Feature Selection for Descriptor Based Classification Models. 1. Theory and GA-SEC Algorithm

Jörg K. Wegner,* Holger Fröhlich, and Andreas Zell

Zentrum für Bioinformatik Tübingen (ZBIT), Universität Tübingen, Sand 1, D-72076 Tübingen, Germany

Received October 24, 2003

The paper describes different aspects of classification models based on molecular data sets with the focus on feature selection methods. Especially model quality and avoiding a high variance on unseen data (overfitting) will be discussed with respect to the feature selection problem. We present several standard approaches and modifications of our Genetic Algorithm based on the Shannon Entropy Cliques (GA-SEC) algorithm and the extension for classification problems using boosting.

INTRODUCTION

This work gives an introduction to the creation of classification models on molecular data sets and testing their model quality. The focus lies on the *NP-complete*[1] feature (descriptor) selection problem, with a special emphasis on aspects most interesting for the chemoinformatics, bioinformatics, and machine learning community.

This paper is divided into two main parts. A theory part, giving a short introduction of the basic principles of hypothesis testing and the generalization error.[2−5] The notion of entropy, especially in the context of machine learning, is explained, which is an important element of the presented GA-SEC hybrid-feature selection algorithm.[6] Furthermore, the difference between *feature extraction* and *feature selection* is presented and also the *feature selection-filter* and *-wrapper* methods used in our work. Finally, the creation of ensemble models, which can be used in combination with our feature selection algorithm, conclude the theory part. The second part explains some relevant points to be considered, when preparing chemical data sets and describes the new GA-SEC variants for classification in more detail.

In general, there exist 2^Npossibilities to pick an optimal feature subset of any size, where $N = |D|$ is the number of descriptors in the data space D, which is also called the descriptor space. When choosing a descriptor subset D_s of size $N_s = |D_s|$ out of a descriptor set of size N requires the evaluation of

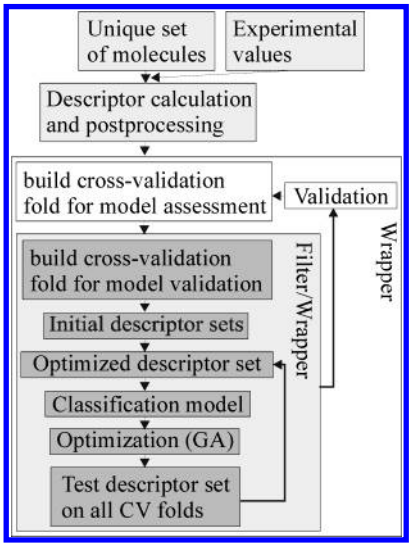
$$\binom{N}{N_s} = \frac{N!}{N_s!(N - N_s)!} \tag{1}$$

subsets.[7]

The complete glossary of mathematical symbols used and two huge QSAR feature selection benchmark data sets can be found in the Supporting Information.

There exist two basic principles for feature selection. One is the *filter approach*, which picks only a good feature subset once, the other is the *wrapper approach*, which tries to optimize the feature subsets by solving the *combinatorial optimization* problem.[8] We will present filter and wrapper methods for the feature selection used in our work and our modified GA-SEC hybrid-feature selection algorithm.[6] Scheme 1 shows the modified QSAR paradigm[8,9] addressing the feature selection problem using a combined *filter* and *wrapper approach*. It can be seen that the *wrapper approach* contains two additional loops, the optimization loop for picking an optimal descriptor set and the validation loop for the model assessment. So it is obvious, that greedy *filter approaches* are much faster than *wrapper approaches*. *Wrappers* utilize the learning machine of interest as a black box to score subsets of variables according to their predictive power. *Filters* select subsets of variables as a preprocessing step, independently of the chosen predictor.

Scheme 1. Modified QSAR Paradigm[8,9] with Focus on Feature/ Descriptor Selection[a]



[a] The inner loop is necessary for optimizing the feature set and the outer validation loop for assessing the model quality.

THEORY − INDUCTION

Hypothesis Testing. Inductive inference or hypothesis testing may be expressed as the following: ***given a data set D and a set of hypotheses H, choose the hypothesis that***

* Corresponding author phone: +49-7071-2976455; fax: +49-7071-29-5091; e-mail: wegnerj@informatik.uni-tuebingen.de.

10.1021/ci0342324 CCC: $27.50
Published on Web 02/13/2004

best explains the data. Four approaches to the problem of multiple hypotheses are common today:[2,10] Epicurus' principle of multiple explanations; Occam's principle of the simplest explanation (known as Occam's razor); Bayesian inference;[2,10] and Vapnik's structural risk minimization.[4] The first two principles are used for the initialization of our feature selection algorithm, creating model ensembles and selecting the models.

Epicurus' Principle of Multiple Explanations: *if more than one theory is consistent with the data, keep them all*. The Greek philosopher of science Epicurus maintained that if several explanations are equally in agreement with a phenomena, we must keep them all for two reasons.[2] First, by making use of multiple explanations it may be possible to achieve a higher degree of precision. Second, it would be unscientific to choose one explanation over another when both explain the phenomena equally well.

The Principle of indifference considers events to be equally probable if we have not the slightest knowledge of the conditions under which each of them is going to occur. For the case of a die, this actually coincides with the so-called "maximum entropy principle",[11] which states that we should choose probabilities p_i for face I to be the outcome of a trial, $I = 1,2, ..., 6$ such that $-\sum p_i \log p_i$ is maximized under the only constraint $\sum p_i = 1$. Here we obtain precisely $p_i = 1/6$.

Occam's Razor Principle: Let the generalization error of a model be its error rate on unseen examples, and the training-set error be its error on the examples it was derived with. Then the formulation of the razor that is perhaps closest to Occam's original intent is as follows:[3] *Given two models with the same generalization error, the simpler one should be preferred because simplicity is desirable in itself*. Vapnik's[4] theory of structural risk minimization shows that the generalization ability of a class of classification models is not a function of the number of features but of its *VC dimension* (capacity). The *VC dimension* of a hypothesis class H over an instance space D is the size of the largest subset of D for which H can generate all possible binary labelings. Although the number of features are sometimes related, in general they are not.

Generalization Error. Because feature selection is a particular form of model selection: *"...the good practice of dividing the available data into separate training and test sets should not be forgotten"*.[12] Given a finite data set M, we would like to estimate the future performance of a classifier induced by the given data set for getting a good tradeoff between bias and variance (Bias-Variance-Decomposition).[5,8]

The bias of a method that estimates a parameter $\hat{c}_j = \hat{f}(\vec{m}_j)$ (predicted value) using the inducer $\hat{f}(\vec{m}_j)$ is defined as the expected estimated value ($E[\hat{c}_j]$) minus the value of $c_j = f(\vec{m}_j)$ (true value)

$$Bias(M,\vec{m}_j) = E[\hat{c}_j] - c_j = E[\hat{f}(\vec{m}_j)] - f(\vec{m}_j) \qquad (2)$$

where $\vec{m}_j$ is a descriptor set for a molecule with index j. An unbiased estimation method is a method that has zero bias. The variance of a method is the statistical variance of the estimate:

$$Var(M,\vec{m}_j) = E[(\hat{f}(\vec{m}_j) - E[\hat{f}(\vec{m}_j)])^2] \qquad (3)$$

If M is the input to the inducer (training set) the effectiveness of the inducer at $\vec{m}_j$ is

$$(\hat{f}(\vec{m}_j) - f(\vec{m}_j))^2 = (c_j - f(\vec{m}_j))^2 \qquad (4)$$

Taking the performance with respect to the training set M (i.e., averaging over all possible training sets of the given size) we obtain the expected risk

$$Risk(M,\vec{m}_j) = E[(c_j - \hat{f}(\vec{m}_j))^2 | \vec{m}_j] \qquad (5)$$

which represents the ability to yield a good performance for all the possible situations (all c_j, $\vec{m}_j$ pairs) and is thus called the generalization error. To assess the generalization error of a given model independently of the training set, it is suitable to consider the expected generalization error:

$$Y = f(M,\vec{m}_j) + \epsilon$$

$$E(\epsilon) = 0$$

$$Var(\epsilon) = \sigma_\epsilon^2$$

$$Risk(M,\vec{m}_j) = E[(c_j - \hat{f}(\vec{m}_j))^2 | \vec{m}_j]$$

$$Risk(M,\vec{m}_j) = \sigma_\epsilon^2 + [E[\hat{f}(\vec{m}_j)] - f(\vec{m}_j)]^2 + E[(\hat{f}(\vec{m}_j) - E[\hat{f}(\vec{m}_j)])^2]$$

$$Risk(M,\vec{m}_j) = IrreducibleError + Bias^2(M,\vec{m}_j) + Var(M,\vec{m}_j) \qquad (6)$$

Accordingly, the generalization error can be separated into three components: the irreducible noise level, the squared bias induced by the choice of a model, and the variance coming from the data sampling. A large bias causes simple models with a low variance. On the other side a small bias, following the training points almost exactly, can cause a high variance for unseen data (overfitting).

The often used ***holdout*** method[6] divides the data into two mutually exclusive subsets called the training and the test set or holdout set. Because this method causes a high bias for a small test set or a wide confidence interval for a huge test set, we will use the ***k-fold-cross-validation*** method.[8] The data set M is randomly split into k mutually exclusive subsets (the folds) M_1, M_2, ..., M_k of approximately equal size. The inducer is trained and tested k times; each time $t \in \{1,2,...,k\}$, it is trained on $M \backslash M_t$ (M without M_t) and tested on M_t. If M_t is the test set which includes the instance $I_j = \langle \vec{m}_j, c_j \rangle$, then the accuracy is

$$acc_{CV}(M) = \frac{1}{m} \sum_{I_j \in M} \delta(f(M \backslash M_t, \vec{m}_j), c_j) \qquad (7)$$

with $\delta(i,j) = 1$ if $i = j$ and 0 otherwise.

Empirical tests have shown that for model selection 5-fold or 10-fold cross-validation gives a good tradeoff between bias and variance.[8] ***Leave-one-out*** (LOO) ***cross-validation***, where the number of folds is equal to the number of samples available, can be used in the inner loop (Scheme 1) to guide the search of the feature selection, but it should not be used to compare feature selection methods.[8,12] It was also shown

that cross-validation mostly outperforms the metric-based model selection methods, and there could be a benefit when using the meta-model-selection methods combining the cross-validation and the metric based model selection methods with lower variance.[13]

THEORY − ENTROPY

The entropy principle can be found in publications covering the information content,[6,14−17] the relation to statistical thermodynamics,[18] analytical chemistry,[19,20] crystal structure estimation,[21] information theory,[11,22,23] text classification,[24] dimensionality reduction,[25,26] and general machine learning algorithms.[27,28]

Entropy and Information Theory. For characterizing the information content of discrete probability distributions $P_i = \{p_{i,1},p_{i,2},...,p_{i,k}\}$ for one descriptor $\vec{d}_i$ we can use a generalized entropy measure, like Rényi's entropy[11,28,29]

$$H_{R\alpha}(P_i) = \frac{1}{1-\alpha}\log\left(\sum_{k=1}^{B} p_{i,k}^{\alpha}\right),\ \alpha > 0,\ \alpha \neq 1 \tag{8}$$

Using the L'Hospital's rule[30] $\lim_{\alpha\to\alpha_0} f(\alpha)/g(\alpha) = \lim_{\alpha\to\alpha_0} f'(\alpha)/g'(\alpha)$, using $\alpha_0 = 1$, we obtain the relation between the Shannon entropy $H_{SE}(P_i)$ and the Rényi entropy $H_{R\alpha}(P_i)$:

$$H_{SE}(P_i) = \lim_{\alpha\to 1} H_{R\alpha}(P_i) = -\sum_{k=1}^{B} p_{i,k}\log_2 p_{i,k}$$

$$H_{R\alpha}(P_i) \geq H_{SE}(P_i) \geq H_{R\beta}(P_i), \text{ if } 1 > \alpha > 0 \text{ and } \beta > 1 \tag{9}$$

Applied to the information content of a discretized descriptor $\vec{d}_i$ into B bins we obtain

$$H_{SE}(\vec{d}_i,B) = -\sum_{k=1}^{B} p(\vec{d}_i(k))\log_2(\vec{d}_i(k))$$

$$p_{ik} = p(\vec{d}_i(k)) = c_{ik}/\sum_{l=1}^{M} c_{il} \tag{10}$$

where p_{ik} is the probability of a data point or "count" c_{ik} to adopt a value within a specific data interval k with B bins and M is the number of molecules. Here we chose $B = 20$. In this fashion, $H_{SE}(P_i)$ values for different data sets can be directly compared, provided a uniform binning scheme can be defined. As discussed in previous papers $H_{SE}(P_i)$ values alone may not be sufficient to select descriptors with significant discriminatory power.[14] Furthermore this method is characterized by a strong tendency to oversample remote areas of the feature space and to produce unbalanced designs.[17]

A direct way for comparing two discrete probability distributions $P_i = \{p_{i,1},p_{i,2},...,p_{i,k}\}$ and $P_j = \{p_{j,1},p_{j,2},...,p_{j,k}\}$ is to use the generalized Jensen's measures of directed divergence[11] to avoid highly correlated features, which are in this case two discrete descriptor distributions (histograms)

$$D_{J,\lambda}(P_i,P_j) = H_E(\lambda P_i - (1-\lambda)P_j) - \lambda H_E(P_i) - (1-\lambda)H_E(P_j) \tag{11}$$

with $D_{J,\lambda}(P_i,P_j) \geq 0$ and $D_{J,\lambda}(P_i,P_j) = 0$ if $P_i = P_j$ and H_E is an entropy measure. When we use the Shannon entropy as an entropy measure, we obtain the generalized Jensen−Shannon measure of directed divergence[11,23,24] and for $\lambda = 1/2$ the Jensen−Shannon measure

$$D_{JS}(P_i,P_j) = H_{SE}\left(\frac{1}{2}(P_i - P_j)\right) - \frac{1}{2}(H_{SE}(P_i) - H_{SE}(P_j)) \tag{12}$$

which is a well known definition and is an analogue to the recently introduced differential entropy definition.[6,14,15]

The success for picking the best feature set with $|D| + 1$ features is less certain than using $n = |D|$ features, under the assumption that the picking probability is uniformly distributed over g experiments:[22]

$$H_{SE}(p(d_1^{(n)}), \cdots, p(d_g^{(n)})) \leq H_{SE}(p(d_1^{(n+1)}), \cdots, p(d_g^{(n+1)}))$$

$$H_{SE}\left(\frac{1}{n}, ..., \frac{1}{n}\right) \leq H_{SE}\left(\frac{1}{n+1}, ..., \frac{1}{n+1}\right)$$

$$-\sum_{i=1}^{g}\frac{1}{n}\log\frac{1}{n} \leq -\sum_{i=1}^{g}\frac{1}{n+1}\log\frac{1}{n+1} \tag{13}$$

In other words, the average entropy rates of subsets $H_{SE}(D_s)$ in bits of a randomly drawn subset with $|D_s|$ features decreases monotonically with the size of the subsets. If one feature is more probable than others, the result of the experiment is less uncertain

$$\sum_{i=1}^{g-1}\frac{1}{n}\log\frac{1}{n} + \frac{1}{m}\log\frac{1}{m} \leq \sum_{i=1}^{g}\frac{1}{n}\log\frac{1}{n}$$

$$\frac{1}{m}\log\frac{1}{m} \leq \frac{1}{n}\log\frac{1}{n} \tag{14}$$

where n is the number of times that the feature is picked over g experiments, m is the number of times of the feature which was picked more often ($m > n$). In other words, we will need less decisions for picking the final feature set, if we start with a more probable feature. We can avoid a huge success uncertainty when starting with a small number of features and using features which are more probable to be selected.

Unfortunately this does not grant the best feature set and does not guarantee the best hypothesis.[3] So following Shannon's first theorem,[22] good models can be found with a higher probability, but it does not make them more predictive.[10] There can still exist a hypothesis with higher generalization ability with using much more features.[3,4] Practical algorithms, however, are not given access to the underlying distribution, because we are still using only a sampling of the real world distribution, so most practical algorithms attempt to fit the data by solving the *NP complete* optimization problems of finding the smallest feature set with the highest generalization ability,[8] e.g. using genetic algorithms[6,31] or tabu search.[32,33]

Additionally, following Jensen's inequality for convex functions[22]

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2),\ 0 \leq \lambda \leq 1 \tag{15}$$

we obtain the chain rule for entropies

$$H_{SE}(p(\vec{d}_1^*), \cdots, p(\vec{d}_{N_s*}^*)) \leq \sum_{i=1}^{N_s^*} H_{SE}(p(\vec{d}_i^*)) \qquad (16)$$

with equality if and only if the $\vec{d}_i^*$ are independent. $N_s^* = |D_s^*|$ is the number of features in the best feature set found D_s^*, $\vec{d}_i^*$ is the descriptor i in D_s^*. Hence when initializing our feature selection algorithm with the entropy values of the single features (right-hand of eq 16), the uncertainty can be much greater than the entropy of the best feature set found $H_{SE}(p(\vec{d}_i^*), \cdots, p(\vec{d}_{N_s*}^*))$, when features depend on each other.

THEORY − FEATURE EXTRACTION/SELECTION

There exist two ways for reducing the data input space. One is to extract features by building linear and nonlinear combinations of a lower dimension of the input features which is called *feature extraction*. The other alternative is to select the features with respect for their generalization ability, which is called *feature selection*.

Feature Extraction. Feature extraction is not the topic of this paper, but we will give a short overview to show the principle difference between these methods. In *feature extraction*, we try to find the best linear or nonlinear combination of features to fulfill a dimensionality reduction criteria, e.g. in PCA each eigenvalue represents a portion of variation in the data and the eigenvalues are ranked by their ability to account for the variation in the data.[34]

There exists feature extraction[34−37] methods using Rényi entropy,[26] principal component analysis (PCA),[38] PCA-GA,[31,32,39] PCA-SA,[40] CROMsel,[41] hierachical discriminant regression (HDR),[42] independent component analysis (ICA),[43] multidimensional scaling (MDS),[44] nonlinear mapping (NLM),[45,46] partial least squares (PLS),[47−53] and kernel PCA.[54,55]

Feature Selection: Introduction. The problem of *feature selection* is that of finding a subset of the original features of a data set,[56−58] such that an induction algorithm that is run on data containing only these features generates a classifier with the highest possible accuracy. Overviews[8,56,59] over feature selection are already available, and the problem of overfitting for feature selection[8,12] was already addressed. Typical algorithms use genetic algorithms (GA),[6,60−62] support vector machines (SVM),[63−65] entropy,[23,24] decision trees (recursive partitioning),[66,67] tabu search,[32,33] stochastic proximity embedding (SPE),[68] model selection metrics,[13] artificial neural networks (ANN),[69] grafting,[70] multitask learning (MTL),[71] feature rankings,[57,58] and text classifiers.[24,72,73]

From a purely theoretical standpoint, the question is not of much interest.[3] The optimal Bayes rule is monotonic, i.e., adding features cannot decrease the accuracy, and hence restricting the induction algorithm to a subset of features is never advised. From the practical standpoint this problem is highly interesting, because feature subset selection is a *NP complete* problem.[1] So the objective is 5-fold:[12,70] first, improving the prediction performance of the predictors; second, providing a better understanding of the underlying process; third, providing faster and more cost-effective predictors. [Not all model types have an improved performance for a smaller feature set used, because they may depend only on the number of instances/molecules used.[4,54,55]]; fourth, serving as a bridge between the harsh reality of the real world, and the cozy idealistic environments inhabited by most machine learning algorithms;[12] fifth, avoiding irrelevant features for similarity analysis. For example, it was shown that inferior similarities were obtained, when fingerprints were applied on the complete molecule and not only on the biologically relevant substituents.[74,75]

Feature Selection: Filter Approach. For a numeric attribute, the feature must first be discretized into several intervals, using, for example, the entropy-based discretization method,[27] because we want to compare discrete probability distributions using the entropy. The number of the features to select must be defined.

The *information gain* (mutual information) $I_G(\vec{d}_i,\vec{c})$ evaluates the worth of a feature $\vec{d}_i$ with respect to the class information $\vec{c}$. The related *gain ratio* measure $I_R(\vec{d}_i,\vec{c})$ and *symmetrical uncertainty* $I_{SU}(\vec{d}_i,\vec{c})$ uses other entropy based normalization factors.[58]

$$I_G(\vec{d}_i,\vec{c}) = H_{SE}(\vec{d}_i) - H_{SE}(\vec{d}_i|\vec{c}) = H_{SE}(\vec{c}) - H_{SE}(\vec{c}|\vec{d}_i) \qquad (17)$$

$$I_R(\vec{d}_i,\vec{c}) = \frac{I_G(\vec{d}_i,\vec{c})}{H_{SE}(\vec{d}_i)} \qquad (18)$$

$$I_{SU} = \frac{2I_G(\vec{d}_i,\vec{c})}{H_{SE}(\vec{c}) + H_{SE}(\vec{d}_i)} \qquad (19)$$

Because decision trees (recursive partitioning)[27] also use discrete features and the *information gain* $I_G(\vec{d}_i,\vec{c})$ as decision criteria, they can be used for feature selection, also.[66,67]

The Relief algorithm[76,77] assigns a relevance weight R_f to each feature, which is meant to denote the relevance of the feature for the target concept. It is a randomized algorithm which finds all weakly relevant features but does not help with redundant features. The OneR algorithm[78,79] can be regarded as a one level decision tree, which tests only one attribute and ranks features with R_{oneR}. Another method to measure the association between two features in a contingency table is based on the ***Chi-squared test***[80−82]

$$\chi^2 = \sum_{j=1}^{c'}\sum_{k=1}^{B} \frac{(c'_{jk} - E_{jk})^2}{c'_{jk}} \qquad (20)$$

where B is the number of intervals, c' is the number of classes, and $E_{ij} = (c'_{*k} \cdot c'_{j*})/|M|$ is the expected frequency of c'_{jk} where c'_{jk} is the number of samples in the kth interval and the jth class, c'_{*k} is the number of samples in the kth interval, c'_{j*} is the number of samples in the jth class, and $|M|$ is the total number of samples (here molecules). Larger χ^2 values reflect more important features. The degree of freedom is $(B - 1)(c' - 1)$. For instance, if $B = 2$ (binary feature) and $c' = 2$ (binary classification problem), the degree of freedom is one and the χ^2 value at the 5% significance level is 3.841. If our χ^2 value is larger than that, the probability is less than 5% that discrepancies this large are attributable to chance, and we are led to reject the null hypothesis of independence between the feature and the class values.

Feature Selection: Wrapper Approach. A wide range of search strategies can be used,[56,83] including best-first,[25] simulated annealing,[40] and genetic algorithms.[6,62] We used the recently introduced Recursive Feature Elimination (RFE)[84] method, which uses Support Vector Machines (SVM),[4,54,55,85,86] to compare our GA-SEC algorithm with. Support Vector Machines (SVM) were initiated by Vapnik[4] with the introduction of the *structural risk minimization principle*, which defines a tradeoff between the approximation quality of a given data set and the complexity of the approximating function. The main aspect of a SVM is the "kernel-trick", which projects the data into a high dimensional (possibly infinite) feature space where a simple linear learning machine can be applied.[56,86] Recursive Feature Elimination (RFE)[84] is a wrapper method which performs a backward feature elimination: The idea is to find the $|D_s^*|$ features which lead to the largest margin of class separation. This combinatorial problem is solved in a greedy fashion. In the 2-class case the algorithm begins with the set of all features and successively eliminates the feature which induces the smallest change in the cost function

$$W^2(\alpha^*) = \sum_{i=1}^{|M|} \alpha_i^* - \frac{1}{2}\sum_{i,j=1}^{|M|} \alpha_i^*\, \alpha_j^* c_i c_j k(\vec{m}_i,\vec{m}_j) \qquad (21)$$

where α^* is the tuple of the Lagrangian multipliers which are obtained by solving the SVM problem and $k(\vec{m}_i,\vec{m}_j)$ is an entry in the kernel matrix. In the literature often the terms $x_j = \vec{m}_j$ and $y_j = c_j$ are used. As for SVM's W^2 is a measure of the predictive ability (and is inversely proportional to the margin), the algorithm at each step eliminates the feature which keeps this quantity small. Assuming that the change of the set of support vectors (and hence of the tuple α^*) when removing only one feature is negligible, this is done by performing the following iterative procedure over all features $|D|$:[63]

- Given a tuple α^* of Lagrangian multipliers as a solution of the SVM learning algorithm, calculate for each feature t

$$w_{(-t)}^2(\alpha^*) = \sum_{i,j=1}^{|M|} \alpha_i^*\, \alpha_j^* c_i c_j k(\vec{m}_i^{(-t)},\vec{m}_j^{(-t)}) \qquad (22)$$

where $\vec{m}_i^{(-t)}$ means that the tth feature from the training vector $\vec{m}_i$ has been removed.

- Remove the feature with the smallest value $D_w(t) = |w^2(\alpha^*) - w_{(-t)}^2(\alpha^*)|$ and retrain the SVM with the reduced set of features.

RFE originally was designed to solve 2-class problems only, but extensions to a multiclass versions are possible.[63] Finally, RFE computes a ranking of the selected features. The number of the features to select must be defined.

THEORY − ENSEMBLE MODELS

Building Models. Following Epicurus's principle and taking model diversity into account, we should combine multiple predictors, to obtain more precise results. Obviously, combining the output of multiple predictors is useful only if there is a disagreement between them, which follows Occam's razor to avoid multiple entities. Two principles are combining unweighted ensembles[87−91] or voting meta algorithms such as bagging[92] or boosting.[93]

Chart 1. Our Hash Code Calculation Method Uses the Modified Morgan Algorithm[100,101] Implemented in JOELib[99] Which Has Some Analogues to the Jochum−Gasteiger Canonical Numbering Algorithm[9,102] [a]

```
create canonicalized molecule using modified Morgan algorithm

hash := 31*(number of rotatable bonds) + (number of SSSR rings);

for all atoms

  hash := 31*hash + atomic number of actual atom;

  hash := 31*hash + heavy valence of actual atom;

  hash := 31*hash + implicite valence of actual atom;

  hash := 31*hash + 100*(partial charge of actual atom);
end
```

[a] The number of SSSR rings is the number of the smallest set of smallest rings.[101,107,108]

The ***Bagging*** algorithm (**B**ootstrap **agg**regat**ing**) votes classifiers generated by different bootstrap samples (replicates).[92]

Boosting was introduced for boosting the performance of weak classifiers. The most important boosting algorithm is AdaBoost.M1 (**Ada**ptive **Boost**ing) for two class problems, with variants called M2, MH for multiclass problems and MR for regression problems.[93] Like Bagging, the AdaBoost algorithm generates a set of classifers and votes them. The AdaBoost algorithm generates the classifers sequentially, while bagging can generate them in parallel. AdaBoost also changes the weights of the training instances provided as input to each inducer based on classifers that were previously built. [In difference feature selection can be regarded as a binary weighting of the features/descriptors.] It was already shown that boosting outperforms bagging[94,95] and that boosted decision trees (recursive partitioning)[87,96,97] perform as well as or close to support vector machines (SVM). Zhang et al. showed this on examples for classifying gene sequences.[98]

METHODS

Data Preparation. It is clear that all models we build should use a representative data set without duplicates, or our model will not have a valid generalization ability.[81] To eliminate duplicates we calculated the hash code for molecules based on the basic atom properties in JOELib[99] using a modified Morgan algorithm[100,101] with some analogues to the Jochum−Gasteiger canonical numbering algorithm.[9,102] Additionally we used the canonicalized SMILES[100,103−105] code for molecules with cis/trans and E/Z information for calculating the SMILES hash code. More complex hash code calculation methods can be applied to reduce the number of mappings of nonidentical molecules with the same hash code, but these statistical tests depend on the data set used.[106] Additionally, hash codes are limited to the size of the integer used, here to $2^{31} - 1$ possible unique numbers. Because we inspected molecules where a duplicate hash code occurred by the graph based equality of the atoms and bonds and then "by hand", avoiding multiple hash code mappings for nonidentical molecules can reduce work but cannot completely avoid detailed equality checks. For these data sets we did not find any noncorrect hash code mappings. Potentially similar molecules occur when the hash code for

Chart 2. The SMILES[100,103−105] Hash Code Uses Also the Modified Morgan Algorithm[100−102] for a Unique Renumbering of the Molecule, Which Contains Also E/Z and Cis/Trans Information and Is More Specific than the Plain Hash Code Which Does Not Use Stereo- or Chirality-Atom-Properties

```
us := create canonicalized/unique SMILES
for all characters in us
  hash := 31*hash + actual character in us;
end
```

two molecules is identical, so we inspected all instances of identity when this happened. We used well-known data sets for logP,[6,109,110] logS,[6,111−116] and Human Intestinal Absorption (HIA)[42,117,119] from the literature to show that only one of these three data sets contained no duplicate molecules.

We combined the test, training, and validation data sets and found 100 duplicates in the Wang data set and three more duplicates, because the molecular structures were wrong. The Huuskonen data set contained only four duplicates and contained six uracil derivates with wrong atom types caused by Corina's 3D structure generation.[120] The complete list of duplicate molecules and incorrect structures can be found in the Supporting Information. Especially the 84 duplicated molecules in the test set and the 15 duplicated molecules in the validation set of the LogP data set[6,109,110] lead to invalid numbers for the generalization ability.

For comparing models it should be guaranteed that the descriptors are using all the same atom type, aromaticity- and hybridization-model. Because many programs use text definitions for the atom types[99,121] we recommended using the same definitions or the same data processing workflow to avoid bad prediction results for new molecules.

Furthermore the data sets should not contain missing descriptor values or at least common techniques should be used to avoid missing values.[10] All descriptor values $\vec{d}_i$ were normalized.[117,118] We did this by using the z-transformed descriptor distribution with a mean of zero and a standard deviation of one[81]

$$\sigma_i = \sigma_i^{(|M|-1)} = \sqrt{\frac{\sum_{m=1}^{|M|} (d_{i,m} - \overline{d}_i)^2}{|M| - 1}} \tag{23}$$

$$\hat{d}_{i,m} = \frac{d_{i,m} - \overline{d}_i}{\sigma_i} \tag{24}$$

where i is the index number of the descriptor and m is the index number of the molecule, $\overline{d}_i$ is the average $d_{i,m}$ over all molecules $|M|$ and $\sigma_i = \sigma_i^{(|M|-1)}$ is the standard deviation of descriptor $\vec{d}_i$. $|D|$ is the number of all available features/descriptors. D and M are descriptor or molecule based formulations of the data set or instance space.

Feature Selection: Hybrid Approach. The GA-SEC algorithm[6] tries to avoid the often criticized "brute-force" method of standard-wrapper approaches[56] by taking the problem of overfitting also into account, caused by coarse search strategies.[12] So we use the Shannon entropy (as given in eq 10) to find features which contain a high information content. Then we apply a divergence measure by using the differential Jensen−Shannon entropy (as defined in eq 12) combined with a clique detection algorithm to find initial feature sets, called Shannon entropy cliques (SEC), which have a high information content and are little correlated to each other. When using a genetic algorithm (GA)[122,123] as wrapper and the SEC for initializing the GA population $P_{GA}(0)$ as filter, we obtain the already introduced hybrid GA-SEC feature selection algorithm.[6]

Table 1. Number of Duplicate and Wrong Molecules in Already Published LogS,[6,111−116] LogP,[6,109,110] and HIA[42,119] Data Sets[a]

	doublets/size			incorrect structures/size		
data set	train	test	validation	train	test	validation
LogS	2/1016	2/253	1/21	6/1016	0/253	0/21
LogP	1/1853	84/138	15/19	1/1853	2/138	0/19
HIA	0/67CV	0/9CV	0/10	0/67CV	0/9CV	0/10

[a] For the HIA data set eight fold cross-validation (CV) were used.

Because we ignored until now the class information of our classification problem, the variants presented here also use multiple filter (MF) approaches to improve the performance of the standard GA-SEC algorithm.

By taking feature information content, feature diversity, and feature prediction ability with respect to our classification problem into account, we try to avoid a high final feature set uncertainty (see eqs 8−16) for reducing the number of evaluation steps. A matrix $M_{adj}(i,j)$ is defined

$$M_{adj}(i,j) = \begin{cases} 1 & \text{if } M_{SE}(i,j) > SE_{cut} \text{ and } M_{JS}(i,j) > D_{cut} \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

where $M_{SE}(i,j) = H_{SE}(P_i) \cdot H_{SE}(P_j)$ is the quadratic information content, $M_{JS}(i,j) = D_{JS}(P_i,P_j)$ is the diversity information for a descriptor pair, SE_{cut} is the minimally allowed quadratic information content value, and D_{cut} is the minimally allowed Jensen−Shannon entropy value. The $M_{adj}(i,j)$ matrix can be used to find descriptors with a high information content which will correlate little with other descriptors by using a maximum clique detection algorithm.[124,125] A maximum complete subgraph (clique) is a complete subgraph that is not contained in any other complete subgraph.[124] Complete means that every node of the clique is connected to every other node of the clique. In our case a clique will be the descriptor subset which has a high information content and where every descriptor is maximally diverse to any other descriptor in this clique.

A subset-selection graph G_d is created from $M_{adj}(i,j)$ by using the Bron-Kerbosch (BK)[124] clique detection algorithm. We call the G_d subsets of the $M_{adj}(i,j)$ matrix the Shannon entropy cliques (SEC). The run time of clique detection algorithms depends strongly on the edge/node density in graphs and an overview was recently published.[126] If one chooses $SE_{cut} = 0$ and $D_{cut} = \infty$ the initialization is analogous to a standard genetic algorithm wrapper approach, with picking features randomly. An overview over the different GA-initialization methods is given in Table 2. *GA-SEC* is the algorithm we have presented already, without using any additional class information.[6] *GA-SEC-MF* (multiple filters) uses the default initialization of the population $P_{GA}(0)$and adds randomly the best descriptors found by the different filter approaches using the class information (see eqs 17−20). *GA-SEC-MFP* (multiple filters for prescreening) reduces the adjacency matrix for finding uncorrelated descriptors to the best descriptors found by the different filter approaches.

Table 2. Definitions of the Alternatives for the GA-SEC Algorithm[6]

$P_{GA}(0)$ initialization method (filter)	description
GA-SEC **G**enetic **A**lgorithm based on **S**hannon **E**ntropy **C**liques	1. $M_{adj}(i,j) = 1$, if $M_{SE}(i,j) > SE_{cut}$ and $M_{JS}(i,j) > D_{cut}$, otherwise $M_{adj}(P_i,P_j) = 0$ 2. $P_{GA}(0)$ from randomly picked G_d of size s_{Clique}
GA-SEC-MF **G**enetic **A**lgorithm based on **S**hannon **E**ntropy **C**liques and **M**ultiple **F**ilters	1. $P_{GA}(0)$ from randomly picked G_d of size s_{Clique} 2. And picking n_{Filter} randomly from p_{Filter} of the greatest $I_G(\vec{d}_i,\vec{c})$, $I_R(\vec{d}_i,\vec{c})$, $I_{SU}(\vec{d}_i,\vec{c})$, R_f, R_{OneR}, χ^2 values 3. Replace six individuals with six s_{Clique} best of $I_G(\vec{d}_i,\vec{c})$, $I_R(\vec{c}_i,\vec{c})$, $I_{SU}(\vec{d}_i,\vec{c})$, R_f, R_{OneR}, χ^2
GA-SEC-MFP **G**enetic **A**lgorithm based on **S**hannon **E**ntropy **C**liques and **M**ultiple **F**ilters for **P**rescreening	1. $M_{adj}(i,j) = 1$, if $M_{SE}(i,j) > SE_{cut}$, $M_{JS}(i,j) > D_{cut}$, and P_i or P_j is one of the n_{Filter} highest $I_G(d_i,\vec{c})$ or $I_R(d_i,\vec{c})$ or $I_{SU}(\langle ital\rangle \vec{d}_i,\vec{c})$ or R_f or R_{OneR} or χ^2 values, otherwise $M_{adj}(P_i,P_j) = 0$ 2. $P_{GA}(0)$ from randomly picked G_d of size s_{Clique} 3. Replace six individuals with six s_{Clique} best of $I_G(\vec{d}_i,\vec{c})$, $I_R(\vec{d}_i,\vec{c})$, $I_{SU}(\vec{d}_i,\vec{c})$, R_f, R_{OneR}, χ^2
GA-MF **G**enetic **A**lgorithm based on **M**ultiple **F**ilters without using clique detection	1. $P_{GA}(0)$ picking n_{Filter} randomly from p_{Filter} of the greatest $I_G(\vec{d}_i,\vec{c})$, $I_R(\vec{d}_i,\vec{c})$, $I_{SU}(\vec{d}_i,\vec{c})$, R_f, R_{OneR}, χ^2 values 2. Replace six individuals with six s_{Clique} best of $I_G(\vec{d}_i,\vec{c})$, $I_R(\vec{d}_i,\vec{c})$, $I_{SU}(\vec{d}_i,\vec{c})$, R_f, R_{OneR}, χ^2

Chart 3. The Hybrid GA-SEC Algorithm for Feature Selection[a]

```
// GASEC hybrid filter-wrapper feature selection
t := 0;
initPopulation(P_GA(0));      // filter approach
evaluate(P_GA(0));
repeat                        // wrapper approach
   P_GA' := selectForVariation(P_GA(t));
   recombine(P_GA');
   mutate (P_GA');   // mutation operator
   evaluate(P_GA');
   P_GA(t + 1) := selectForSurvival(P_GA(t), P_GA');
   t := t + 1;
until terminate = true;
```

[a] The initialization uses the *filter approach*, and the loop of the genetic algorithm represents the feature selection *wrapper approach* (see Table 2).

GA-MF initializes $P_{GA}(0)$ directly without using Shannon entropy clique detection and taking only the best descriptors found by the filters into account. All multiple filter (*MF*) variants contain always one individual with size s_{Clique} with the best descriptors found by the plain filter approaches. An additional scheme containing the filter and GA wrapper with additional descriptions is available in the Supporting Information.

The number of features to select is identified automatically, in contrast to the previously presented RFE.

In part 2 of this two-part paper we show how these GA-SEC algorithms can be used to find the most relevant features for predicting human intestinal absorption (HIA) coefficients out of a large data set of 2934 descriptors.

CONCLUSIONS

We presented basic principles for hypothesis testing (induction) and the generalization error in context of the feature selection problem. The entropy terms were described in detail to appreciate previous publications in this area and avoid misleading terms. The difference between *feature extraction* and *feature selection* and the further distinction between *feature selection wrappers* and *feature selection filters* were presented to show that there are already many different approaches available, which can be helpful for selecting and understanding molecular features.

Finally, we presented three new hybrid feature selection algorithms GA-SEC and its two variants GA-SEC-MF and GA-SEC-MFP. We are taking also the class information of our classification problem into account, to reduce the uncertainty of the features used to be picked for the final feature set. This set has the highest generalization ability and a small number of features used. In contrast to standard filter approaches and the RFE wrapper approach, the number of features to be selected must not be defined for applying our GA-SEC algorithm and its variants.

ACKNOWLEDGMENT

We thank ALTANA Pharma AG, Konstanz, Germany for financial support and providing us with the HIA data set. Additionally we thank Simon G. Wiest for fruitful discussions.

Supporting Information Available: Glossary of mathematical symbols, the names of the duplicate compounds, and the corrected molecular structures in SDF format and a detailed scheme for our hybrid GA-SEC algorithm. Furthermore, it contains two huge QSAR feature selection benchmark data sets with ∼3000 descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Davies, S.; Russell, S. Np-completeness of searches for smallest possible feature sets, *Proceedings of the 1994 AAAI Fall Symposium on Relevance.* AAAI Press: New Orleans, 1994; pp 37−39.
(2) Li, M.; Vitanyi, P. Inductive Reasoning and Kolmogorov Complexity. *J. Comput. System Sci.* **1992**, *44*, 343−384.
(3) Domingos, P. The Role of Occam's Razor in Knowledge Discovery. *Data Min. Knowledge Discov.* **1999**, *3*, 409−425.
(4) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, U.S.A., 1995.
(5) Goutte, C. Statistical Learning and Regularisation for Regression, Dissertation, Paris, France, 1997.
(6) Wegner, J.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077−1084.
(7) Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169−177.
(8) Kohavi, R. Wrappers for Performance Enhancement and Oblivious Decision Graphs, Dissertation, Stanford University, 1995.

(9) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: D-69469 Weinheim, Germany, 2000; ISBN 3-52-29913-0.
(10) Trigg, L. Designing Similarity Functions, Dissertation, University of Waikato: New Zealand, 1997.
(11) Kapur, J. N. *Measures of information and their applications*; John Wiley & Sons: New Delhi, India, 1994; ISBN 0-470-22064-3.
(12) Reunanen, J. Overfitting in Making Comparisons Between Variable Selection Methods. *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1371−1382.
(13) Bengio, Y.; Chapados, N.; Extensions to Metric-Based Model Selection. *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1209−1227.
(14) Stahura, F. L.; Godden, J. W.; Bajorath, J. Differential Shannon Entropy Analysis Identifies Molecular Property Descriptors that Predict Aqueous Solubility of Synthetic Compounds with High Accuracy in Binary QSAR Calculations. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 550−558.
(15) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. Distinguishing between Natural Products and Synthetic Molecules by Descriptor Shannon Entropy Analysis and Binary QSAR Calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1245−1252.
(16) Godden, J. W.; Bajorath, J. Differential Shannon Entropy as a Sensitive Measure of Differences in Database Variability of Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060−1066.
(17) Agrafiotis, D. K. On the Use of Information Theory for Assessing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 576−580.
(18) Kay, J. J. Self-Organization In Living Systems, Dissertation, University of Waterloo: Waterloo, Ontario, Canada, 1984.
(19) Eckschlager, K.; Danzer, K. *Information theory in analytical chemistry*; Wiley-Interscience: New York, U.S.A., 1994; ISBN 0-471-59507-1.
(20) Eckschlager, K. Stepanek, V. *Information theory as applied to chemical analysis*; Wiley-Interscience: New York, U.S.A., 1979; ISBN 0-471-04945-X.
(21) Burger, K. Neue Möglichkeiten der Kristallstrukturbestimmung aus Pulverdaten durch die Nutzung resonanter Streuung von Röntgenstrahlung und der 'Maximum Entropy' Methode, Dissertation, Universität Tübingen, Germany, 1997; Shaker Verlag: ISBN 3-8265-3018-7.
(22) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; John Wiley & Sons: New York, 1991; ISBN 0-471-06259-6.
(23) Lin, J. Divergence Measures on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145−151.
(24) Dhillon, I. S.; Mallela, S.; Kumar, R. A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification (Kernel Machines Section). *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1265−1287.
(25) Globerson, A.; Tishby, N. Sufficient Dimensionality Reduction (Kernel Machines Section), *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1307−1331.
(26) Torkkola, K. Feature Extraction by Non-Parametric Mutual Information Maximization (Kernel Machines Section). *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1415−1438.
(27) Witten, I. H.; Eibe, F. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*; Morgan Kaufmann: 1999; ISBN 1-55860-552-5.
(28) Principe, J. C.; Xu, D.; Fisher, J. W. Information-Theoretic Learning. In *Unsupervised Adaptive Filtering*; Wiley-Interscience, 1999; Chapter 7, pp 265−319, ISBN 0-471-37941-7.
(29) Zyczkowski, K. Rényi Extrapolation of Shannon Entropy. *Open Sys., Inf. Dyn.* **2003**, *10*, 297−310.
(30) Bronstein, I. N.; Semendjajew, K. A. Taschenbuch der Mathematik, Teubner, Stuttgart, Germany, 1991.
(31) Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M. Genetic Algorithm Applied to the Selection of Factors in Principal Component-Artificial Neural Networks: Application to QSAR Study of Calcium Channel Antagonist Activity of 1,4-Dihydropyridines (Nifedipine Analogous). *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1328−1334.
(32) Baumann, K.; Albert, H.; Korff, M. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. part I. search algorithm, theory and simulations. *J. Chemom.* **2002**, *16*, 339−350.
(33) Baumann, K.; Korff, M.; Albert, H. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. part ii. practical applications. *J. Chemom.* **2002**, *16*, 351−360.
(34) Malinowski, E. R. *Factor Analysis in Chemistry*; Wiley-Interscience: New York, 2002; ISBN 0-471-13479-1.
(35) Nilson, J. Multiway Calibration in 3D QSAR, Dissertation, University Groningen, Groningen, Sweden, 1999.
(36) Carreira-Perpiñán, M. A. Continuous latent variable models for dimensionality reduction and sequential data reconstruction, Dissertation, University of Sheffield, UK, 2001.
(37) Varmuza, K. Multivariate Data Analysis in Chemistry. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 3, pp 1098−1134, ISBN 3-527-30680-3.
(38) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a Preferred Set of Molecular Descriptors for Compound Classification Based on Principal Component Analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 669−704.
(39) Xue, L.; Bajorath, J. Molecular Descriptors for Effective Classification of Biologically Active Compounds Based on Principal Component Analysis Identified by a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 801−809.
(40) Gohlke, H.; Dullweber, F.; Kamm, W.; März, J.; Kissel, T.; Klebe, G. Prediction of Human Intestinal Absorption using a combined 'Simulated Annealing/Back-propagation Neural Network' Approach. *Rational Approaches Drug Des.* **2001**, 261−270.
(41) Luçić, B.; Trinajstic, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 121−132.
(42) Hwang, W. S.; Weng, J. Hierachical Discriminant Regression. *IEEE Trans. Pattern Analysis Machine Intelligence* **2000**, *22*, 1−6.
(43) Hyvärinen, A.; Oja, E. Independent Component Analysis: A Tutorial. *Neural Networks* **2000**, *13*, 411−430.
(44) Agrafiotis, D. K.; Rassokhin, D. N.; Lobanov, V. S. Multidimensional scaling and visualization of large molecular similarity tables. *J. Comput. Chem.* **2001**, *22*, 488−500.
(45) Rassokhin, D. N.; Lobanov, V. S.; Agrafiotis, D. K. Nonlinear mapping of massive data sets by fuzzy clustering and neural networks. *J. Comput. Chem.* **2001**, *22*, 373−386.
(46) Agrafiotis, D. K.; Lobanov, V. S. Nonlinear Mapping Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1356−1362.
(47) Cho, S. J.; Cummins, D.; Bentley, J.; Andrews, C. W.; Tropsha, A. An Alternative to 3D QSAR: Application of Genetic Algorithms and Partial Least Squares to Variable Selection of Topological Indices. Submitted for publication in *J. Comput. Aided Mol. Des.*
(48) Dunn, W. J.; Wold, S.; Edlund, U.; Hellberg, S.; Gasteiger, J. Multivariate Structure−Activity Relationships Between Data from a Battery of Biological Tests and an Ensemble of Structure Descriptors: The PLS Method. *Quant. Struct.-Act. Relat.* **1984**, *3*, 131−137.
(49) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(50) Bergström, C. A. S.; Strafford, M.; Lazorova, L.; Avdeef, A.; Luthman, K.; Artursson, P Absorption Classification of Oral Drugs Based on Molecular Surface Properties. *J. Med. Chem.* **2003**, *46*, 558−570.
(51) Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative Structure−Activity Relationship Analysis of Functionalized Amino Acid Anticonvulsant Agents Using k-Nearest-Neighbor and Simulated Annealing PLS Methods. *J. Med. Chem.* **2002**, *45*, 2811−2823.
(52) Stanton, D. T. On the Physical Interpretation of QSAR Models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1423−1433.
(53) Eriksson, L.; Antti, H.; Holmes, E.; Johansson, E.; Lundstedt, T.; Shockcor, J.; Wold, S. Partial Least Squares (PLS) in Cheminformatics. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 3, pp 1134−1166, ISBN 3-527-30680-3.
(54) Schöllkopf, B. Support Vector Learning, Dissertation, University of Berlin, Oldenbourg Verlag: Germany, 1997.
(55) Schölkopf, B.; Smola, A. J. *Learning with Kernels − Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, 2002; ISBN 0-262-19475-9.
(56) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection (Kernel Machines Section). *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1157−1182.
(57) Stoppiglia, H.; Dreyfus, G.; Dubois, R.; Oussar, Y. Ranking a Random Feature for Variable and Feature Selection. *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1399−1414.
(58) Hall, M. Correlation-based feature selection for machine learning, Dissertation, University of Waikato, Australia, 1999.
(59) Belanche, L.; Molina, L. C.; Nebot, A. Feature Selection Algorithms: A Survey and Experimental Evaluation. In *2002 IEEE International Conference on Data Mining (ICDM'02'*, Institute of Electrical and Electronics Engineers, Maebashi City, Japan, 2002, 306.
(60) Yasri, A.; Hartsough, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218−1227.
(61) Hoffman, B. T.; Kopajtic, T.; Katz, J. L.; Newman, A. H. 2D QSAR Modeling and Preliminary Database Searching for Dopamine Transporter Inhibitors Using Genetic Algorithm Variable Selection of Molconn Z Descriptors. *J. Med. Chem.* **2000**, *43*, 4151−4159.
(62) Ozdemir, M. Evolutionary computing for feature selection and predictive data mining, Dissertation, Rensselaer Polytechnic Institute Troy, New York, U.S.A., 2002.

(63) Weston, J.; Elisseeff, A.; Schölkopf, B.; Tipping, M. Use of the Zero-Norm with Linear Models an Weston, J.; Elisseeff, A.; Schölkopf, B.; Tipping, M. Use of the Zero-Norm with Linear Models and Kernel Methods'. *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1439−1461.

(64) Bi, J.; Bennett, K.; Embrechts, M.; Breneman, C.; Song, M. Dimensionality Reduction via Sparse Support Vector Machines (Kernel Machines Section). *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1229−1243.

(65) Rakotomamonjy, A. Variable Selection Using SVM-based Criteria. *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1357−1370.

(66) Godden, J. W.; Furr, J. R.; Bajorath, J. Recursive Median Partitioning for Virtual Screening of Large Databases. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 182−188.

(67) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.

(68) Agrafiotis, D. K.; Xu, H. A Geodesic Framework for Analyzing Molecular Similarities. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 475−484.

(69) Rivals, I.; Personnaz, L. MLPs (Monolayer Polynomials and Multi-Layer Perceptrons) for Nonlinear Modeling Isabelle. *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1383−1398.

(70) Perkins, S.; Lacker, K.; Theiler, J. Graftin: Fast, Incremental Feature Selection by Gradient Descent in Function Space. *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1333−1356.

(71) Caruana, R.; Sa, V. R. Benefitting from the Variables that Variable Selection Discards. *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1245−1264.

(72) Bekkerman, R.; El-Yaniv, R.; Tishby, N.; Winter Y. Distributional Word Clusters vs Words for Text Categorization (Kernel Machines Section). *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1183−1208.

(73) Forman, G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *J. Machine Learning Res.* (special issue on Variable and Feature Selection) **2003**, *3*, 1289−1305.

(74) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in diversity profiling and combinatorial series design. *Molecular Diversity* **1999**, *4*, 1−22.

(75) Willet, P. Subset-Selection Methods For Chemical Databases. In *Molecular Diversity in Drug Design*; Dean, M., Lewis, R. A., Eds.; Kluwer Academic Publishers: Dordrecht, Netherlands, 1999; ISBN 0-7923-5980-1.

(76) Kira, K.; Rendell, L. *A practical approach to feature selection.* In *Proceeedings of the Ninth International Workshop on Machine Learning (ML92)*; Sleeman, D., Edwards, P., Eds.; Morgan Kaufman: San Mateo, CA, 1992; pp 249−256.

(77) Kononenko, I. Estimating Attributes: Analysis and Extensions of RELIEF. In *Proceedings of the European Conference on Machine Learning*; Bergadano, F., Raedt, L. D., Eds.; 1994.

(78) Holte, R. C. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning* **1993**, *3*, 63−91.

(79) Holmes, G.; Nevill-Manning, C. G. Feature Selection via the Discovery of Simple Classification Rules. In *Proceedings of the International Symposium on Intelligent Data Analysis (IDA-95)*; Baden-Baden, Germany, 1995.

(80) Liu, H.; Li, J.; Wong, L. A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. *Genome Informatics* **2002**, *13*, 51−60.

(81) Altman, D. G. *Practical statistics for medical research*; Chapmann & Hall/CRC: New York, U.S.A., 1991; ISBN 0-412-27630-5.

(82) Wu, S.; Flach, P. A. Feature selection with labeled and unlabeled data. In Bohanec, M., Kasek, B., Lavrac, N., Mladenic, D., Eds.; *ECML/PKDD'02 workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*; University of Helsinki: 2002; pp 156−167.

(83) Kohavi, R.; John, G. Wrappers for feature selection. *Artificial Intelligence* **1997**, *97*, 273−324.

(84) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **2002**, *46*, 389−422.

(85) Czerminnski, R.; Yasri, A.; Hartsough, D. Use of Support Vector Machine in Pattern Classification: Application to QSAR Studies. *Quant. Struct.-Act. Relat.* **2001**, *20*, 227−240.

(86) Cristianini, N.; Taylor, J. S. *An Introduction to Support Vector Maschines − and other kernel-based learning methods.* Cambridge University Press: Cambridge, UK, 2000; ISBN 0-521-78019-5.

(87) Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. Decision forest: Combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 525−531.

(88) Poland, J.; Zell, A. Different Criteria for Active Learning in Neural Networks: A Comparative Study. In *Proceedings of the 10th European Symposium on Artificial Neural Networks (ESANN 2002);* Verleysen, M., Evere/Belgium, 2002; pp 119−124.

(89) Manallack, D. T.; Tehan, B. G.; Gancia, E.; Hudson, B. D.; Ford, M. G.; Livingstone, D. J.; Whitley, D. C.; Pitt, W. R. A Consensus Neural Network-Based Technique for Discriminating Soluble and Poorly Soluble Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 674−679.

(90) Maclin, R.; Shavlik, J. W. Combining the predictions of multiple classifiers: Using competitive learning to initialize neural networks. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*; Montreal, Canada, 1995; pp 524−530.

(91) Alpayd, E. Techniques for combining multiple learners. In *Proceedings of Engineering of Intelligent Systems EIS'98;* Teneriffe, Spain, 1998; pp 6−12.

(92) Breimann, L. Bagging Predictors. *Machine Learning* **1996**, *24*, 123−140.

(93) Freund, Y.; Schapire, R. A short introduction to boosting. *J. Jpn. Soc. Artif. Intel.* **1999**, *14*, 771−780.

(94) Agrafiotis, D. K.; Ceden, W.; Lobanov, V. S. On the use of neural network ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.

(95) Schapire, R. E.; Freund, Y.; Bartlett, P. L.; Lee, W. S. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals Statistics* **1998**, *26*, 1651−1686.

(96) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.

(97) Cho, S. J.; Shen, C. F.; Hermsmeier, M. A. Binary Formal Inference-Based Recursive Modeling Using Multiple Atom and Physicochemical Property Class Pair and Torsion Descriptors as Decision Criteria. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 668−680.

(98) Yuan, X.; Yuan, X.; Buckles, B. P.; Zhang, J. *A Comparison Study of Decision Tree and SVM to Classify Gene Sequence, 19th International Conference on Data Engineering (ICDE'03)*; March, Bangalore, India, 2003.

(99) JOELib, http://joelib.sourceforge.net/.

(100) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures − A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(101) Figueras, J. Ring Perception Using Breadth-First Search. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 986−991.

(102) Ivanciuc, O. Canonical Numbering and Constitutional Symmetry. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 1, pp 139−160, ISBN 3-527-30680-3.

(103) Weininger, D. SMILES: a Chemical Language for Information Systems. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(104) Weininger, D. SMILES 2: Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(105) Weininger, D. SMILES-A Language for Molecules and Reactions. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 1, pp 80−102, ISBN 3-527-30680-3.

(106) Ihlenfeldt, W. D.; Gasteiger, J. Hash Codes for the Identification and Classification of Molecular Structure Elements. *J. Comput. Chem.* **1994**, *15*, 793−813.

(107) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Theoretical Aspects of Ring Perception and Development of the Extended Set of Smallest Rings Concept. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 187−206.

(108) Downs, G. M. Ring Perception. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 1, pp 161−177, ISBN 3-527-30680-3.

(109) Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculating Partition Coeffcients. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615−621.

(110) Wang, R.; Gao, Y.; Lai, L. Calculating partition coeffcient by atom-additive method. *Perspect. Drug Discov. Des.* **2000**, *19*, 47−66.

(111) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.

(112) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.

(113) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.

(114) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2002**, *43*, 429−434.
(115) Liu, R.; So, S. S. Development of Quantitative Structure−Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633−1639.
(116) Livingstone, D. J.; Ford, M. G.; Huuskonen, J. J.; Salt, D. W. Simultaneous prediction of aqueous solubility and octanl/water partition coefficient based on descriptors derived from molecular structure. *J. Comput.-Aided. Mol. Des.* **2001**, *15*, 741−752.
(117) Mazzatorta, P.; Benfenati, E.; Neagu, D.; Gini, G. The importance of scaling in data mining for toxicity prediction. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1250−1255.
(118) Tounge, B. A.; Pfahler, L. B.; Reynolds, C. H. Chemical Information Based Scaling of Molecular Descriptors: A Universal Chemical Scale for Library Design and Analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 879−884.
(119) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.
(120) Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Chemical Information in 3D-Space. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030−1037.
(121) OpenBabel, http://openbabel.sourceforge.net/.
(122) Clark, D. E. *Evolutionary Algorithms in Molecular Design*; Wiley-VCH: 2000; ISBN 3-527-30155-0.
(123) Homeyer, A. Evolutionary Algorithms and their Applications in Chemistry. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 3, pp 1239−1280, ISBN 3-527-30680-3.
(124) Bron, C.; Kerbosch, J. Finding all cliques of an undirected graph. *Comm. ACM.* **1973**, *16*, 575−577.
(125) Bomze, I.; Budinich, M.; Pardalos, P.; Pelillo, M. The maximum clique problem. In *Handbook of Combinatorial Optimization*; Du, D.-Z., Pardalos, P. M., Eds.; Kluwer Academic Publishers: Boston, MA, 1999; Vol. 4.
(126) Gardiner, E. J.; Holliday, J. D.; Willett, P.; Wilton, D. J.; Artymiuk, P. J. Selection of reagents for combinatorial synthesis using clique detection. *Quant. Struct.-Act. Relat.* **1998**, *17*, 232−236.

CI0342324