Toxicology Analysis by Means of the JSM-method

V.G. Blinova, D.A. Dobrynin, V.K. Finn, S.O. Kuznetsov, and E.S. Pankratova

All-Russia Institute for Scientific and Technical Information (VINITI),
Usievicha 20, 125219 Moscow, Russia

{blinova, dobr, finn, serge, pankr}@viniti.ru

Abstract A learning model based on the JSM-method of hypotheses generation is used for learning from positive and negative examples and prediction of toxicity of chemical compounds from the data set of Predictive Toxicology Challenge (PTC'2001). As a data language the Fragmentary Code of Substructure Superposition (FCSS) is used.

1. Learning Model

In this work we use a learning model from [3] and data representation [1, 2, 6] for the analysis of data from [7]. The learning model from [3], called JSM-method in honour of John Stuart Mill, complies with the common paradigm of learning from positive and negative examples: given descriptions of positive and negative examples with respect to a goal attribute, construct a generalization of the positive examples that would not cover any negative example. For original definitions of the JSM-method (in terms of an infinite-valued first-order predicate calculus extended with quantifiers over tuples of variable length) we refer to [3]. Here, for the sake of simplicity, we present a fragment of it (to be used for the Predictive Toxicology Challenge [7]) in terms of the Formal Concept Analysis (FCA) as we did, e.g., in [5]. First, we recall some basic notions of the Formal Concept Analysis [4].

A *(formal) context* is a triple of sets $K = (G, M, I)$, where G is called a set of objects, M is called a set of attributes, and $I \subseteq G \times M$ is a relation. For $g \in G$ and $m \in M$ gIm is interpreted as «object g has attribute m». For $A \subseteq G$ and $B \subseteq M$ *derivation operators* are defined as follows:

$A' = \{m \in M \mid \forall g \in A\ (gIm)\}$;

$B' = \{g \in G \mid \forall m \in B\ (gIm)\}$.

A *(formal) concept* of a formal context (G, M, I) is a pair (A, B), where $A \subseteq G$, $B \subseteq M$, $A' = B$, and $B' = A$. The set A is called the (*formal*) *extent* and B the (*formal*) *intent* of the concept (A, B).

Let w be a *goal* attribute, different from attributes from the set M (we call them *structural* attributes). For example, in the toxicology analysis w correspond to toxicity and the structural attributes from M correspond to particular molecular substructures.

Input data for learning are given by sets of positive, negative, and undetermined examples with respect to the goal attribute w. The undetermined examples are to be classified by means of the learned rules. In terms of FCA, this situation can be described by three contexts: a positive context $K_+ = (G_+, M, I_+)$, a negative context $K_- = (G_-, M, I_-)$, and an undetermined context $K_\tau = (G_\tau, M, I_\tau)$. Here G_+, G_- and G_τ are sets of positive, negative, and undetermined examples, respectively; M is a set of *structural* attributes; $I_\varepsilon \subseteq G_\varepsilon \times M$, $\varepsilon \in \{+, -, \tau\}$ are relations that specify the structural attributes of positive, negative, and undetermined examples, respectively. The derivation operators in these three contexts are denoted by superscripts $+, -, \tau$,

respectively. Now, a positive hypothesis (called a *counterexample forbidding hypothesis* in [3]) is defined in the following way.

If intent h_+ of a concept of the positive context K_+ is not contained in the intent of any negative example (i.e., $\forall\, g_- \in G_-,\ h_+ \not\subset g_-^{-}$), then it is called a *positive hypothesis* with respect to. the goal attribute *w*. *Negative hypotheses* are defined similarly: If intent h_- of a concept of the negative context K_- is not contained in the intent of any positive example (i.e., $\forall\, g_+ \in G_+,\ h_- \not\subset g_+^{+}$), then it is called a *negative hypothesis* with respect to the goal attribute *w*.

Hypotheses are used for the classification of undetermined examples from G_τ. If intent g_τ^τ of an undetermined example $g_\tau \in G_\tau$ contains a positive hypothesis h_+ (i.e., $g_\tau^\tau \supseteq h_+$), we say that h_+ is *for the positive classification of* g_τ. A *hypothesis for negative classification of* g_τ is defined similarly: If intent g_τ^τ contains a negative hypothesis h_- (i.e., $g_\tau^\tau \supseteq h_-$), we say that h_- is *for the negative classification of* g_τ. If there is a hypothesis for positive classification of g_τ and no hypothesis for negative classification of g_τ, then g_τ is *classified positively*. *Negative classifications* of g_τ are defined similarly, i.e., if there is a hypothesis for its negative classification and there is no hypotheses for its positive classification. If g_τ^τ does not contain any negative or positive hypothesis, then no classification is made. If g_τ^τ contains both positive and negative hypotheses, then the classification is said to be contradictory.

Note that JSM-method provide for more possibilities of definition of hypotheses and classifications.

2. Descriptor Language for Representing Chemical Compounds

Toxicity of a chemical compound, as any other biological activity, depends on the character of weak bonds that arise between the compound and the biological receptor during their interaction. It is well-known that the possibility of these bonds depend on π-electrons of the compound. That is why as a descriptor language we used the *fragmentary code of substructure superposition* (FCSS) [1, 2, 3]. By means of this language a chemical compound is described as a set of substructures that are centers of localization of π-electrons. The description of a chemical compound by means of FCSS language is often more relevant for the study of biological activities than that by means of structural formula and its simplifications.

Now we give a short description of the FCSS language. First, active or *descriptor centers* (DC) – atoms or groups of atoms that can be centers of "weak" interaction – are distinguished in a chemical compound. They are atoms and groups of atoms that contain movable π- and d-electrons or a whole electrostatic charge, i.e., all heteroatoms (N, O, S, P, halloids, metals, etc.), carbon pairs connected with multiple (double, triple) bonds and aromatic cyclic systems as a whole. The lists of descriptor centers of FCSS is given in Tables 1 and 2.

The elements of FCSS descriptor language fall into two categories: linear and cyclic descriptors. A linear descriptor is given by a pair of DCs connected by a chain of carbon pairs. For these chains it is also essential whether there is a conjugation (common d- and π-electrons) between atoms of these chains. A linear FCSS descriptor has the following form:

Descriptor Center 1	Chain length (in carbon atoms)	Descriptor Center 2	Conjugation attribute (binary)

A linear descriptor is given by seven digits: two digits for each descriptor center (first goes a DC with the smaller number) and the chain length. Conjugation attribute is given by a single bit (1 if there is conjugation and 0 otherwise).

Cyclic descriptors have the following structure:

«Head»	«Body»	«Tail»
Geometric form of the cyclic system	The number of π-electrons in the conjugated system	Location of heteroatoms

The Head gives the size of the cycle (the number of atoms) in the case of a monocycle and the size and location of separate simple cycles in case of polycyclic systems.

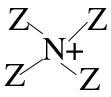
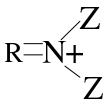
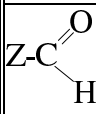
Table 1

List of FCSS Descriptor Centers of the first type

Atom	Valences	DC Number	Atom	Valences	DC Number
Li	1	43	Ga	3	43
Be	1	43	Ge	4	43
B	3	53	As	3,5	51
N-	2	00	As+	4	51
O-	1	15	Se	2,4,6	54
O+	3	16	Br	1	31
F	1	32	Br	1	48
Na	1	43	Rb	1	43
Mg	2	43	Sr	2	43
Al	3	43	Y	3	43
Si	2,4	52	Zr	4	43
P	3,4,5	47	Nb	2,5	43
P+	4	47	Mo	2,4,6	43
S	6	23	Ag	1,2	43
S+	3,4	23	Cd	2	43
Cl	1	31	Sn	2,4	43
K	1	43	Sb	3,5	51
Ca	2	43	Te	2,4,6	54
Sc	3	43	I	1	31
Ti	4	43	I	1	49
V	2,3,4,5	43	Ba	2	43
Cr	2,3,4,6	43	Pt	2	43
Mn	2,4,7	43	Au	1,2	43
Fe	2,3	43	Hg	1,2	43
Co	2	43	Ti	3	43
Ni	2	43	Pb	2,4	43
Cu	1,2	43	Bi	2,3,5	43
Zn	2	43			

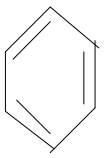
Table 2

List of FCSS Descriptor Centers of the second type

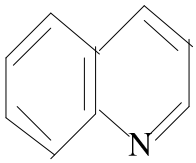
DC	Valence	Code	DC	Valence	Code	DC	Valence	Code
	4	01	O=R	2	13	R-OH	2	11
	4	07	Z-SH	2	21	R-O-R	2	12
Z-NH-Z	3	02	R-S-R	2	22		2	14
	3	03	S=R	2	25	$Z\text{-}CH_3$	4	41
R=NH	3	04	R≡N	3	06	R≡CH	4	41, 80
R=N-R	3	05	$R{=}CH_2$	4	41, 80			

Z denotes any atom, R denotes any atom except for H.

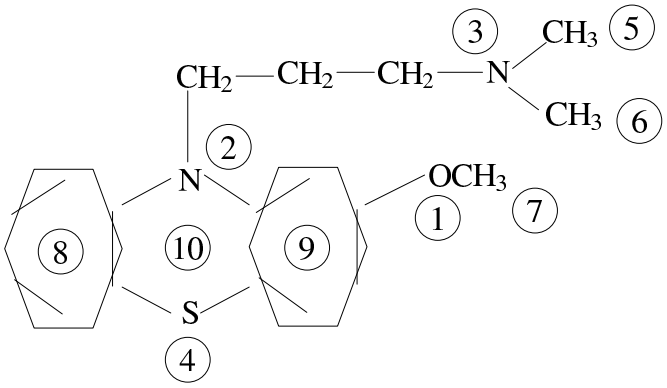
Example:



6,06

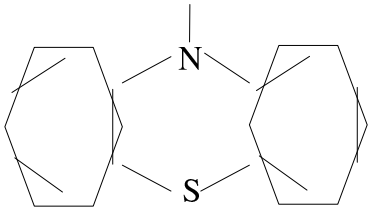


66,10M1



DCs are given by numbers. The cyclic part is coded in the following way:

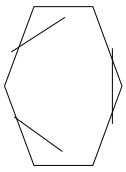
8 9 10 Is coded as a whole cyclic system



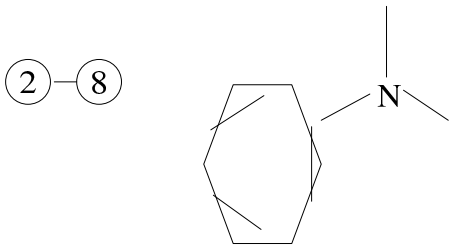
66B6, 16 N6S13

Aromatic parts of the cyclic system are coded separately.

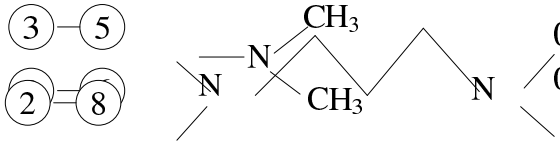
8 9 ,



6,06

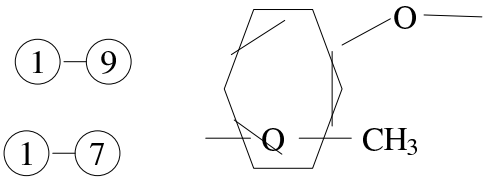


0300331



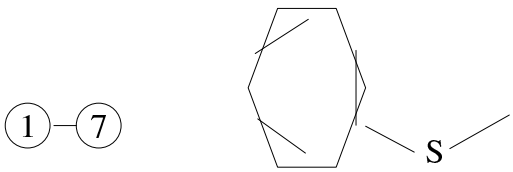
0301410

0303030



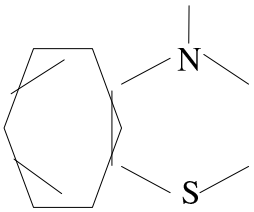
1200331

1201410

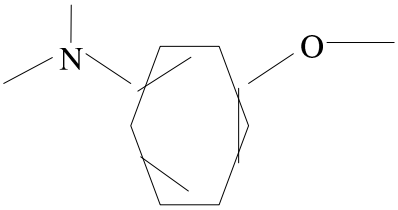


2200311

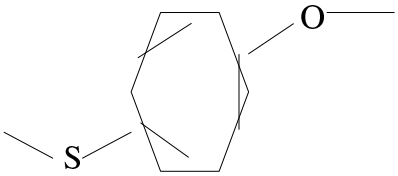
Descriptors in the bensole ring:



0362221



0363120



1264221

As a result we obtain the following code of the compound:

6,06	0300331	0301410
66B6,16N6S13	0303030	1200331
	0362221	1201410
	0363120	22200331
	1264221	

3. Predictions for Toxicology Data Set

Predictions for toxicology data set were made on the basis of the classification model from Section 1 with the use of FCCS descriptors as structural attributes (set *M*). Positive and negative examples were provided in [7].

The goal attribute was toxicological activity for each of the four sex/species groups. Positive and negative classifications were considered as the corresponding predictions for toxicity. Contradictory classifications were considered as ambiguous and were ignored.

The evaluation of results of ROC analysis show that the predictions made were always among the optimal ones (not also that we did not use the possibility of presenting a multiple, i.e., a disjunctive prediction model). The ROC analysis testifies to the «conservatism» of the JSM-predictions: it makes fairly good number of correct predictions and makes almost no errors. Besides the relevance of the FCSS language we can explain this by two features of the learning model: first, we strictly forbid the counterexamples of a hypothesis and, second, a hypothesis itself, as a «similarity» of some positive examples, is in a certain sense their *least* general generalization.

References

1. Avidon V.V., Pomerantsev A.B. Structure-activity relationship oriented languages for chemical structure representation, *J. Chem. Inf. Comput. Sci.*, vol.22, no. 4, (1982) 207-214.
2. Blinova V., Dobrynin D., Languages for Representing Chemical Conpounds for Intelligent Systems of Chemical Design, *Automated Documentation and Mathematical Linguistics*, no. 3 (2000).
3. Finn, V.K., Plausible Reasoning in Systems of JSM Type, *Itogi Nauki i Tekhniki, ser. Informatika*, vol. 15 (1991) 54–101.
4. Ganter, B., Wille, R., *Formal Concept Analysis. Mathematical Foundations*, Springer, (1999).
5. Ganter B., Kuznetsov S.O., Formalizing Hypotheses with Concepts, *Proc. of the 8th International Conference on Conceptual Structures* (ICCS 2000), Lecture Notes in Artificial Intelligence, vol. 1867 (2000) 342-356.
6. Leibov A.E., Automatic Coding of Chemical Structures by Means of FCSS Codes. *Itogi Nauki i Tekhniki, ser. Informatika*, vol. 15 (1991) 141-158 [in Russian].
7. PKDD'01 Predictive Toxicology Challenge: http://www.informatik.uni-freiburg.de/~ml/ptc/