A Hybrid Feature Selection Algorithm for the QSAR Problem

Marian Viorel Crăciun, Adina Cocu, Luminița Dumitriu, and Cristina Segal

Department of Computer Science and Engineering,
University "Dunărea de Jos" of Galați, 2 Științei, 800146, Romania
{mcraciun, cadin, lumi, csegal}@ugal.ro

Abstract. In this paper we discuss a hybrid feature selection algorithm for the Quantitative Structure Activity Relationship (QSAR) modelling. This is one of the goals in Predictive Toxicology domain, aiming to describe the relations between the chemical structure of a molecule and its biological or toxicological effects, in order to predict the behaviour of new, unknown chemical compounds. We propose a hybridization of the ReliefF algorithm based on a simple fuzzy extension of the value difference metric. The experimental results both on benchmark and real world applications suggest more stability in dealing with noisy data and our preliminary tests give a promising starting point for future research.

1 Introduction

Predictive Toxicology (PT) is one of the newest targets of the Knowledge Discovery in Databases (KDD) domain. Its goal is to describe the relationships between the chemical structure of chemical compounds and biological and toxicological processes. This kind of relationships is known as Structure-Activity Relationships (SARs) [1]. Because there is not a priori information about the existing mathematical relations between the chemical structure of a chemical compound and its effect, a SAR development requires close collaboration between researchers from Toxicology, Chemistry, Biology, Statistics and Artificial Intelligence – Data Mining and Machine Learning domains [2], in order to obtain reliable predictive models.

In real PT problems there is a very important topic which should be considered: the huge number of the chemical descriptors. Data Mining, as a particular step of the Knowledge Discovery in Databases (KDD) process [3] performs on the data subset resulted after the pre-processing procedure. Irrelevant, redundant, noisy and unreliable data have a negative impact, therefore one of the main goals in KDD is to detect these undesirable properties and to eliminate or correct them. This assumes operations of data cleaning, noise reduction and feature selection because the performance of the applied Machine Learning algorithms is strongly related with the quality of the data used.

Besides the removal of the problematic data, feature selection could also have importance in the reduction of the horizontal dimension and, consequently, of the hypothesis space of the data set: less attribute are more comprehensible, the smaller dimension of the input space allows faster training sessions, or even an improvement in predictive performance.

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part I, LNCS 3991, pp. 172 – 178, 2006.

In the literature there are at least three broad trends in dealing with the problem of features selection: filter, wrapper and embedded methods. The filter based approach evaluate the quality of the attributes separately of a machine learning algorithm and taken into account the contribution of the attribute individually (e.g. the Information Theory based measures: Information Gain, Gain Ratio) or in the context of the other attributes from the training set (e.g. Relief [4], ReliefF [5], RReliefF [6]). Contrary, the wrapper methods use the machine learning algorithm with the purpose of quality evaluation of the attributes. The embedded techniques are nothing else than machine learning algorithms having the ability to extract most suited attributes learning the training data in the same time (e.g. Decision Trees, Bayesian Networks). In real world application wrapper and embedded methods seams to select the proper attributes offering superior performance. But, they are strongly dependent on the learner. On the other side, the filter techniques present an insight of the training data set and its statistical properties.

This paper presents an upgrading of the classical Relief algorithm using fuzzy theory [7]. This simple patch allows the algorithm to evaluate fuzzy attributes as well as the categorical and numerical ones. Also, if the fuzzyfication of the training data set is possible the algorithm will be less sensitive on noise. This Fuzzy Relief is tested and compared with few other methods using a well known data set and a toxicological set.

2 Distance and Fuzzy Sets

Usually, any similarity measure stands on a distance function or a metric. But when this is extended to measure the distance between two subsets of a metric space, the triangular inequality property is sometimes lost.

Let be $X \neq \Phi$ a set and $d: X \times X \to \boldsymbol{R}_+$ a distance. Then, in this situation D could measure the distance between two sets:

$$D(A,B) = \begin{cases} \inf_{x \in A, y \in B} d(x,y), \ if \ A \neq \Phi, B \neq \Phi \\ 0, \ if \ A = \Phi \ or \ B = \Phi \end{cases}, \quad A, B \subset X \ . \tag{1}$$

Nevertheless, even this is not properly a distance function; it might be used to measure the distance between two fuzzy sets:

$$d(A,B) = \int_0^1 D(A^\alpha, B^\alpha) d\alpha \ . \tag{2}$$

where A and B are two fuzzy sets and A^α, B^α are their α-cuts:

$$A^\alpha = \{t \in X \mid \mathrm{A(t)} \geq \alpha\}, \ \alpha > 0 \ . \tag{3}$$

There are many ways to measure the distance between two fuzzy sets [8], [9]:

– using the difference between the centres of gravity,

$$d(A,B) = |CG(A) - CG(B)|, \ CG(A) = \frac{\int_U x \mu_A(x) dx}{\int_U \mu_A(x) dx} \ . \tag{4}$$

– Hausdorff distance,

$$D(A,B) = \sup_{\alpha\in[0,1]} \max\{|\, a_2(\alpha) - b_2(\alpha)\,|, |\, a_2(\alpha) - b_2(\alpha)\,|\}\ . \tag{5}$$

where α-cuts $A^{\alpha} = [a_1(\alpha), a_2(\alpha)]$ and $B^{\alpha} = [b_1(\alpha), b_2(\alpha)]$, $\alpha\in[0,1]$,

– C_{∞} distance,

$$C_{\infty} = ||A - B||_{\infty} = sup\{|A(u) - B(u)| : u\in U\}\ . \tag{6}$$

– Hamming distance,

$$H(A,B) = \int_{U} |\, A(x) - B(x)\,|\, dx\ . \tag{7}$$

With the support of one of these quasi-distance function, the similarity between fuzzy sets (linguistic variables) can be easily measured.

3 Fuzzy Relief

The Relief (Relief, ReliefF, RReliefF) family methods evaluates the contribution of the values of each attribute in the training data set to distinguish between most similar instances in the same class and in opposite (different) class. Using the difference function each attribute is scored being penalized if the values of the attribute are different for the instances in the same class and recompensed if the values of the attribute are different for the instances in the opposite class.

In order to extend the difference function to handle complex application with fuzzy attributes where the values might be overlapping, one could use one of the distance functions presented in the previous section. In this way, the evaluation of the similarity will be less noise sensitive and more robust. The similarity is determined using the distance (difference) function between the values $v^{(i,j)}$ and $v^{(i,k)}$ of the A_i attribute:

$$\mathrm{diff}(i, t_j, t_k) = \begin{cases} 0, & \text{if } A_i \text{ categorial and } v^{(i,j)} = v^{(i,k)} \\ 1, & \text{if } A_i \text{ categorial and } v^{(i,j)} \neq v^{(i,k)} \\ \dfrac{d(v^{(i,j)}, v^{(i,k)})}{Max_i - Min_i}, & \text{if } A_i \text{ numerical or fuzzy} \end{cases} \tag{8}$$

4 Experimental Results

The method proposed in the previous section is compared with few other classical methods: context dependent (such us ReliefF [5]) and myopic (such as Information Gain, Gain Ratio, and Chi Squared Statistics [11]) which evaluate the individual quality of each attribute.

The data sets used in the experiments were: the classical Iris set [12] and a real world toxicological data set provided by our partners in the framework of a national research project:

Iris data set
This well known data set is very often used as a benchmark and numerous publications reported very good results on it. It consists from 150 iris plant instances described by 4 attributes: sepal length, sepal width, petal length, and petal width. The plants are equal distributed in 3 classes: Iris Setosa, Iris Versicolour and Iris Virginica.

Toxicological data set
The set contains 268 organic chemical compounds characterized by 16 chemical descriptors and their toxic effect (the lethal dose) against 3 small mammals (rat, mouse, and rabbit). The compounds are grouped in 3 toxicity classes (low, medium, high) using the lethal doses. The class distribution is not equal this time. There are more chemical compounds in the most toxic class (156) than in the other two (72 and 40, respectively).

The data sets are artificially altered with different levels of noise between 0 and 25%. Not all the attributes were affected by noise. Only the values of 50% from the characteristics, the most predictive ones, discovered by the majority of the feature selection methods on the noiseless sets, were randomly modified. (e.g. for the Iris dataset the last two attributes are strongly correlated with the class and will be affected by noise, for the toxicology data set, the first eight descriptors are not altered).

The feature selection methods used in this set of experiments are:

Information Gain (IG) – evaluates the worth of an attribute A by measuring the information gain with respect to the class C:

$$Gain(A) = I(A;C) = H_C - H_{C|A}\,. \tag{9}$$

$$Gain(A) = -\sum_k p_{k\cdot} \log p_{k\cdot} - \sum_j p_{\cdot j} \sum_k p_{k|j} \log p_{k|j}\,. \tag{10}$$

Gain Ratio (GR) – evaluates the worth of an attribute by measuring the gain ratio (the information gain normalized by the entropy of the attribute) with respect to the class:

$$GainR(A) = \frac{\mathrm{Gain(A)}}{\mathrm{H_A}}\,. \tag{11}$$

Chi squared statistics (Chi) – evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class.

ReliefF – evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.

Bayesian Networks (BN) – evaluates the importance of one attribute by observing how much the predictive performance of the model drops if we remove the corresponding variable. An important feature makes the prediction accuracy of the model

to drop down when it is left out. On the other hand, if it removing a feature not affect to much the performance of the selected model, then it is less important. The predictive performance of the classifier is estimated with leave-one-out (LOO) cross validation method [13, 14].

FuzzyRelief – is the *Relief* algorithm working with fuzzy variables (after the fuzzyfication of the training data).

In order to compare the different filters for attribute selection, the well known Machine Learning method k-Nearest Neighbours (kNN), with k=10 and inverse distance weighting is used. The results of the experiments are obtained with *Weka* [11] and *B-Course* [15] software. Generally, the predictive performances of machine learning methods decrease and the importance of feature selection appears when the uncertainty (noise) in data increases.

The next two tables show the cross-validated (10-fold) performances of kNN in different circumstances. In both cases, first row contains the predictive performance obtained using all the available attributes for comparison purposes. The next rows include the kNN performance obtained in combination with different feature selection techniques: only half of the attributes, those with the higher ranks, are used for predictions.

Table 1 illustrates the predictive performance on the *Iris* dataset. The results of kNN working with or without feature selection methods are similar for both clean and very low noise level (5%) data. For noise level higher than 10%, all the attributes are a better choice in all situations. This is explainable taken into account the fact that the unaltered characteristics can compensate the noisy ones. The proposed hybrid selection technique, *FuzzyRelief*, demonstrates an analogous behaviour with almost all the other feature selection techniques (except *BN*).

Performances on the toxicological data are shown in Table 2. In this case the strength and flexibility of *FuzzyRelief* are more obvious. The ability of fuzzy techniques to deal with imperfect data (noise over 10%) in combination with a strong and well known feature selection method such as *ReliefF* yield to the appropriate attributes subset to describe the QSARs. The kNN accuracy in classification is enhanced.

Table 1. Prediction accuracy – Iris data set

Feature selection + kNN	**Noise level**					
	0%	5%	10%	15%	20%	25%
none+kNN	95%	92%	89%	90%	83%	81%
IG+kNN	95%	93%	82%	86%	79%	79%
GR+kNN	95%	93%	82%	86%	79%	73%
Chi+kNN	95%	93%	82%	86%	80%	79%
BN+kNN	95%	93%	86%	89%	80%	73%
ReliefF+kNN	95%	91%	83%	83%	79%	79%
FuzzyRelief+kNN	95%	93%	83%	83%	79%	79%

Table 2. Prediction accuracy – toxicological data set

Feature selection + kNN	Noise level					
	0%	5%	10%	15%	20%	25%
none +kNN	63%	56%	59%	57%	62%	56%
IG+kNN	57%	57%	57%	57%	62%	57%
GR+kNN	57%	57%	57%	57%	62%	57%
Chi+kNN	57%	57%	57%	57%	62%	57%
BN+kNN	59%	58%	57%	62%	63%	61%
ReliefF+kNN	65%	60%	59%	55%	61%	60%
FuzzyRelief+kNN	63%	59%	59%	64%	63%	62%

5 Conclusions

The results so far show the strength and the flexibility of *Fuzzy Relief* when possible uncertain data is used for training predictive data mining techniques. Its ranking performance proved in the case studies presented is comparable and some times better than the performances of other filter methods especially when the data is affected by noise. One of its drawbacks is the time needed to evaluate the similarity between the linguistic variables (between the fuzzy sets). Of course, depending on the target problem of the KDD process, the tradeoffs between the quality of the solutions and data, the dimensionality of the data sets and the available time will dictate the right strategy.

The future research will be focused in evaluating different similarity measures between the different fuzzy attributes and in testing the method on more real world data mining applications. Also, another future interest will be focused on evaluating different classifying methods in combination with *FuzzyRelief.*

Acknowledgement

This work was partially funded by the PRETOX Romanian project under the contract 407/2004 MENER.

References

1. Hansch, C. Hoekman, D., Leo, A., Zhang, L., Li, P., The expanding role of quantitative structure-activity relationship (QSAR) in toxicology, Toxicology Letters 79 (1995) 45-53
2. Y-t. Woo, A Toxicologist's View and Evaluation, Predictive Toxicology Challenge (PTC) 2000-2001
3. Fayad, U.M., Piatesky-Shapiro, G., Smyth, P., Uthurusamy, eds. Advances in Knowledge Discovery and Data Mining, (AAAI/MIT Press Menlo Park, CA), (1996)
4. Kira, K., Rendell, L. A., A practical approach to feature selection, In International Conference on machine learning, Morgan Kaufmann , (1992) 249-256
5. 5 Kononenko, I., Estimating attributes: Analysis and Extension of Relief. In Proc. of ECML'94, the Seventh European Conference in Machine Learning, Springer-Verlag, (1994) 171-182

6. Robnik Sikonja, M. and Kononenko, I., An adaptation of Relief for attribute estimation in regression, In Fisher, D., editor, Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97), Morgan Kaufmann Publishers (1997) 296–304
7. Zadeh, L.A., Fuzzy Logic and Approximate Reasoning, Synthese, 30 (1975) 407-428
8. Pal, Sankar K., Shiu Simon C. K., Foundations of Soft Case-Based Reasoning, John Wiley and Sons, (2004)
9. Fuller, R., Introduction to Neuro-Fuzzy Systems, Advances in Soft Computing Series, Springer-Verlag Berlin, (1999)
10. Wilson, D.R., Martinez, T.R., Improved Heterogeneous Distance Functions, Journal of Artificial Intelligence Research (JAIR), 6-1 (1997) 1-34
11. Witten, I.H., Frank E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, San Francisco, CA (1999)
12. Fisher, R.A., Iris Plant Database
13. Domingos P., Pazzani M., Beyond Independence: Conditions for the optimality of the simple bayeisan classifier, Proceeding of the Thirteenth International Conference on Machine Learning, (1996)
14. Elkan Charles, Naïve Bayesian Learning, Technical Report, University of California, (1997)
15. Myllymäki P., Silander T., Tirri H., Uronen P., B-Course: A Web-Based Tool for Bayesian and Causal Data Analysis. International Journal on Artificial Intelligence Tools, Vol 11, No. 3 (2002) 369-387