, when a dataset contains more than dozens of attributes.

This paper reports an attempt to introduce attribute selection to the mining of SAR from chemical graphs. The next section briefly describes the overview of the analysis, the basic introduction to the mining method as well as the problems encountered. The selection of categorical attributes is done by using correlation coefficients among them, the definition of which is given in Section 3. Results of an application to the chemical graph mining are shown in Section 4, where the effects of attribute selection is discussed referring the quality of rules judged from a chemist's point of view.

2 Mining Chemical Graphs by the Cascade Model

2.1 Overview of the Dopamine D2 Antagonists Analysis

Dopamine is a neurotransmitter in the brain. Neural signals are transmitted via the interaction between dopamine and proteins known as dopamine receptors. There are five different receptor proteins, D1 – D5, each of which has a different biological function. Certain chemicals act as an antagonist to these receptors. An antagonist binds to a receptor, but does not function as a neurotransmitter. Therefore, it blocks the function of the dopamine molecule.

We used the MDDR database developed by Prous Science and MDL as the data source [4]. It contains 1,349 chemical structure records that describe dopamine (D1, D2, D3, and D4) antagonist activities. The sample problem used in this paper is to discover the structural characteristics responsible for D2 antagonist activity, which is known to be the hardest problem among 4 antagonist activities.

Figure 1 shows the brief scheme of the analysis. All structural formulae of chemicals are stored in a SDF file, a common data exchange format used in computer chemistry. Then, the molecular orbital calculation by means of MM-AM1-Geo software is applied to derive three electronic properties: HOMO, LUMO and Dipole. LogP values are calculated by ClogP program to give the hydrophobic property of molecules. On the other hand, we extract many linear fragments contained in chemical graphs, and the presence/absence of these fragments in a molecule is used as other type of attributes. This type of linear fragments was first introduced by Klopman [5], followed by the developments by Okada [6] and Kramer et al [7]. Linear fragments are expressed by constituent elements and bond types like "c3H:c3---C4H-N3", and they are used as attribute names. Current fragment generation algorithm is to be published [8].

Obviously, the number of all possible fragments is too large. Therefore, the length of linear fragments was limited to be shorter than 9, and one of the terminal atoms of a fragment was restricted to be a heteroatom or a carbon constituting a double or triple bond. Then, we got 8041 fragments, which was too many to be analyzed by the current implementation of the cascade model. Therefore, we selected 114 fragments, of which ratio of appearances in compounds was in the range: 15%-85%.

Application of the cascade model to the table generated rules that characterized the structures of D2 antagonists. Resulting rules were interpreted and evaluated by chemists.

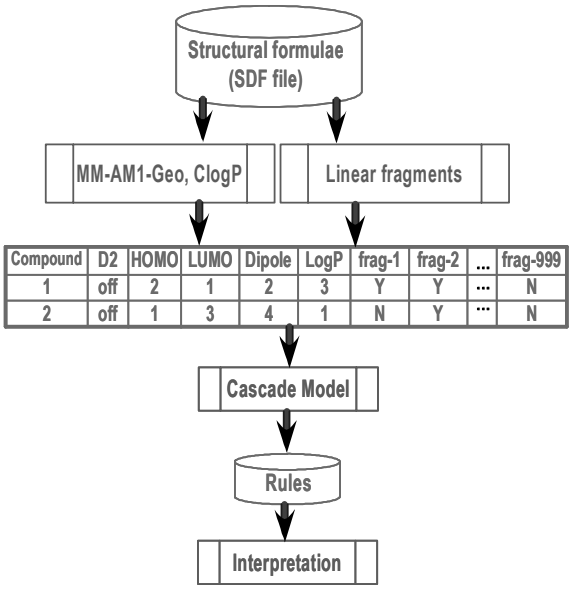

Fig. 1. Flow of chemical graph mining by the cascade model.

2.2 The Cascade Model and the Datascape Survey

The cascade model can be considered an extension of association rule mining [1]. The method creates an itemset lattice in which an [attribute: value] pair is used as an item to constitute itemsets. Links in the lattice are selected and interpreted as rules. That is, we observe the distribution of the RHS (right hand side) attribute values along all links, and if a distinct change in the distribution appears along some link, then we focus on the two terminal nodes of the link. Consider that the itemset at the upper end of a link is {[A: y]} and an item [B: n] is added along the link. If a marked activity change occurs along this link, we can write a rule:

```
Cases: 200 ==> 50 BSS=12.5
IF [B: n] added on [A: y]
THEN [Activity]:    .80 .20  ==>  .30 .70 (y n)
THEN [C]:           .50 .50  ==>  .94 .06 (y n)
```

where the added item [B: n] is the main condition of the rule, and the items at the upper end of the link {[A: y]} are considered preconditions. The main condition changes the ratio of the active compounds from 0.8 to 0.3, while the number of supporting instances decreases from 200 to 50. *BSS* means the between-groups sum of squares, which is derived from the decomposition of the sum of squares for a categorical variable. Its value can be used as a measure of the strength of a rule. The second "THEN" clause indicates that the distribution of the values of attribute [C] also changes sharply with the application of the main condition. This description is called the *collateral correlation*.

Recently, new facilities for *datascape survey* are introduced in order to reduce the number of rules [9], and to denote the details of data distribution specified by a rule [10]. Interpretation of rules became easier by these functions.

2.3 Attribute Selection Problem

There is no reason to justify the selection of 114 fragments appearing in 15%-85% of the compounds. In fact, chemists could notice other important fragments contributing to the D2 activity by browsing structural formulae. However, if we use more attributes, the combinatorial explosion in the lattice size prohibits the analysis. The past experience suggested that the upper limit of the attributes was 100-150.

On the other hand, analysts often encountered a pair of fragments with the same number of supporting compounds like O1=S4-c3:c3H and S4-c3:c3H. The latter support must always be larger or equal to that of the former, since the latter is a substructure of the former. Then, the equality of these supports means that they appear exactly in the same compounds, and the selection of both fragments as attributes is redundant. That is, the correlation coefficient between these two attributes is 1.0.

Omission of an attribute from such pairs is expected to enable the analysis using more attributes with lower supports. Furthermore, attribute pairs do not need to be completely correlated. We can omit an attribute if it is in a highly correlated pair. Therefore, we decided to introduce a correlation coefficient between a pair of attributes, and to use it as a criterion to omit/keep attributes.

3 A Correlation Coefficient between Categorical Variables

Correlation coefficient is a well-known concept in the world of numerical attributes. Recently, we introduced a generalized covariance using a vector expression for the value difference [11], and a uniform treatment of covariance became possible among numerical and categorical variables. Here, we briefly mention a special case to define a correlation coefficient between a pair of binary attributes.

Gini successfully defined the variance of categorical data [12]. He first showed that the following equality holds for the variance of a numerical variable x_i.

$$V_{ii} = \left(\sum_a \left(x_{ia} - \bar{x}_i \right)^2 \right) \Big/ n = \frac{1}{2n^2} \sum_a \sum_b \left(x_{ia} - x_{ib} \right)^2 \quad , \tag{1}$$

where V_{ii} is the variance of the i-th variable, x_{ia} is the value of x_i for the a-th instance, and n is the number of instances.

Then, he introduced the distance definition (2) into value differences of (1), and got the categorical variance expression (3), which is well known as Gini-index.

$$x_{ia} - x_{ib} \begin{cases} = 1 & \text{if } x_{ia} \neq x_{ib} \\ = 0 & \text{if } x_{ia} = x_{ib} \end{cases} , \tag{2}$$

$$V_{ii} = \frac{1}{2n^2} \sum_a \sum_b (x_{ia} - x_{ib})^2 = \frac{1}{2} \left(1 - \sum_r p_i(r)^2 \right) \quad . \tag{3}$$

Extension of the above definition to the covariance fails, if we simply change $(x_{ia} - x_{ib})^2$ to $(x_{ia} - x_{ib})(x_{ja} - x_{jb})$. We employed a regular simplex expression to values of a categorical variable, and used a vector expression, $\overrightarrow{x_{ia}x_{ib}}$, instead of a scalar, $x_{ia} - x_{ib}$, in the variance definition. Our proposal for V_{ij} definition was the sum of inner products of $\overrightarrow{x_{ia}x_{ib}}$ and $\overrightarrow{x_{ja}x_{jb}}$, where two regular simplexes for x_i and x_j are rotated to give the maximum value for V_{ij}. Its definition is given by the subsequent formulae,

$$V_{ij} = \max\left(Q_{ij}(L)\right) \quad , \tag{4}$$

$$Q_{ij}(L) = \frac{1}{2n^2}\sum_a\sum_b \left\langle \overrightarrow{x_{ia}x_{ib}} \middle| L \middle| \overrightarrow{x_{ja}x_{jb}} \right\rangle \quad . \tag{5}$$

Here, $\boldsymbol{L}$ is an orthonormal transformation applicable to the value space. The bracket notation, $\langle \boldsymbol{e}|\boldsymbol{L}|\boldsymbol{f}\rangle$, is evaluated as the scalar product of two vectors $\boldsymbol{e}$ and $\boldsymbol{Lf}$ (or $\boldsymbol{L}^{-1}\boldsymbol{e}$ and $\boldsymbol{f}$). If the lengths of the two vectors, $\boldsymbol{e}$ and $\boldsymbol{f}$, are not equal, zeros are first padded to the vector of the shorter length.

		x_j: u	v	
x_i	r	n_{ru}	n_{rv}	$n_{r\cdot}$
	s	n_{su}	n_{sv}	$n_{s\cdot}$
		$n_{\cdot u}$	$n_{\cdot v}$	n

We apply this definition to the simplest 2 x 2 contingency table shown at the left, where $n_{r\cdot}$ *and* $n_{\cdot u}$ show marginal distributions. Straightforward application of (5) to this table gives the following expressions for V_{ii}, V_{jj} and V_{ij}, and a correlation coefficient R_{ij} is given by (9).

$$V_{ii} = n_{r\cdot}n_{s\cdot}/n^2 = \tfrac{1}{2}\left(1 - (n_{r\cdot}/n)^2 - (n_{s\cdot}/n)^2\right) \quad . \tag{6}$$

$$V_{jj} = n_{\cdot u}n_{\cdot v}/n^2 = \tfrac{1}{2}\left(1 - (n_{\cdot u}/n)^2 - (n_{\cdot v}/n)^2\right) \quad . \tag{7}$$

$$V_{ij} = \frac{\left|n_{ru}n_{sv} - n_{rv}n_{su}\right|}{n^2} \quad . \tag{8}$$

$$R_{ij} = \frac{V_{ij}}{\sqrt{V_{ii}V_{jj}}} \quad . \tag{9}$$

The numerator in (8) is the critical term used to represent the extent of dependency between two variables. In fact, the correlation coefficient is 1.0 (0.0) for completely dependent (independent) data, respectively.

4 Results and Discussion

We applied the attribute selection scheme to the dopamine D2 antagonist problem. All generated fragments reached 8041 kinds. First, we selected a fragment as an attribute, if the probability of its appearance satisfied the following condition,

$$edge < P(\text{fragment}) < 1.0 - edge \quad . \tag{10}$$

When *edge* was set to 0.01, 0.02, 0.03, 0.05, 0.10 and 0.15, numbers of selected fragments were 1698, 1056, 730, 377, 176, and 114, respectively. We employed the presence/absence of these fragments as the initial attribute set $\{x\}$.

4.1 Attribute Selection using Correlation Coefficients

The procedure of the attribute selection is as follows.

1. Calculate correlation coefficients among all attribute pairs: x_i and x_j, and put the pair into a list: *pairs*, when it satisfies the condition: $R_{ij} > min\text{-}R_{ij}$.
2. Sort *pairs* in the descending order of R_{ij}.
3. Pop *pairs,* and get a pair: x_i and x_j.
4. Omit an attribute (x_i or x_j) from $\{x\}$, if both attributes are members of $\{x\}$.
5. Repeat steps 3 and 4 until every pair in *pairs* is examined.

When we omit an attribute from a correlated attribute pair at step 4, a longer fragment is kept in the attribute set. It is because an analyst can get more ideas from the longer attribute name, when it appears in a rule.

Figure 2 shows the numbers of selected attributes in log scale for 6 *edge* values, changing $min\text{-}R_{ij}$ value to 1.0, 0.99, 0.97, 0.95, 0.90, 0.85, 0.80, 0.75, and 0.70. Here, no attribute selection is carried out at $min\text{-}R_{ij}$ = 1.0, and attributes in perfectly correlated pairs are omitted at $min\text{-}R_{ij}$ = 0.99.

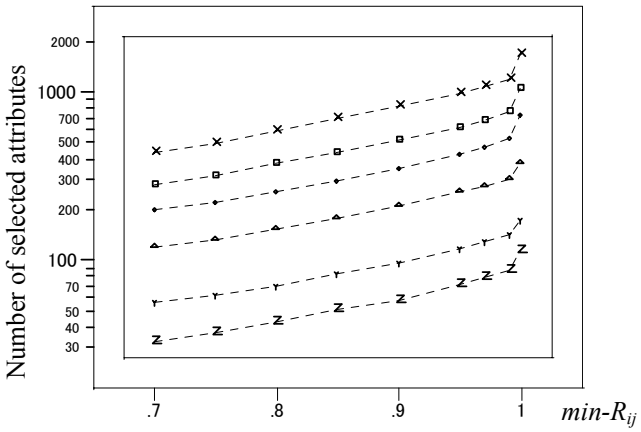

Fig. 2. Numbers of selected attributes changing $min\text{-}R_{ii}$ value.

Shapes of plots in the figure do not depend on *edge* values. Another interesting point is steep slopes found at the right end of the plots. It means about 20-30% of attributes in the initial attribute sets are completely correlated in the chemical graph mining using linear fragments. Roughly speaking, about half attributes are omitted at $min\text{-}R_{ij}$ = 0.90. Therefore, we can conclude that the attribute selection scheme using correlation coefficients works well in reducing the number of attributes.

4.2 Effects to Lattice Size

Lattice expansion in the cascade model is controlled by a parameter, *thres*. The smaller *thres* value we use, the larger number of nodes in the lattice we examine. Ordinary values of *thres* in this application have been in the range: 0.15-0.2, and the

number of nodes in the lattice was in the range: 5,000 – 30,000 using 100-150 attributes. We examined the lattice size changing *edge* and *min-*R_{ij} values, for three *thres* values: 0.15, 0.175 and 0.20.

Figure 3 shows rough contour maps of the number of nodes (*#nodes*) in the lattice, where y-axis is *min-*R_{ij} and x-axes are number of selected attributes (*#attributes*) in (A) and *edge* in (B), respectively. The calculated points are shown by '+' in the figure, but the points resulted in the combinatorial explosion of the lattice are not depicted. The lowest contour line (*#nodes* = 3000) is indicated by arrows.

Contour lines in (A) are all more or less parallel to y-axis in large *min-*R_{ij} values, but they trailed to the bottom right corner in small *min-*R_{ij} values. This fact shows that the lattice size does not arise sharply when we use more uncorrelated attributes. In fact, we could use 400-500 attributes selected from more than 1000 attributes.

The contour lines in Figure 3 (B) are drawn from the upper right to the bottom left corners. The meaning of this fact can be seen by the inspection of two ⟠ points and two + points near the gray contour in the top right map. The data for these four points are summarized in Table 1.

Table 1. Calaculated results for 4 points near a gray contour (*thres*=0.15)

Point	*edge*	*min-*R_{ij}	*#attributes*	*#nodes*	*#detected*	*#rules*	*score*
P1	0.02	0.70	287	4992	23	6 (3)	3
P2	0.05	0.80	155	5983	39	8 (4)	3
P3	0.10	0.90	130	5223	72	9 (4)	2
P4	0.15	0.99	88	6265	97	14 (5)	2

This table shows that similar number of nodes emerge from a wide range of *#attributes* (88 – 287). That is, a correlated attributes set grows the lattice size, while an uncorrelated set depresses it. As a result, the attribute selection using a lower value of *min-*R_{ij} proved to be very useful in reducing the lattice size.

4.3 Evaluation of Rules

A matter of great importance is not the number of selected attributes nor the lattice size, but the quality of rules. *#detected* column in Table 1 shows the number of detected links with large *BSS* values, where the optimization to a rule starts. *#rules* column denotes the number of resulting rules. Also shown in parentheses is the number of principal rules after the rules organization step [9]. These numbers tend to increase as we use higher *min-*R_{ij} values. Appearance of many rules does not always lead to good knowledge discovery. For example, there exist many highly correlated attributes in the calculation at P4, which might be a cause of redundant rules. In fact, the increase in the number of principal rules is very limited.

Here, we introduce an evaluation scheme of rules. Analysts noticed three important substructures relevant to the D2 antagonist activity, browsing various rules. They consisted of an aromatic ether, a tertiary amine separated from an aromatic ring by 3 single bonds, and a CO group bonded to an amine. The appearance of these features in rules was used to judge their quality. That is, we search three features in the main condition of principal rules, and the numbers of the found features were employed as the *score* of a rule set. When a feature appears only in a relative rule, we count it as

0.5. Note that the appearance of a feature is counted only once. Therefore, the highest *score* of resulting rules is 3. This evaluation scheme is rough, as the true mechanism for the appearance of D2 antagonist activity is not known yet. But we can expect that this *score* will be a guide to judge the quality of rule sets.

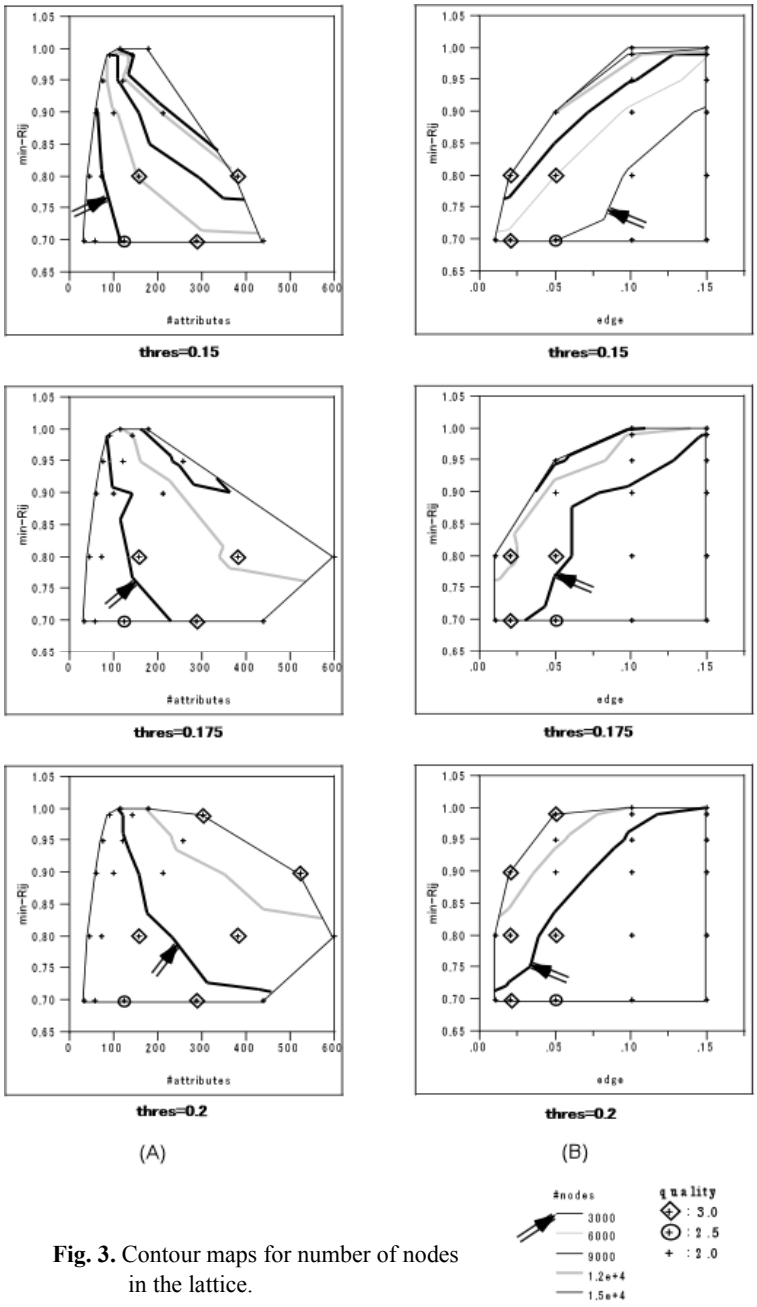

Fig. 3. Contour maps for number of nodes in the lattice.

The last column of Table 1 shows this *score* for four calculations. We can notice that the number of rules has no meaning from this viewpoint. The next problem is to find adequate values leading to a good rule set for the three parameters: *edge*, *min*-R_{ij} and *thres*.

The calculated points with *score*=3, 2.5, 2 are shown by ⟡, ⊕ and +, respectively in Figure 3. The distribution of high score points in Figure 3 (A) indicates that neither the number of attributes nor the size of the lattice have direct relationships to the *score* of a rule set. On the other hand, Figure 3 (B) shows that high score rule sets result from calculations at *edge* = 0.02 and 0.05. The attribute selection using *min*-R_{ij} = 0.8 seems to give better rule sets.

Then, the suggested plan of mining is to employ relatively lower *edge* values, followed by the selection of attributes using *min*-$R_{ij} \cong 0.8$. The effect of *thres* value seems to be limited, as far as the objective of mining is to grasp rough characteristics of chemical graphs.

5 Concluding Remarks

The attribute selection scheme introduced in this paper is essentially a method to cope with collinearity among explanation attributes. Lots of researches have been done to solve this problem in the regression analysis. They include various attribute selection schemes, canonical regression method and partial least squares.

Among mining methods for categorical data, reduct concept in the rough set solved this problem clearly [13]. However, its implementation cannot treat thousands of attributes. Another approach from the mining community is the closed itemset concept in the association rule mining [14, 15]. It is used to compute long frequent itemsets fast, and it is also applied in the filtering of rules to omit redundant ones. However, this method is useful only when a pair of attributes correlates completely. Even if a correlation coefficient is larger than 0.99, the method cannot be applied to data with a noise instance.

The cascade model has also encountered the collinearity problem. The method first incorporated collateral correlations in a rule expression. It illustrates attributes with high correlations to the main condition, and helps an analyst to interpret rules greatly [3]. Further, correlated attributes results in the generation of a pair of rules, covers of which overlap considerably to each other. This problem was solved by the organization of rules into a principal rule and its relative rules [9]. The attribute selection introduced in this paper has been shown to be useful in the reduction of the lattice size. Moreover, the omission of a correlated attribute cuts self-evident collateral correlations, and it also reduces the number of relative rules. Therefore, the load of an analyst has been reduced. All these functions work for partially correlated attributes, and it offers a superior framework than those given by the closed itemset.

The comprehensive analysis of ligands for dopamine receptor proteins are now under progress using the proposed system. They include not only discriminations of antagonists, but also those among agonists. Also under investigation are factors that distinguish antagonists and agonists. The results will be a model work in the field of qualitative SAR analysis.

Acknowledgements

The author wishes to thank Dr. Masumi Yamakawa and Dr. Hirotaka Niitsuma of Kwansei Gakuin University for their valuable discussions.

References

1 Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. Proc. VLDB (1994) 487-499
2 Okada, T.: Rule Induction in Cascade Model based on Sum of Squares Decomposition. Principles of Data Mining and Knowledge Discovery (Proc. PKDD'99), LNAI 1704, Springer-Verlag (1999) 468-475
3 Okada, T.: Efficient Detection of Local Interactions in the Cascade Model. In: Terano, T. et al (eds.) Knowledge Discovery and Data Mining PAKDD-2000. LNAI 1805, Springer-Verlag (2000) 193-203
4 MDL Inc.: http://www.mdl.com/products/knowledge/drug_data_report/index.jsp
5 Klopman, G.: Artificial Intelligence Approach to Structure-Activity Studies. J. Amer. Chem. Soc. 106 (1984) 7315-7321
6 Okada, T.: Discovery of Structure Activity Relationships using the Cascade Model: The Mutagenicity of Aromatic Nitro Compounds. J. Computer Aided Chemistry, 2 (2001) 79-86
7 Kramer, S., De Raedt, L., Helma, C.: Molecular feature mining in HIV data. In: Proc. of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-01) (2001) 136-143
8 Okada, T., Yamakawa M., Niitsuma, H.: Spiral Mining using Comprehensible Attributes Generated from Molecular Structures. Submitted to post-congress proceedings of Active Mining 2003
9 Okada, T.: Datascape Survey using the Cascade Model. In: Satoh, K. et al. (eds.) Discovery Science 2002. LNCS 2534, Springer-Verlag (2002) 233-246
10 Okada, T.: Topographical Expression of a Rule for Active Mining. In: Motoda, H. (ed.) Active Mining. IOS Press, (2002) 247-257
11 Okada, T.: A Note on Covariances for Categorical Data. In: Leung, K.S. et al (eds.) Intelligent Data Engineering and Automated Learning - IDEAL 2000. LNCS 1983, Springer-Verlag (2000) 150-157
12 Gini, C.W.: Variability and Mutability, contribution to the study of statistical distributions and relations, Studi Economico-Giuridici della R. Universita de Cagliari (1912). Reviewed in: Light, R.J., Margolin, B.H.: An Analysis of Variance for Categorical Data. J. Amer. Stat. Assoc. 66 (1971) 534-544
13 Pawlak Z.: Rough sets: Theoretical aspects of reasoning about data. Dordrecht: Kluwer (1991)
14 Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient Mining of Association Rules Using Closed Itemset Lattices. Information Systems, 24 (1) (1999) 25-46
15 Zaki, M.J., Hsiao, C.J.: CHARM: An efficient algorithm for closed itemset mining. In: Proc. SDM'02, SIAM (2002) 457-473