Association rule derivation for side effects of medical supplies and its application

H. Shiroyama, Y. Zuo & E. Kita
Graduate School of Information Science, Nagoya University, Japan

Abstract

In drug discovery, it is very important to predict the side effect of the drug accurately. The prediction algorithm of the drug side effect is presented in this study. This algorithm is based on the concept of the structure-activity relationship. Firstly, the drug side effects are gathered from the registration of medical products by using text mining. Next, the chemical structure information of the drug is obtained from the PubChem data base. Then, the association rules between the chemical structure and the side effects are defined. The associate rules are applied to the prediction of the side effect of 10 chemical products.
Keywords: drug, side effect, association rule, PubChem, text mining.

1 Introduction

Several drugs (medicines) have been developed every year. While new drugs are very useful for improving illness and injuries, they sometimes have terrible side effects. Therefore, it is very important for the prediction of the drug side effects in the drug discovery.

A new drug discovery is a very time-consuming process. The drug discovery is mainly composed of four steps; basic study, non-medical study, medical study and approval and production. In the basic study, the potential chemical products are developed. In non-medical study, the effect of the products is confirmed in animal experiment and so on. In medical study, the effect of the products is provided for patients and health persons. Since the side effects of the potential products are confirmed in non-medical and medical studies, the drug discovery needs a long time and enormous cost.

Therefore, some researchers have studied the prediction algorithm of the drug side effect before non-medical and medical studies. Enslein et al. used


WIT Transactions on Biomedicine and Health, Vol 15, © 2011 WIT Press
www.witpress.com, ISSN 1743-3525 (on-line)
doi:10.2495/EHR110381

multi-regression analysis and discriminant analysis for predicting the side effect [1–3]. Moriguchi et al. have developed adaptive least square (ALS) method and Fuzzy ALS method for huge toxicity data discovery [4–6]. Gilles Klopman has developed the system named as"MULTICASE" which is based on the concept of the quantitative structure activity relationship (QSAR) [7, 8].

In this study, the association rules are used for predicting the drug side effect. This algorithm is based on the concept of structure-activity relationship (SAR). Known drug side effects are gathered from the registration of medical products by using the text mining. The chemical structures of drugs are obtained from the PubChem data base. The association rules between the chemical structures and the side effects are defined. The activity of the side effects is evaluated from the association rules. In the numerical example, the present algorithm is applied for predicting six side effects of 10 chemical products.

The remaining part of this paper is organized as follows. The association rule algorithm is shown in section 2. The present algorithm is explained in section 3. In section 4, the algorithm is applied for predicting side effects of 10 chemical products. The conclusions are summarized again in section 5.

2 Association rule

2.1 Definition of association rule

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is introduced for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The itemsets X and Y are called as antecedent and consequent of the rule, respectively.

The support $supp(X)$ of the association rule $X \rightarrow Y$ is defined as the proportion of transactions in the data set which contain the itemsets X and Y.

$$Support = \frac{\sigma(X \cup Y)}{M} \tag{1}$$

where $\sigma(X \cup Y)$ denotes the total number of transactions which contain the itemset X and Y and M the total number of the transactions.

The confidence of the rule is defined as the portion of the transactions containing the itemsets X and Y and the transactions containing the itemset X alone.

$$\text{Confidence} = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{\text{Support}(X \rightarrow Y)}{Support(X)} \tag{2}$$

The association rule is usually described as follows.

Antecedent → Consequent (Support = α,Confidence = β)

2.2 A priori algorithm

Total number of the association rules increases exponentially according to the increase of the transactions and itemsets. It is very time-consuming to calculate the support and the confidence. For this purpose, A priori algorithm was used in this study [9] since it can calculate the support and the confidence in the real-time.

3 Side effect evaluation

3.1 Structure-activity relationship

The present algorithm is based on the concept of Structure-Activity Relationship (SAR).

Structure activity relationship (SAR) is the relationship between the chemical or three-dimensional structure of a molecule and its biological activity. The analysis of SAR enables the determination of the chemical groups responsible for evoking a target biological effect in the organism. This allows modification of the effect or the potency of a bioactive compound (typically a drug) by changing its chemical structure.

This method was refined to build mathematical relationships between the chemical structure and the biological activity, known as quantitative structure-activity relationships (QSAR).

3.2 Side effect information

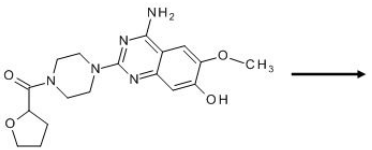
The drug side effect information is gathered from the registration of medical products by using the text mining. In Japan, the drug effect information of 17,000 drugs is distributed as HTLM data by Japan Pharmaceutical Information Center (JPIC). The use of the text mining technique extracts the drug side effect from the HTML data of the registration of medical products.

In this study, we will focus on the side effect for liver, kidney and blood and therefore, gather the information on Aspartate aminotransferase (AST) increase, Alanine aminotransferase (ALT) increase, Blood Urea Nitrogen (BUN) increase, Creatinine (CRE) increase, Red Blood Cell (RBC) decrease, and White Blood Cell (WBC) decrease.

3.3 Chemical structure information

The drug chemical structures are obtained from PubChem database [10] as the description of the simplified molecular input line entry specification (SMILES).

PubChem is a database of chemical molecules and their activities against biological assays. The system is maintained by the National Center for Biotechnology Information (NCBI).



molecular_weight	c
logP	c
mole_refraction_index	b
Aromatic_carbon	y
Aromatic_nitrogen	y
amino_group	y
carbonyl_group	y
carboxyl_group	n
four_halogens	n
hydroxy_group	y
sulfonyl_group	n

Figure 1: Information of chemical structural formula.

The original SMILES specification was developed in the late 1980s [11]. It has since been modified and extended by others, most notably by Daylight Chemical Information Systems Inc.

We focus on the hydrophobic property (ClogP) and the molar refraction (CMR) of the drug chemical structure. Since the hydrophobic property (ClogP) is one of important indexes for the bioactivity and the bioaccumulation of the drugs, it is the essential factor for QSAR. The molar refraction (CMR) is strongly related to the volume of the molecules and to London dispersive forces that has important effect in drug-receptor interaction.

Once the SMILES information of the drug is obtained through PubChem database, the hydrophobic property (ClogP) and the molar refraction (CMR) of the drug is evaluated through the Bio-Loom [12].

3.4 Algorithm

The association rules are defined as follows.

1. The side effects of known drugs are gathered from the registration of medical products by using the text mining.
2. The drug chemical structures are obtained from PubChem database.
3. The association rules are defined from the information of the side effect and chemical structures.
4. The numbers of the antecedent and the consequent of the rules are counted.

The association rules are used for predicting the drug side effect as follows.

1. The itemset of the drug chemical structures is given.
2. The rules conforming the itemset are gathered.
3. The activity evaluation parameter P of the rule set is evaluated.

$$P = \frac{\sum_i^N N2_i}{\sum_i^N N1_i} \tag{3}$$

where $N1_i$ and $N2_i$ denote total number of items in the antecedent and the consequent of the rules, respectively. N the total number of the rules conforming the itemset.

4. It is shown that the side effect with the parameter $P > P_s$ is active. The threshold P_s is specified as $P_s = 0.6$ in the following numerical examples.

Table 1: Association rule.

ID	rule	Ant.	Con.
1	Aromatic_carbon = y carbonyl_group = y → AST=y	33	29
2	Aromatic_carbon = y carbonyl_group = y hydroxy_group = y → AST=y	33	29
3	mole_refraction_index = c carbonyl_group = n hydroxy_group = y → AST=y	38	31
4	molecular_weight = c carbonyl_group = y hydroxy_group = y → AST=y	39	31
5	Aromatic_nitrogen = y carbonyl_group = n → AST=y	40	31
6	Aromatic_carbon = y Aromatic_nitrogen = y carbonyl_group = y → AST=y	40	31
7	molecular_weight = b carboxyl_group = y → AST=y	30	23
8	molecular_weight = b carbonyl_group = n hydroxy_group = y → AST=y	30	23
9	logP=a Aromatic_carbon = y carbonyl_group = n → AST=y	37	28
10	mole_refraction_index = c carbonyl_group = y → AST=y	56	42

A simple example is shown in Fig. 1 and Table 1. As shown in Fig. 1, the chemical structures are obtained from the drug information through the PubChem. The association rules conforming the itemset are listed (Table 1). The association rules with ID = 1, 2, 4 and 6 conform the chemical structure of the unknown drug.

Table 2: Prediction result of drug ID = 1 to 5.

ID	Side effect	Confidence	Prediction	Actual
1	AST increase	0.654952803	active	active
	ALT increase	0.659446587	active	active
	BUN increase	0.549799017	inactive	active
	CRE increase	0.550037249	inactive	active
	RBC decrease	0.565320665	inactive	inactive
	WBC decrease	0.608465608	active	active
2	AST increase	0.591299678	inactive	active
	ALT increase	0.589204945	inactive	active
	BUN increase	0.55704698	inactive	inactive
	CRE increase	0.522154648	inactive	inactive
	RBC decrease	0.583883752	inactive	inactive
	WBC decrease	0.569427527	inactive	active
3	AST increase	0.648256421	active	active
	ALT increase	0.640763463	active	active
	BUN increase	0.55152027	inactive	inactive
	CRE increase	0.541868255	inactive	inactive
	RBC decrease	0.556363636	inactive	inactive
	WBC decrease	0.58747698	inactive	active
4	AST increase	0.624987293	active	active
	ALT increase	0.619428779	active	active
	BUN increase	0.524590164	inactive	active
	CRE increase	0.527687296	inactive	inactive
	RBC decrease	0.566509115	inactive	inactive
	WBC decrease	0.588581024	inactive	inactive
5	AST increase	0.649959724	active	inactive
	ALT increase	0.648337029	active	inactive
	BUN increase	0.55469217	inactive	inactive
	CRE increase	0.541937581	inactive	inactive
	RBC decrease	0.577151335	inactive	inactive
	WBC decrease	0.587978142	inactive	inactive

Therefore, the activity evaluation parameter P is calculated as follows.

$$P = \frac{29 + 29 + 31 + 31}{33 + 33 + 39 + 40} = 0.82 \tag{4}$$

Table 3: Prediction result of drug ID = 6 to 10.

ID	Side effect	Confidence	Prediction	Actual
6	AST increase	0.624987293	active	active
	ALT increase	0.652334657	active	active
	BUN increase	0.547340425	inactive	active
	CRE increase	0.551418981	inactive	active
	RBC decrease	0.565320665	inactive	active
	WBC decrease	0.608465608	active	active
7	AST increase	0.612885386	active	inactive
	ALT increase	0.619428779	active	inactive
	BUN increase	0.524590164	inactive	active
	CRE increase	0.527687296	inactive	inactive
	RBC decrease	0.566509115	inactive	inactive
	WBC decrease	0.588581024	inactive	inactive
8	AST increase	0.630815473	active	active
	ALT increase	0.632175861	active	active
	BUN increase	0.607453416	active	active
	CRE increase	0.553736875	inactive	inactive
	RBC decrease	0.558091286	inactive	inactive
	WBC decrease	0.577898551	inactive	active
9	AST increase	0.645465612	active	active
	ALT increase	0.646773705	active	active
	BUN increase	0.550055006	inactive	active
	CRE increase	0.547775947	inactive	active
	RBC decrease	0.564935065	inactive	inactive
	WBC decrease	0.613981763	active	active
10	AST increase	0.605838524	active	active
	ALT increase	0.605319041	active	active
	BUN increase	0.519916143	inactive	inactive
	CRE increase	0.532163743	inactive	inactive
	RBC decrease	0.561151079	inactive	inactive
	WBC decrease	0.6	active	active

4 Numerical example

The side effects of ten chemical products are predicted by the present algorithm; AST increase, ALT increase, BUN increase, CRE increase, RBC decrease, and WBC decrease. The products are numbered as ID $= 1, 2, \ldots,$ and 10, respectively.

When the confidence of the side effect is greater than 0.6, it is concluded that the side effect is active.

The results are shown in Tables 2 and 3. For example, the product ID $= 1$ is Levofloxacin. In Levofloxacin, five side effects except for RBC decrease are active. Table 2 shows that the present algorithm can predict four out of six side effects accurately. Totally, the prediction accuracy is 66.7%.

5 Conclusion

In the drug discovery, it is very important to predict the side effect of the drug accurately. The prediction algorithm of the drug side effect was described in this paper. The use of text mining gathers the drug side effects from the registration of medical products and then, the chemical structure information of drug is obtained from the PubChem data base. Then, the association rules between the chemical structure and the side effect of the drug are defined.

In numerical example, the present algorith was applied for predicting six side effects of ten drugs. The results show that the prediction accuracy of the algorithm is 66.7% totally. In this study, the side effects are gathered from the registration of medical products. Since the activity of the side effects depends on the gender, the age, and so on, the registration of medical products does not have enough information. Therefore, we are planning to update the association rule for improving the prediction accuracy.

References

[1] K. Enslein and P. N. Craig. A toxicity estimation model. *Journal of Environmental Pathology and Toxicology*, 2:115–121, 1978.

[2] K. Enslein and P. N. Craig. Carcinogenesis: A predictive structure-activity model. *Journal of Toxicology and Environmental Health*, 10:521–530, 1982.

[3] K. Enslein, T. R. Lander, M. E. Tomb, and W. G. Landis. A structure-activity model. *Teratogenesis, Carcinogenesis, and Mutagenesis*, 9:503–513, 1983.

[4] I. Moriguchi, S. Hirono, Q. Liu, Y. Matsushita, and T. Nakagawa. Fuzzy adaptive least squares and its use in quantitative structure-activity relationships. *Chemical and Pharmaceutical Bulletin*, 38:3373–3379, 1990.

[5] I. Moriguchi, S. Hirono, Y. Matsushita, Q. Liu, and I. Nakagome. Fuzzy adaptive least squares applied to structure-activity and structure-toxicity correlations. *Chemical and Pharma-ceutical Bulletin*, 40:930–934, 1992.

[6] I. Moriguchi, S. Hirono, Q. Liu, and I. Nakagome. Fuzzy adaptive least squares and its application to structure-activity studies. *Quantitative Structure-Activity Relationships*, 11:325–331, 1992.

[7] Gilles Klopman. Chemical reactivity and the concept of charge- and frontier-controlled reactions. *Journal of the American Chemical Society*, 90(2):223–234, 1968.
[8] H. S. Rosenkratz and G. Klopman. A structural analysis of the genotoxic and carcinogenic potentials of cyclosporin a. *Muta-genesis*, 7(2):115–118, 1992.
[9] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, 1994.
[10] Pubchem. http://pubchem.ncbi.nlm.nih.gov/.
[11] Smiles. http://www.daylight.com/dayhtml/doc/theory/theory.smiles. html.
[12] Bio-loom. http://www.biobyte.com/bb/prod/bioloom.html.