World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:1, No:11, 2007

Model Discovery and Validation for the Qsar Problem using Association Rule Mining

Luminita Dumitriu, Cristina Segal, Marian Craciun, Adina Cocu, and Lucian P. Georgescu

***Abstract*—** There are several approaches in trying to solve the Quantitative Structure-Activity Relationship (QSAR) problem. These approaches are based either on statistical methods or on predictive data mining. 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. Predictive data mining techniques use either neural networks, or genetic programming, or neuro-fuzzy knowledge. These approaches have a low explanatory capability or non at all. This paper attempts to establish a new approach in solving QSAR problems using descriptive data mining. This way, the relationship between the chemical properties and the activity of a substance would be comprehensibly modeled.

.

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

I. 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. 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)

Among data mining techniques the most used ones are based on neural networks [6] or on neuro-fuzzy approaches [5] or on genetic programming [4]. All these approaches predict the activity of a chemical compound, without being able to explain the predicted value.

A descriptive data mining technique recently applied to chemical compound classification is frequent sub-structure mining [2], a modified version of association rule mining.

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

To insert images in *Word,* position the cursor at the insertion point and either use Insert | Picture | From File or copy the image to the Windows clipboard and then Edit | Paste Special | Picture (with “Float over text” unchecked).

II. Descriptive Data Mining

A. Association rules

The description of the association rules mining was first given by Agrawal et al. [1]. 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

Manuscript received January 25, 2006. This work was supported in part by the Romanian Research Development and Innovation Natioanal Plan under Grant 4281/2004.

L. Dumitriu, C. Segal, M. Craciun, and A. Cocu are with the Computer Science and Engineering Dept., at the "Dunarea de Jos" University, str. Domneasca nr. 111, Galati, 800201, Romania (phone: +40 236 460182; fax: +40 236 460182; e-mail: Luminita.Dumitriu@ugal.ro).

Lucian P. Georgescu., is with the Chemistry Dept., at the "Dunarea de Jos" University, str. Domneasca nr. 111, Galati, 800201, Romania.

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∪B set. The strength of an association A⇒B is measured by the confidence of the rule determined as support (A∪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⊂S and support (S)/support (A)≥min_conf, A⇒S-A is an association rule. Therefore, classically finding association rules consists in two stages:

1) 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;
2) 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.

B. Formal Concept Analysis

The theory of formal concept analysis was introduced by Wille [7], and correlated with association rules mining by Zaki and Ogihara [8]. 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 ° s is proven to be a closure operator.

The context (T, I, ⊂) 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⊆T and Y⊆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, ⊂) 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 (X_1, Y_1) and (X_2, Y_2) of concepts, the relation $(X_1, Y_1) \geq (X_2, Y_2)$ means that $Y_1 \subset Y_2$ and $X_1 \supset X_2$. 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 $(c_1, c_2, conf)$ is associated with a pair of concepts that satisfy $c_1 \geq c_2$, where conf is the precision determined as support(Y_2)/ support(Y_1).

Association rules are represented at the intent level of a concept, as $Y_1 \Rightarrow Y_2 - Y_1$, with c_2 frequent and p≥min_conf. Whenever Y_1 is a pseudo-intent and Y_2 is its intent, we have a global implication rule, with conf=1 (due to the same support).

Note. If (c_1, c_2, p) and (c_2, c_3, q) are implication rules, $(c_1, c_3, p*q)$ is also an implication rule.

C. Frequent sub-structure mining

While most of QSAR-related techniques use the chemical compound properties data to predict activity, the approach described in [2] applies to two types of representation for chemical compounds:

1) the topological representation that sees a chemical compound as an undirected graph, having atoms in the vertices and bonds in the edges and
2) the geometric representation that sees a chemical compound as an undirected graph with 3D coordinates attached to the vertices.

The frequent sub-structure mining attempts to build, just like frequent itemsets, frequent connected sub-graphs, by adding vertices step-by step.

The main difference from frequent itemset mining is that graph isomorphism has to be checked, in order to correctly compute candidate support.

The purpose of frequent sub-structure mining is the classification of chemical compounds.

III. Our Approach

We are developing an approach that:

1) attempts describing, and not predicting, the relationship between the quantitative structure and the activity,
2) attempts describing the QSAR and not classifying the substances.

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.

D. 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. It will be considered in the Experimental results section.

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. For the moment, we do not have enough experiments to prove this is the best way to do it.

E. Association Rule processing

The pre-processed DM data is used by the SFERA

benchmark (System for Finding and Extending Rules of Association [3]). The system's outcomes are:

1) the frequent concepts;
2) the association rules that are partial implication rules and
3) the pseudo-intents along with their associated concepts, the base of global implication rules.

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

F. 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.

A global implication has the form:

pseudo-intent → concept,

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

We filter the rules that comprise the target attribute only in the conclusion. We will call this set of rules R_T.

G. 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. Either the compound satisfies the premises, or it doesn't.

If it does, C_i is then checked against the R_j conclusion ignoring the target attribute cluster id. 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 Ci 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 Rj, this means that an exception is raised and it will be marked as NOK. We will discuss later the significance of raised exception later.

If C_i doesn't satisfy the premises of R_j, the rule is skipped. This case will be denoted as N/A.

This is a straightforward case, the most favorable one. This situation is not guaranteed to occur, so if RT is void we are considering predicting in a more elaborated way as a future direction of work.

In the end, we analyze the results per rule and compound. We will obtain the results as in Table 1.

TABLE I
TENTATIVE PREDICTION RESULTS

Compound	R_1	R_2	...	R_m
C_1	OK – id_{11}	N/A	...	NOK
C_2	NOK	OK– id_{22}	...	OK– id_{2m}
...				
C_n	N/A	NOK	....	OK– id_{nm}

H. Result interpretation

A last phase consists in observing the OK distribution for a compound in the result table, using the following criteria for the corresponding situations:

1) N/A for all rules. This means that the compound belongs to a category that was not appropriately represented in the training set. This situation can occur either because similar compounds do not exist, which is highly unlikely to happen, or it is not interesting from a toxicological point of view. Either way, the compound is submitted to a chemist's attention. This situation is labeled UNKNOWN;
2) One or more NOK in the context of N/As express the fact that the generating rules are too specific to the training set, thus rule generalization should be considered and validated by the chemists. This situation is also labeled UNKNOWN;
3) The same (OK, cluster id) pair in the context of N/As and NOKs leads to a successful classification; This situation is labeled by the cluster id's value;
4) Several (OK, cluster id) pairs is interpreted as follows: the most frequent cluster id is presumed to be the classification result; if adjacent cluster ids are present it is presumed that the rule can be used for a set of cluster an it is labeled GEN, but if dispersed cluster ids are present (for example, low and high degree of toxicity for the same compound) the rules involved are considered to be too general, thus rule specialization should be taken into account and the label is SPEC. In the GEN case the most frequent cluster_id is considered as result, while in the SPEC case the results are considered as unreliable.

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

IV. Experimental Results and Model Validation

We are presenting here the experimental results on a database of 260 pre-clustered compounds split 20% in D_M and 80% D_T.

The lethal dose attribute values were separated in 4 non-overlapping toxicity clusters.

We have conducted experiments on various data transformations. The descriptor set consisting of the element masses in every compound was used as: original data; 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 total molar mass of the compound and substructure data – comprising the count of substructures (as – CH_3) and the count of relative position of benzoic bonds.

We have found the following:

1) the presence data of element mass were irrelevant - we mostly obtained NOKs;
2) the percentage data had too few relevant results, allowing prediction;
3) the original data were more relevant, as 50% of the compounds in D_T had relevant prediction information – a cluster id result or a GEN result;

4) the substructure representation offered the most expressive results.

In what concerns the model validation for the substructure-based representation, we have confronted the model results with the associated cluster id and found the results in Table 2. It is to be mentioned that the high toxicity class had the best representation – nearly 3 times more compounds were labeled "high" than each of the other categories.

We have observed the following:

1) a third of the very high toxicity compounds could not be classified, most of them have been classified as highly or very highly toxic, just 3% are unacceptable results – meaning a very highly toxic compound classified as medium or low;
2) a fifth of the high toxicity compounds could not be classified, most of them have been correctly classified, a sixth have been classified either as very high or as medium, just 4% are unacceptable results – meaning a highly toxic compound classified as low;
3) unfortunately, the performance degrades considerably for the medium and low toxicity compounds.

TABLE II
MODEL VALIDATION RESULTS

Result type / Cluster id	UNKNOWN [%]	Correct Cluster id [%]	GEN [%]	SPEC [%]
1 – very high	31.25%	15.63%	**50.00%**	3.13%
2 - high	19.79%	**62.50%**	13.54%	4.17%
3 – medium	12.50%	0.00%	**71.88%**	15.63%
4 - low	35.00%	5.00%	5.00%	**55.00%**

Furthermore, we validated the model against the training set, following the same procedure. We have found the results in Table 3. We observed the same behavior applied to the training set, too.

TABLE III
TRAINING SET VALIDATION RESULTS

Result type / Cluster id	UNKNOWN [%]	Correct Cluster id [%]	GEN [%]	SPEC [%]
1 – very high	22.22%	11.11%	**66.67%**	0.00%
2 - high	8.00%	**80.00%**	12.00%	0.00%
3 – medium	12.50%	0.00%	**87.50%**	12.50%
4 - low	50.00%	0.00%	0.00%	**50.00%**

The behavior coincidence made us think that the predefined clusters are not completely accurate, since the high toxicity set was considerably larger than the others. We have computed an over-fitting factor, expressing the fact that the predicted cluster id is the one of the largest cluster instead of the right one. We have found out that there is an extremely strong correlation, almost 100%, between the behavior shown in tables 2 and 3 and the over-fitting factor. In fact, the compounds in the very high, medium and low toxicity clusters were mostly predicted to belong to high cluster, due to the large representation of this cluster.

This way, we have found out that the original clustering was inappropriate. Thus, we are now considering a new approach in what concerns classifying the data into toxicity classes.

II. V. Conclusion and Future Work

The conclusion to be drawn from our experiments is that for a well represented class the promising results may be misleading and we should consider a new way of classifying data into toxicity classes as well as a larger database, with larger representation for the extreme classes.

Our main contribution relies in the facts that until now only computational means or neural network-based methods were used for QSAR. 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. Also, being able to validate the model against the training set, allows us to model data starting from the correct premises.

Our research has several directions for the future:

1) performing a 5 fold cross validation, in order to verify the prediction accuracy;
2) considering a rule generalization and specialization approach;
3) building a compound taxonomy, in order to class-specifically predict activity based on class specific models;
4) integrating the partial implication rules in our approach;
5) creating a methodology that facilitates the prediction using descriptive mining.

References

[1] 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.

[2] Deshpande, M., Kuramochi, M., Wale, N. and George Karypis, G., (2005) "Frequent Substructure-Based Approaches for Classifying Chemical Compounds" in IEEE Transaction on Knowledge and Data Engineering, Vol 17(8): 1036-1050

[3] Dumitriu, L., (2002) "Interactive mining and knowledge reuse for the closed-itemset incremental-mining problem", Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, ed. U. Fayyad, Vol 3:2, pp. 28-36, ian. 2002, http://www.acm.org/sigkdd/ explorations.

[4] Langdon, W. B. and Barrett, S. J., (2004) "Genetic Programming in Data Mining for Drug Discovery", in Evolutionary Computing in Data Mining, Springer, 2004, Ashish Ghosh and Lakhmi C. Jain, 163, Studies in Fuzziness and Soft Computing, 10, ISBN 3-540-22370-3, pp. 211--235.

[5] Neagu, C.D., Benfenati, E., Gini, G., Mazzatorta, P., Roncaglioni, A., (2002) "Neuro-Fuzzy Knowledge Representation for Toxicity Prediction of Organic Compounds", in Proceedings of the 15th European Conference on Artificial Intelligence, Frank van Harmelen (Ed.):, ECAI'2002, Lyon, France, July 2002. IOS Press 2002: pp. 498-502

[6] Wang, Z., Durst, G., Eberhart, R., Boyd, D., Ben-Miled, Z., (2004) "Particle Swarm Optimization and Neural Network Application for QSAR", in the Proceedings of the 18th International Parallel and Distributed Processing Symposium (IPDPS 2004), 26-30 April 2004, Santa Fe, New Mexico, USA. IEEE Computer Society 2004, ISBN 0-7695-2132-0.

[7] 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.

[8] 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.