DESCRIPTIVE MINING FOR THE QSAR PROBLEM

Luminita DUMITRIU, Marian CRACIUN, Cristina SEGAL, Lucian GEORGESCU

Computer Science and Engineering Department, "Dunarea de Jos" University, str. Domneasca nr. 111, Galati, 800201, Romania

Abstract: There are several approaches in trying to solve the Quantitative Structure-Activity (QSAR) problem. These approaches are based either on statistical methods or on predictive data mining using neural networks. Among the statistical methods, one should consider regression analysis, pattern recognition (such as cluster analysis, factor analysis and principal components analysis) or partial least squares. These approaches have a low explanatory capability or non at all. This paper attempts to establish a new approach in solving QSSAR problems using descriptive data mining. This way, the relationship between the chemical properties and the activity of a substance would be comprehensibly modeled.

Keywords: Quantitative Structure-Activity Relationship, data mining, association rules, classification.

1. INTRODUCTION

The concept of Quantitative Structure-Activity Relationship (QSAR) has been introduced by Hansch and co-workers in the 1960s. Investigating the relationship between the structure and the activity of chemical compounds (SAR) supports understanding the activity of interest and allows the prediction of the activity of new compounds based on knowledge of the chemical structure alone. These predictions can be achieved by quantifying the SAR.

In the 1950's, Hansch using regression analysis succeeded to correlate biological activity with molecular properties. Nowadays, more sophisticated statistical methods or forms of pattern recognition, such as cluster analysis, factor analysis and principal components analysis, have been used in the search for patterns between biological and physical data.

Pattern recognition techniques, like multivariate statistics, along with principal component analysis (PCA) are data dimension reduction and transformation techniques from multiple experiments to the underlying patterns of information [Ebert 1984, Van de Waterbeemd 1989 a, b]. Partial least squares (PLS) is used for performing the same operations on the target properties. The predictive ability of this method can be tested using cross-validation on the test set of compounds

The aim of QSAR techniques is to find correlations between any property or form of activity, biological activity in general, and the properties of a set of molecules. However, in its most general form, QSAR is supposed to cover correlations independent of actual physicochemical properties. The goal is to connect the activities and properties by some known mathematical function, F:

Biological activity = F (Structure Properties)

The quality of any QSAR depends on the quality of the modeled data. The quality of the data relies on

This paper was recommended for publication by Viorel MINZU

multiple readings for a given observation, for which the variation of data on the same compound should be much smaller than the variation over the series.

2. DESCRIPTIVE DATA MINING

2.1. Association rules

The description of the association rules mining was first given by Agrawal et al. (1993). The set of items or attributes are designated by the literals I = { I1, I2, …, In}. A record (or transaction) contains some of the items of I, for the transactional data base case, or contains their presence information, for the relational data base case. We will denote this relation through the inclusion operator, $\subset$. The input data for the mining algorithms consists in a set of records. Any set of items of I is called an itemset. An association rule is a relation between itemsets, $A \Rightarrow B$, where A and B are contained in some transaction, and $A \cap B = \varnothing$. A is the antecedent of the rules, and B is the consequent.

An itemset is associated with a measure of frequency, called support, and support (X) denotes the ratio between the number of records that contain X and the total number of records in the data set. For a rule, the support measure refers to the $A \cup B$ set. The strength of an association $A \Rightarrow B$ is measured by the confidence of the rule determined as support $(A \cup B)$/support (A).

Mining association rules is finding all the rules that exceed two user-specified thresholds, one for support, min_sup, and one for confidence, min_conf. An itemset that exceeds the support threshold is a large itemset. Let S be a large itemset, for any $A \subset S$ and support (S)/support (A)≥min_conf, $A \Rightarrow S\text{-}A$ is an association rule. Therefore, classically finding association rules consists in two stages:

- Discovering all large itemsets. This stage is classically split into two parts: candidate-generation step, of an incremental manner, and large item selection, counting the support of the candidates and pruning the ones that are not large;
- Determining the rules with enough confidence.

The main algorithms are sequential or parallel, running on the entire data set or only on a training set, use different approaches to reduce the number of data base scans or the amount of storage memory.

2.2. Formal Concept Analysis

The theory of formal concept analysis was introduced by Wille (1982), and correlated with association rules mining by Zaki and Ogihara (1998). Let I be the set of items and let T be the set of records. Let s be a mapping between the power set of I and the power set of T, which associates to a set of itemsets all records that contain at least one of them. Let t be a mapping between the power set of T and the power set of I that associates to a set of records all itemsets contained in them. The composition $c = t \circ s$ is proven to be a closure operator.

The context (T, I, $\subset$) and the mappings s and t define a Galois connection between $\wp(I)$ and $\wp(T)$.

A concept in this context is a pair (X, Y) of closed sets, where $X \subseteq T$ and $Y \subseteq I$, with t(X)=Y and s(Y)=X (according to this, c(X)=X and c(Y)=Y, so X and Y are closed sets). X is the extent of the concept, while Y is the intent of the concept.

Every context (T, I, $\subset$) can be associated with a Galois lattice of concepts, with join and meet operators derived from the closure operator, c. The Galois lattice can be represented by a Hasse diagram. Between a pair (X1, Y1) and (X2, Y2) of concepts, the relation (X1, Y1)≥ (X2, Y2) means that $Y1 \subset Y2$ and $X1 \supset X2$. A frequent concept has support(X) ≥min_sup. All frequent itemsets are uniquely determined by the frequent concepts. There can be frequent itemsets that are not closed sets, but they are included in closed sets and are sharing the same support. These itemsets do not need to be generated (though, classical algorithms do generate them). They are called pseudo-intents.

A partial implication rule (c1, c2, conf) is associated with a pair of concepts that satisfy c1≥c2, where conf is the precision determined as support(Y2)/ support(Y1).

Association rules are represented at the intent level of a concept, as $Y1 \Rightarrow Y2\text{-}Y1$, with c2 frequent and p≥min_conf. Whenever Y1 is a pseudo-intent and Y2 is its intent, we have a global implication rule, with conf=1 (due to the same support).

Note. If (c1, c2, p) and (c2, c3, q) are implication rules, (c1, c3, p*q) is also an implication rule.

3. OUR APPROACH

We consider a database D of chemical descriptors having a target attribute *A* (activity).

Our approach considers a part of D, denoted D_M, to be used for descriptive mining and the rest, denoted D_T, to test the predictive power of the results obtained by mining.

3.1. Data and target attribute pre-processing

The original data, except for the associated target attribute, can be subject to different transformations, but for the moment we ignore this aspect.

The target attribute (in our case the lethal dose) comes either as a value or as an interval. This attribute is subjected to a clustering method in order to transform it in cluster number to whom the attribute value is a member.

3.2. Association Rule processing

The pre-processed DM data is used by the SFERA benchmark (System for Finding and Extending Rules of Association). The system's outcomes are:

- the frequent concepts;
- the association rules that are partial implication rules and
- the pseudo-intents along with their associated concepts.

A pair (pseudo-intent, associated concept) represents global implication rules that are equivalent to implications in proposition logic.

3.3. Post processing

The post-processing part creates the conditions for the main contribution to this paper, the tentative prediction.

We start from the implications resulted from SFERA. We ignore for now the partial implications.

An implication has the form:

pseudo-intent→concept,

and both expressions are conjunctions of propositions, involving relationships between D_M attributes and values (equality, set membership, interval membership etc.).

The first step is to filter the rules that comprise the target attribute either in the premises or in the conclusion. The rest of the rules are not important to our approach. We will call this set of rules R_T.

3.4. Tentative prediction

The tentative prediction part of our approach represents the main contribution to this paper.

For each chemical compound C_i in D_T, we check it against the premises of each rule R_j in R_T. We have to remember at this point that the data in D_T does not include the target attribute value. Either the compound satisfies the premises, or it doesn't.

First case: if it does, this means that the premises of R_j do not include the target attribute. C_i is then checked against the R_j conclusion. If C_i satisfies the conclusion of R_j, this means that the proposition involving the target attribute in R_j's conclusion must be true, hence C_i target attribute value can be predicted in cluster membership terms. We will denote this case as OK and memorize the corresponding cluster number. If C_i does not satisfy the conclusion of R_j, this means that an exception is raised.

Second case: if it doesn't this happens either do to the lack of information in D_T regarding the target attribute or due to a common attribute. If a common attribute is involved, then R_j is not applicable for C_i and we consider the next rule. This case will be denoted as N/A. If the target attribute is the only one responsible for R_j premises not being satisfied by C_i, we will assert the hypothesis that the proposition regarding the target attribute is true and check R_j conclusion. If the conclusion is satisfied, we assume the hypothesis is correct, and if it isn't we assume the hypothesis as false. If the hypothesis is correct we will denoted this case as OK and memorize the corresponding cluster number, otherwise we will denote it as NOK and memorize the corresponding cluster number.

In the end, we analyze the results per cluster number and compound: if there are only OKs we mark the result OK, if it is at least one NOK or an Exception we mark it as NOK. If tehre are no results we mark N/A.

We will obtain the results as in Table 1.

Table 1. Tentative prediction results.

Compound	$Cluster_1$	$Cluster_2$	...	$Cluster_m$
C_1	OK	N/A	...	NOK
C_2	NOK	OK	...	OK
...				
C_n	N/A	NOK	....	OK

3.5. Result interpretation

A last phase consists in observing the OK distribution for a compound in the result table. One or more OKs in adjacent clusters is a presumably good result, we will call it relevant, while dispersed OKs or only NOKs and N/As tells us that the quality of the QSAR is not acceptable and original data should be transformed in a different way.

Afterwards, a cross-validation phase can also be used.

4. EXPERIMENTAL RESULTS

We are presenting here the experimental results on a database already separated in D_M and D_T (we had no information for cross-validation purposes), we had structure information and lethal dose for 50 compounds and 20 compounds with no lethal dose information.

The lethal dose attribute values were separated in 6 non-overlapping toxicity classes.

We have conducted experiments on three data transformation:

- original data – comprising the element mass in every compound;
- presence data (as in market basket analysis) – comprising Boolean data reflecting the presence or absence of an element in a compound;
- percentage data – comprising percentages of elements mass in the tmolar mass of the compound.

We have found the following:

- the presence data were irrelevant - we mostly obtained NOKs;
- the percentage data had too few relevant results, allowing prediction;
- the original data are the most relevant, as we can see in Figure 1.

```
---------- Run ----------
Use:
     ToxLevel -s <substance file> -r <rules file> -o <report file>

Processing: 'Substance_file.txt' and 'Rules_file.txt' output 'Report_file.txt'
```

tox-0	tox-1	tox-2	tox-3	tox-4	tox-5	SUBSTANTA
NOK	NOK	NOK	NOK	NOK	---	PPG1013
NOK	OK	NOK	NOK	OK	---	2,4,5-TB
NOK	NOK	NOK	NOK	NOK	---	2,4,5-TES
NOK	NOK	NOK	NOK	NOK	---	benzipram
NOK	NOK	NOK	OK	NOK	---	benzoylprop
NOK	NOK	NOK	OK	OK	---	chlomethoxyfen
NOK	NOK	NOK	NOK	NOK	---	cumyluron
OK	NOK	NOK	OK	NOK	---	dicryl
NOK	NOK	NOK	NOK	NOK	---	DMNP
NOK	NOK	NOK	NOK	NOK	---	dymron
NOK	NOK	NOK	NOK	OK	---	etobenzanid
OK	NOK	NOK	NOK	NOK	---	karsil
OK	OK	OK	NOK	NOK	---	monalide
OK	NOK	NOK	NOK	NOK	---	pentanochlor
OK	NOK	NOK	NOK	OK	---	pethoxamid
NOK	NOK	NOK	NOK	NOK	---	phenisopham
NOK	NOK	NOK	OK	OK	---	phenobenzuron

Fig. 1 Results table for the original data. As one can see 50% of the new substances have relevant prediction information.

5. CONCLUSION AND FUTURE WORK

Our research has several directions for the future:

- integrating the partial implication rules in our approach;
- using cross-validation for the prediction;
- running several experiments in order to validate the approach;
- creating a methodology that facilitates the prediction using descriptive mining.

Our main contribution relies in the facts that until now only computational means or neural network-based methods were used for this purpose. All these methods have low explaining capability or none at all. Being able to predict biological activity by descriptive means leads to building an explicit model for QSAR.

6. ACKNOWLEDGEMENT

This work has been supported by the PRETOX grant no 4281/2004 of the National Plan for Research Development and Innovation, in the framework of the MENER program.

7. REFERENCES

Agrawal, R., Imielinski, T. and Swami (1993) "Mining association rules between sets of items in large databases", in Proceedings of 1993 ACM SIGMOD International Conference on Management of Data, Washington D.C., pp. 207-216.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Inkeri Verkamo, A. (1996) "Fast discovery of association rules", in Advances in Knowledge Discovery and Data Mining, ed. U. Fayyad et al., AAAI Press: Menlo Park, CA, pp. 307-328.

Agrawal, R. and Shim, K. (1996) "Developing Tightly-Coupled Data Mining Applications on a Relational Database System", in Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining, ed. E. Simoudis et al., AAAI Press, Portland, Oregon, pp. 287-295.

Cox, K., Eick, S., Wills, G. and Brachman, R. (1997) "Visual Data Mining: Recognizing Telephone Calling Fraud", in Data Mining and Knowledge Discovery Journal, Kluwer Academic Publishers, vol. 1, pp. 225-231.

Dumitriu, L., Pecheanu, E., Istrate, A. and Segal, C. (2000) “Finding association rules from relational databases”, in Proceedings of the International Conference Data Mining 2000, Cambridge, pp..

Dumitriu, L., Pecheanu, E., Istrate, A. and Segal, C.(2000) "Finding association rules from relational databases via SQL queries", in Control Engineering and Applied Informatics Journal, Mediamira Science Publisher, vol. 2, pp.5964.

Houtsma, M. and Swami, A. (1995) "Set-oriented mining of association rules in relational databases", in Proceedings of the 11th International Conference on Data Engineering, IEEE Computer Society Press, Los Alamitos, pp.25-34.

Kodratoff, Y. (1998) "Research Topics in Knowledge Discovery in Data and Texts", invited talk at the 3rd SIGMOD’98 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), Seattle, WA.

Lin, D-I. and Kedem, Z.M. (1998) "Pincer-Search: a new algorithm for discovering the maximum frequent set", in Proceedings of the 6th International Conference on Extending Database Technology, Lecture Notes in Computer Science, Springer-Verlag, 1377, pp.105-113.

Lin, J-L. and Dunham, M.H. (1998) "Mining association rules: Anti-skew algorithms", in Proceedings of the 14th International Conference on Data Engineering, IEEE Computer Society Press, Los Alamitos, pp.125-133.

Nestorov, S. and Tsur, S. (1997) "Using DB2’s Object Relational Extensions for Mining Association Rules", Technical Report TR 03690, Santa Teresa Laboratories, IBM Corporation.

Park, J.S., Chen, M. and Yu, P.S. (1995) "An effective hash-based algorithm for mining association rules", in Proceedings of 1995 ACM SIGMOD International Conference on Management of Data, San Jose, pp.175-186.

Sarawagi, S., Thomas, S. and Agrawal, R., (1998) "Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications", in Proceedings of 1998 ACM SIGMOD International Conference on Management of Data, Seattle, pp. 343-354.

Savasere, A., Omiecinski, E. and Navathe, S. (1995) "An efficient algorithm for mining association rules in large databases", in Proceedings of the 21st International Conference on Very Large Data Bases, ed. U. Dayal et al., Morgan Kaufmann, Los Altos, pp. 432-444.

Thomas, S. and Sarawagi, S. (1998) "Mining Generalized Association Rules and Sequential Patterns Using SQL Queries", in Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, pp. 344-348.

Toivonen, H., (1996) "Sampling large databases for association rules", in Proceedings of the 22nd International Conference on Very Large Data Bases, ed. T.M. Vijayarama et al., Morgan Kaufmann, Los Altos, pp. 134-145.

Wille, R. (1982) "Restructuring lattice theory: an approach based on hierarchies of concepts", in Ordered Sets, Proceedings of NATO Advanced Study Institute, D. Reidel Publisher Co., pp. 445-470.

Zaki, M.J., Parthasarathy, S.,Ogihara, M. and Li, W., (1997) "New algorithms for fast discovery of association rules", in Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, ed. D. Heckerman et al., AAAI Press, pp.283-29.

Zaki, M.J. and Ogihara, M. (1998) "Theoretical Foundations of Association Rules", in Proceedings of the 3rd SIGMOD’98 Workshop on DMKD, Seattle, WA, pp 7:1-7:8.