Feature vs. Classifier Fusion for Predictive Data Mining – a Case Study in Pesticide Classification

Henrik Boström
School of Humanities and Informatics
University of Skövde
P.O. Box 408, SE-541 28 Skövde
Sweden
henrik.bostrom@his.se

Abstract - ***Two strategies for fusing information from multiple sources when generating predictive models in the domain of pesticide classification are investigated: i) fusing different sets of features (molecular descriptors) before building a model and ii) fusing the classifiers built from the individual descriptor sets. An empirical investigation demonstrates that the choice of strategy can have a significant impact on the predictive performance. Furthermore, the experiment shows that the best strategy is dependent on the type of predictive model considered. When generating a decision tree for pesticide classification, a statistically significant difference in accuracy is observed in favor of combining predictions from the individual models compared to generating a single model from the fused set of molecular descriptors. On the other hand, when the model consists of an ensemble of decision trees, a statistically significant difference in accuracy is observed in favor of building the model from the fused set of descriptors compared to fusing ensemble models built from the individual sources.***

Keywords: feature fusion, classifier fusion, decision fusion, chemoinformatics

1 Introduction

Data mining techniques have become standard tools to develop predictive and descriptive models in situations where one wants to exploit data collected from earlier observations in order optimize future decision making [1]. In the case of predictive modeling, one typically tries to estimate the expected value of a particular variable (called the dependent variable), given the values of a set of other (independent) variables. In the case of a nominal dependent variable (i.e., the possible values are not given any particular order), the prediction task is usually referred to as classification, while the corresponding task when having a numerical dependent variable is referred to as regression. One usually wants the model to be as correct as possible when evaluated on independent test data, and several suggestions for how to measure this have been proposed. For classification, such measures include accuracy, i.e., the percentage of correctly classified test examples, and the area under the ROC curve (AUC), i.e., the probability that a test example belonging to a class is ranked as being more likely belonging to the class than a test example not belonging to the class [2]. Besides the ability to make correct predictions, one is also often interested in obtaining a comprehensible model, so that the reasons behind a particular classification can be understood, and also that one may gain insights into what factors are important for the classification in general. Examples of such comprehensible models are decision trees and rules, e.g. [3], while examples of models not belonging to this group, often called black-box, or opaque, models, include artificial neural networks and support vector machines (see e.g. [4]).

A central issue when developing predictive models from multiple sources is how to best integrate – or fuse – the information from these sources. Should one fuse all available data and then generate a predictive model, or should one generate models from the different sources and then fuse the models?

In this work we address this general question by studying a particular problem within the domain of chemoinformatics – pesticide classification. This problem concerns classifying molecules into being a pesticide or not, based on different molecular descriptors. What is interesting from an information fusion perspective with this and many similar problems in chemoinformatics is that there is no (known) single source of molecular descriptors that always is optimal, but there is a large number of different sources developed for different purposes. Hence, the question for a particular application in this area is how to best combine the information from these sources when generating a predictive model. In this study, we consider two basic strategies for fusing this information: i) fusing the sources before generating the model, and ii) generating models from each source and then fusing the models into a global predictive model. The research question of this study is whether or not the choice of strategy has any impact on the resulting model w.r.t. predictive performance in the domain of pesticide classification.

In the next section, we describe the problem of pesticide classification in a little more detail, and point out the sources of information used in the study. In section three,

we briefly describe the employed predictive data mining techniques as well as the methods for fusing features and classifiers. In section four, we present the experimental setup and the results from the experiment. Finally, in section five we give concluding remarks and point out directions for future work.

2 Data Sources

The particular class of molecules that we consider in this study are so-called *pesticides*, i.e., molecules that may be used for preventing or destroying pests, such as microbes, insects and other organisms that are considered to be harmful.

The set of pesticides used in this study was selected from the e-Pesticide Manual [5]. The structures were converted into SMILES strings [6] and standardized in order to allow further processing such as removing duplicates and descriptor calculations. After removing duplicate structures, we retained 1613 unique compounds with molecular weight (MW) between 100-700 Dalton. A similar sized counter set (i.e., molecules that have not been classified as pesticides) was selected from a set of compounds obtained from external vendors and compiled at AstraZeneca. The selection was done by matching the MW distribution between the two sets using a genetic algorithm, see Figure 1. The algorithm minimizes the F-test statistics between two datasets in order to obtain a non-significant difference between the normal distributions of a desired property. This was done in order to avoid trivial classifications based on molecular size.

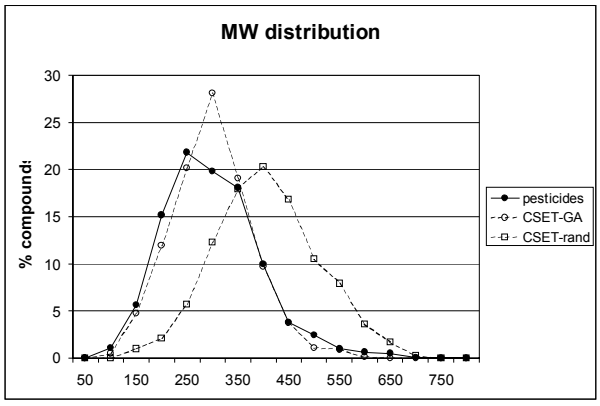

Figure 1. Molecular weight distributions for the pesticides and counter sets. CSET-rand is randomly selected while CSET-GA is selected using a genetic algorithm.

The following 2D molecular descriptors have been calculated for both sets with DRAGON [7]:

- B01: 48 constitutional descriptors (the so-called 0D descriptors independent of molecular connectivity such as MW, number of atoms, sum of atomic properties, etc.)
- B17: 152 chemical functional group counts
- B18: 120 atom centered fragments. Although this set was initially proposed for predicting molecular hydrophobicity (lipophilicity) [8], it has proved to be very useful for several other binary classification tasks [9,10]
- B20: 28 molecular properties calculated from models and empirical descriptors.

For a detailed description of the selected molecular descriptors, see [7] and the references therein.

3 Methods

3.1 Decision trees and ensembles

Techniques for generating decision trees are perhaps among the most well-known methods for predictive data mining. Early systems for generating decision trees include CART [11] and ID3 [12], the latter being followed by the later versions C4.5 [3] and C5.0 [13]. The basic strategy that is employed when generating decision trees is called recursive partitioning, or divide-and-conquer. It works by partitioning the examples by choosing a set of conditions on an independent variable (e.g., the variable has a value less than a particular threshold, or a value greater or equal to this threshold), and the choice is usually made such that the error on the dependent variable is minimized within each group. The process continues recursively with each subgroup until certain conditions are met, such as that the error cannot be further reduced (e.g., all examples in a group belong to the same class). The resulting decision tree is a graph that contains one node for each subgroup considered, where the node corresponding to the initial set of examples is called the root, and for all nodes there is an edge to each subgroup generated from it, labeled with the chosen condition for that subgroup. A decision tree is normally depicted with the root at the top having the ancestor nodes below it – see Fig. 2 for a decision tree generated within the domain of pesticide classification. An example is classified by the tree by following a path from the root to a leaf node, such that all conditions along the path are fulfilled by the example, where the conditions are formed from the variable names directly below each node and from the edge labels (e.g., the condition `B01-nCIC ≤ 1.5` follows from the root and the leftmost edge from the root). The estimated class probabilities at the reached leaf node are used to assign the most probable class to the example. The relative sizes of the sectors of each pie chart in Fig. 2 correspond to estimated class probabilities, where a dark sector corresponds to the probability of belonging to the class of pesticides, and a light sector corresponds to the

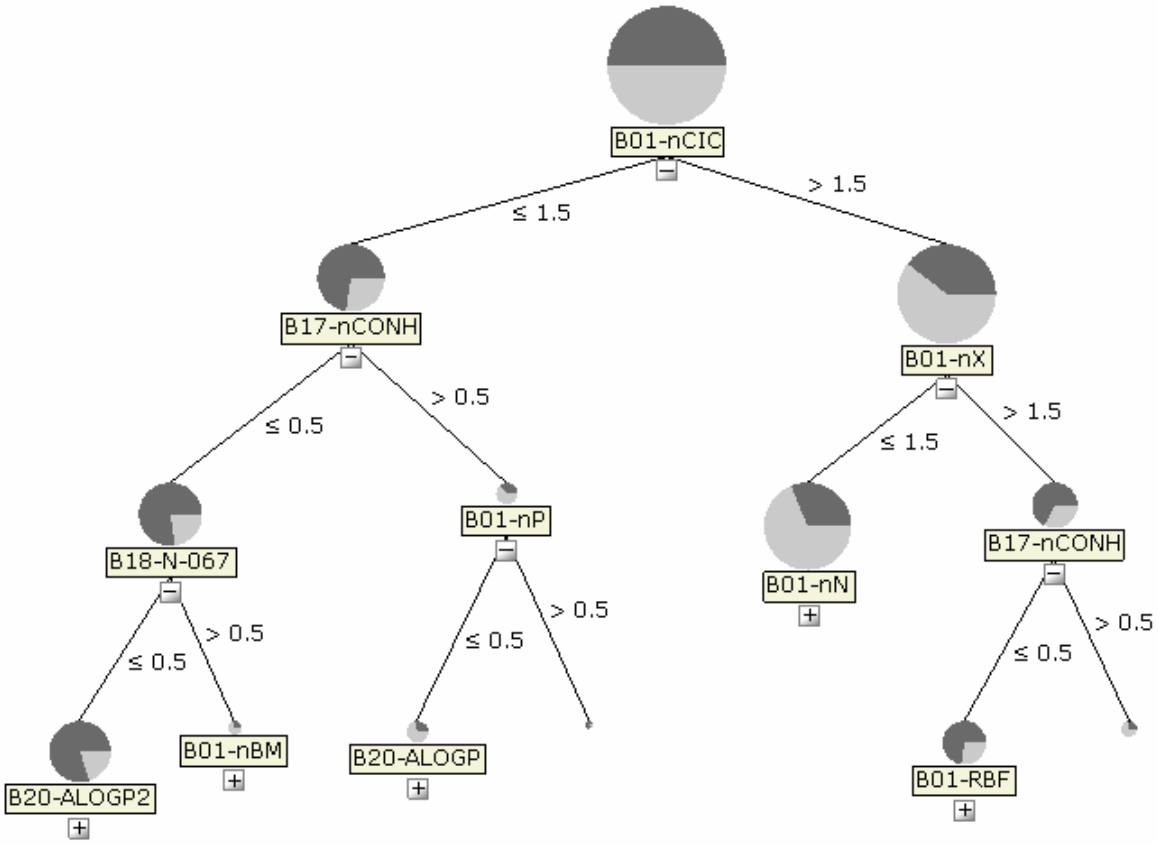

Figure 2. Example of a decision tree for classifying molecules as pesticides (dark color) or non-pesticides (light color) based on chemical descriptors.

probability of belonging to the class of non-pesticides[1].

Decision trees have many attractive features, such as allowing for human interpretation and hence making it possible for a decision maker to gain insights into what factors are important for particular classifications. However, recent research has shown that significant improvements in predictive performance can be achieved by generating large sets of models, or *ensembles*, which are used to form a collective vote on the value for the dependent variable [14]. It can be shown that as long as each single model performs better than random, and the models make independent errors, the resulting error can in theory be made arbitrarily small by increasing the size of the ensemble. However, in practice it is not possible to completely fulfill these conditions, but several methods have been proposed that try to approximate independence, and still maintain sufficient accuracy of each model, by introducing randomness in the process of selecting examples and conditions when building each individual model. One popular method of introducing randomness in the selection of training examples is bootstrap aggregating, or bagging, as introduced by Breiman [15]. It works by randomly selecting n examples with replacement from the initial set of n examples, leading to that some examples are duplicated while others are excluded. Typically a large number (at least 25-50) of such sets are sampled from which each individual model is generated. Yet another popular method of introducing randomness when generating decision trees is to consider only a small subset of all available independent variables at each node when forming the tree. When combined with bagging, the resulting models are referred to as *random forests* [16], and these are widely considered to be among the most competitive and robust of current methods for predictive data mining. The drawback of ensemble models are however that they can no longer be easily interpreted and hence provide less guidance into how classifications are made.

3.2 Feature and classifier fusion

Feature fusion concerns how to generate and select a single set of features for a set of objects to which several sets of features are associated. The purpose of feature fusion is typically to obtain a representation that allows for more effective analysis (cf. fusing pixel information into segments to improve image classification).

[1]The plus and minus signs below each node indicate whether (+) or not (-) the tree could be further expanded in the particular tool that has been used to create the tree [19].

Normally the fused vector results in loss of information. In case all objects have the same sets of features associated to them, one could however perform feature fusion with no loss of information by simply concatenating the feature vectors. Although such a concatenation may clearly not be the most effective method for high-dimensional feature vectors, it may in fact be the most effective alternative for low-dimensional data. Since we in this study only are concerned with moderately-sized feature vectors[2], this method was chosen as a base-line for feature fusion.

The combination of multiple classifiers generated from different sources or in different ways into a global model is often referred to as *classifier fusion* [17]. The purpose of classifier fusion is to improve predictive performance, typically by combining the output of each of the fused classifiers, where the output either is a single class label or a probability distribution over all class labels[3]. One may consider classifier fusion to be a special case of *decision fusion* [18], since the latter normally is not restricted to combining multiple decisions only, but also measures of confidence, probabilities etc.

A large number of approaches have been proposed for decision fusion in general, and for classifier fusion in particular. The latter include both ways of combining the output from multiple classifiers and ways of choosing classifiers to include in the combination, see [17] for an extensive characterization of different techniques for classifier fusion.

In this work, we have chosen to use *Bayes average* as a base-line method for obtaining a class probability distribution from the fused classifiers:

$$P_A(x \in C \mid x) = \frac{\sum_{k=1}^{K} P_k(x \in C \mid x)}{K}$$

where x is an example to be classified, C is one of the classes it may belong to, and P_k, k=1, …, K, are the probability distributions for the K classifiers.

4 Empirical Evaluation

We first state the experimental hypothesis, then describe the experimental setting, and finally present the results together with conclusions.

4.1 Experimental hypothesis

The null hypothesis of this experiment is that the choice of fusion strategy (fusing descriptors before building a model vs. fusing models built from each set of descriptors) has no impact on the predictive performance.

4.2 Experimental setting

We took the entire data set of 3221[4] compounds, of which 50% were classified as pesticides and 50% as non-pesticides, and performed a stratified split into two sets – one consisting of 70% (2255 compounds) to be used for training (model construction) and the other consisting of 30% (966 compounds) for testing. The size of the test set was considered to be sufficiently large by this division to allow detection of any significant differences between the strategies – if the original set of compounds would have been significantly smaller, cross-validation could have been considered as an alternative.

Two techniques for generating predictive models were considered – decision trees and ensembles (random forests) as implemented in the Rule Discovery System, v. 2.5.1 [19]. The former is an example of a technique resulting in a comprehensible model, while the latter results in what can be considered to be a non-comprehensible model (each ensemble in the experiment consists of 50 non-pruned trees).

We considered generating models from each of the four sets of descriptors (B01, B17, B18 and B20) as well as generating a fused model by averaging the class probabilities from each individual model. The resulting models were compared to the model obtained by first fusing all descriptors into a global set, and then generating a predictive model.

[2]The feature vector obtained by concatenating all four molecular descriptor sets contains 348 features.

[3]The methods for generating ensemble models described in the previous section are hence examples of methods for classifier fusion.

[4]Of the original set of 3226 compounds, 5 were removed due to failure to calculate molecular descriptors.

Tree model	B01	B17	B18	B20	Fused Model	Fused Descriptors
No. errors	245	266	250	298	194	236
Accuracy	74.64	72.47	74.12	69.15	79.92	75.57
AUC	0.8244	0.8066	0.8231	0.7564	0.8866	0.8292
No. rules	74	80	76	85	315	48

Table 1. Results for the decision tree method.

Ensemble model	B01	B17	B18	B20	Fused Model	Fused Descriptors
No. errors	200	184	149	275	164	109
Accuracy	79.30	80.95	84.58	71.53	83.02	85.09
AUC	0.8768	0.9041	0.9219	0.8170	0.9167	0.9227
No. rules	16253	14440	17368	18393	66454	15256

Table 2. Results for the ensemble method.

4.3 Experimental results

The predictive performance on the 966 test examples for the decision tree method is presented in Table 1, where the number of errors, accuracy, AUC and number of rules (leaf nodes) is presented for each set of descriptors (B01, B17, B18 and B20) as well as for the fused model, and the model generated from the fused descriptor set.

The fused model clearly outperforms each individual model generated from a single set of descriptors when comparing the predictive performance (when measured both as accuracy and AUC). The fused model furthermore significantly outperforms the model generated from the fused descriptor set – hence demonstrating that the choice of strategy indeed may have an impact on the resulting predictive performance. According to McNemar's test, the double-sided tail probability for the observed difference in accuracy is 0.006, hence allowing the null hypothesis to be safely rejected.

The improved performance is however associated with a cost regarding the comprehensibility, since the size of the fused model is equal to the sum of the sizes of the included models. Interestingly, the tree generated from the fused set of descriptors is the smallest (and still slightly more accurate than the models built from separate descriptor sets). This might be due to that more compact models can be found when combinations of descriptors from different sources are allowed. However, another possible explanation is that the increased dimensionality may lead to more extensive pruning, since the risk of being mislead by spurious correlations when growing the tree increases with the number of dimensions.

The predictive performance on the 966 test examples for the ensemble method is presented in Table 2, where again the number of errors, accuracy, AUC and number of rules (leaf nodes) is presented for each set of descriptors (B01, B17, B18 and B20) as well as for the fused model, and the model generated from the fused descriptor set.

In contrast to the results for the decision tree method, building an ensemble model from a fused set of descriptors is more effective than fusing ensemble models built from the individual descriptor sets. According to McNemar's test, the double-sided tail probability for the observed difference in accuracy is 0.012. This means that the null hypothesis can again be safely rejected – the choice of strategy is indeed important – but this time the best strategy is to fuse the descriptor sets. It should also be noted that in contrast to the experiment with decision trees, the model built from one of the individual descriptor sets (B18) actually outperforms one of the fusion strategies.

5 Concluding remarks

We have investigated two strategies for fusing information from multiple sources when generating predictive models in the domain of pesticide classification – by fusing features from all sources before building a model and by fusing models built from the individual sources. The experiment demonstrated that the choice of strategy does indeed have an impact on the predictive performance. In case the models consist of decision trees, it is clearly more beneficial to combine the predictions from the individual models instead of generating a single model from the fused set of descriptors. However, when the models consist of ensembles of decision trees, it is significantly more effective to build the model from a fused set of descriptors than fusing ensemble models built from the individual sources. Whether this finding holds also for other domains remains to be investigated, but this study has nevertheless demonstrated that significant gains in predictive performance can be obtained by considering different fusion strategies for different types of model. This finding also calls for investigating more elaborate fusion methods than the two basic strategies considered in this study, both regarding fusion of features and fusion of classifiers (cf. [17]). Another line of future research concerns investigating what effect the choice of fusion strategy has on comprehensibility (not only counting the number of rules) and how this affects the possibilities of gaining new insights in the domain of application.

Acknowledgements

Many thanks to Dr. Sorel Muresan, AstraZeneca R&D, Mölndal, Sweden, for providing the data sets, together with descriptions of these, as well as for valuable discussions.

This work was supported by the Information Fusion Research Program (www.infofusion.se) at the University of Skövde, Sweden, in partnership with the Swedish Knowledge Foundation under grant 2003/0104.

References

[1] Witten I. and Frank E., *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*, Morgan Kaufmann (2005)

[2] Provost F., Fawcett T. and Kohavi R., "The case against accuracy estimation for comparing induction algorithms", Proc. Fifteenth Intl. Conf. Machine Learning, (1998) 445-553

[3] Quinlan J.R., *C4.5: Programs for Machine Learning*, Morgan Kauffman, 1993

[4] Hastie T., Tibshirani R. and Friedman J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag (2001)

[5] Tomlin C.D.S. (Ed.), *The e-Pesticide Manual*, 13th edition, BCPC Publications (2003)

[6] James C. A., Weininger D. and Delaney J., *Daylight Theory Manual*, Daylight Chemical Information Systems, http://www.daylight.com/dayhtml/doc/theory/index.html (accessed Feb. 21, 2007)

[7] DRAGON for Windows (Software for Molecular Descriptor Calculations) version 5.4 - 2006, Talete SRL, http://www.talete.mi.it (accessed Feb. 21, 2007)

[8] Ghose A. K., Viswanadhan, V. N. and Wendoloski J. J., "Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods", *Journal of Physical Chemistry A*, vol. 102 (1998) 3762-3772.

[9] Sadowski J. and Kubinyi H., "A scoring scheme for discriminating between drugs and nondrugs.", *J. Med. Chem.*, vol. 41 (1998) 3325-3329.

[10] Muresan S. and Sadowski J., "'In-House likeness': Comparison of large compound collections using artificial neural networks", *J. Chem. Inf. Model.*, vol. 45 (2005) 888-893

[11] Breiman L., Friedman J.H., Olshen R.A. and Stone C.J., *Classification and Regression Trees*, Wadsworth (1984)

[12] Quinlan J.R., "Induction of decision trees", *Machine Learning*, vol. 1 (1986) 81-106

[13] Quinlan J.R., "Data Mining Tools See5 and C5.0", http://www.rulequest.com/see5-info.html (accessed Feb. 21, 2007)

[14] Bauer E. and Kohavi R., "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants", *Machine Learning*, vol. 36 (1999) 105-139

[15] Breiman L., "Bagging Predictors", *Machine Learning*, vol. 24 (1996) 123-140

[16] Breiman L., "Random Forests", *Machine Learning*, vol. 45 (2001) 5-32

[17] Ruta D. and Gabrys B., "An Overview of Classifier Fusion Methods", *Computing and Information Systems*, vol. 7 (2000) 1-10

[18] Dasarathy B.V., *Decision Fusion*, IEEE Computer Society Press (1994)

[19] Rule Discovery System, v. 2.5.1, Compumine AB, http://www.compumine.com/web/public/rds (accessed Feb. 21, 2007)