IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP5129

This paper appears in *Managing Modern Organizations Through Information Technology*, Proceedings of the 2005 Information Resources Management Association International Conference, edited by Mehdi Khosrow-Pour.

Prediction for Compound Activity in Large Drug Datasets Using Efficient Machine Learning Approaches

Larry J. Layne
Division of Government Research, Univ. of New Mexico, Albuquerque, NM 87131, USA, ljlayne@unm.edu

Shibin Qiu
Dept of Electrical & Computer Engineering, Univ. of New Mexico, Albuquerque, NM 87131, USA, squi@unm.edu

ABSTRACT

Modern drug design requires activity prediction within a large number of chemical compounds using their descriptors that are often generated with high-noise in high-dimensional space. Both computational performance and classification quality face great challenges if machine learning algorithms are to be applied successfully. For computational efficiency, we implement the proximal support vector machine (PSVM) since it only depends on linear operations and can be trained faster than support vector machines (SVM) using quadratic optimization. For even larger datasets, we use parallel computing to make the training and classification time acceptable. To improve the classification quality, we implement and compare the SVM, k-nearest neighbor, decision tree and the naive Bayes classifiers. We measure the classification qualities by using the cross-validation accuracies, generalization accuracies, and the false positive and false negative ratios in ROC (receiver operating characteristics) curves. We also conduct feature selection in order to find the most important features and gain insights into the nature of the descriptors of the compounds. Features are easy to select using linear SVMs but the selection may be biased. We use a nonlinear kernel SVM in the feature selection process to achieve a higher ranking quality. To fully understand the properties of the noisy features in the dataset, we experiment with different number of features using the SVM classifier to obtain an optimal number of features.

INTRODUCTION

The problem of drug design is to find drug candidates from a large collection of compounds that will bind to a target molecule of interest. The compounds can be described by using a number of numerical descriptors, or by using structural features directly. The descriptors are drawn from a compound's physicochemical properties. The descriptors found in the quantitative structure activity relationship (QSAR) dataset [9], for instance, encode the size, polarity, hydrogen bond donor, etc., for each of the possible substitutions. The thrombin drug dataset [1], also used for the 2001 KDD cup, uses indicator variables as the descriptors, indicating the presence or absence of certain characteristics derived from shape-based comparison and alignment of compounds. Descriptors may also be used to represent the traditional pharmacophore properties, as is the case for the CDK2 dataset [14]. A compound may have a large number of descriptors and the descriptor space contains numerous dimensions. Activities meaningful for drug design include chemical reactivity, biological activity and toxicity. It is widely believed that the structure of a compound has a direct relationship to its activity. Therefore, QSAR data representation is frequently used for computational models. In this work, we use the QSAR data abstraction to classify drug-binding activities.

Neural networks have been used for the classification of active drug candidates [5]. Genetic algorithms and decision trees have also been applied to drug design [6, 9]. Support vector machines (SVM) have been used for the analysis of pharmaceutical data [4] and for active learning in drug discovery [14]. In this paper, we use the k-nearest neighbor method and the naive Bayes classifier to classify active compounds, in addition to using SVM and decision tree classifiers. We analyze and compare the performance for each of these classifiers. Due to the high dimensionality of the compound descriptor space, feature selection is often conducted to select the most important descriptors and obtain informative insights into the compounds. The support vector regression method was used to select variables [3]. We use the recursive feature elimination (RFE) approach to select feature descriptors [8]. Many researchers use linear kernels for RFE, but we choose nonlinear kernel in the RFE feature selection process in order to obtain better ranking. The large number of features in the drug datasets are mixed with noise and prevent normal machine learning algorithms from accurately and effectively classifying the compounds. We test different numbers of features and estimate the optimal number of features using the ranked features. We test the algorithms on two datasets, the pyrimidines and the triazines QSAR datasets. The former has been tested by a few researchers but the latter has been rarely used due to its large size. The triazines dataset contains more than 10,000 data points and requires a considerable amount of training time. We use the proximal SVM (PSVM) to improve computational performance. PSVM only needs linear algebraic operations and does not rely on quadratic programming, which is generally time consuming, as used by traditional SVMs. To further reduce PSVM training time, we shuffle the dataset and select a subset of the data points. To measure the classification qualities, we use the false positive and false negative ratios in ROC (receiver operating characteristics) curves, in addition to the cross-validation accuracies and generalization accuracies. The thrombin dataset exhibits extremely high dimensionality, as it contains 139,351 descriptors for each compound. For this dataset, we use parallel computing to reduce the training time from hours to minutes.

We organization this paper as follows: In Section 2 we briefly describe the classification algorithms used; in Section 3 we explain the datasets used in this paper and present classification results; study feature selection is presented in Section 4; and, conclusions are made in Section 5.

CLASSIFICATION ALGORITHMS

In this section we briefly describe the classification algorithms used in our study for drug discovery, including support vector machine, naive Bayes classifier, decision tree, and k-nearest neighbor method.

Support Vector Machine

Support vector machines (SVMs) [13] have been successfully applied to a wide range of problems, including object recognition, speaker identification, face detection and text categorization. A support vector

machine finds an optimal separating hyperplane between members of two classes either in the input space or in an abstract feature space. If the boundary between the two classes is nonlinear, a kernel function is used to map the training data nonlinearly into a high dimensional feature space and construct a separating hyperplane with maximum margin in the feature space.

The basic idea of an SVM classifier is to find an optimal maximal margin separating hyperplane between the two classes of data points. SVMs use an implicit nonlinear mapping from the input space to a higher dimensional feature space using kernel functions, in order to classify the data points which are not linearly separable in the input space. The nonlinear support vector machine can be formulated as follows, where we follow the notation used by Mangasarian et al. [7].

$$\min \quad v\frac{1}{2}\|y\|^2 + \frac{1}{2}\left(u^T u + \gamma^2\right) \tag{1}$$

$$\text{subject to} \quad D(K(X,X^T)Du - e\gamma) + y \geq e,$$

where $X \in \Re^{m \times n}$ is the training data matrix, m is the number of data points in the training set and n is the dimension of input space. We use X_i to denote the i^{th} row of matrix X, representing the i^{th} data point. $D \in \Re^{m \times m}$ is the diagonal label matrix, D_{ii} = 1(-1), if the label at the i^{th} data point is 1(-1). $u \in \Re^m$ and $\gamma \in \Re$ are the variables to be solved. n is a weighting parameter for the separation error vector $\gamma \in \Re^m$. $e \in \Re^m$ is a vector of all ones. $K \in \Re^{m \times m}$ is the kernel matrix between X and X^T,

$$K = K(X,X^T). \tag{2}$$

The commonly used Gaussian kernel and polynomial kernel are:

$$(K(X,X^T))_{ij} = exp(-\mu\|X_i - X_j\|^2) \tag{3}$$

$$(K(X,X^T))_{ij} = (1 + < X_i, X_j >)^d. \tag{4}$$

Solving the above optimization problem, we get u, γ and the separating function. For a given data point x, the classifier function (separating surface) is represented as,

$$f(x) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i D_{ii} k(x, x_i) + \gamma\right), \tag{5}$$

where the set of nonzero coefficients of a_i determines the support vectors.

The Proximal Support Vector Machine

The optimization problem of (1) is a quadratic programming problem. Its solution cannot be derived from linear algebraic operations alone. The proximal SVM (PSVM) is derived by changing (1) to (6) below,

$$\min \quad v\frac{1}{2}\|y\|^2 + \frac{1}{2}\left(u^T u + \gamma^2\right) \tag{6}$$

$$\text{subject to} \quad D(K(X,X^T)Du - \gamma^a) + y = e.$$

The solution to (6) is purely linear and defines the PSVM classifier function as below.

$$f(x) = sgn((K(x^T,X^T)K(X,X^T)+e^T)Dv), \tag{7}$$

where

$$v = (I/v + D(KK^T + ee^T)D)^{-1}e$$

$$= (I/v + GG^T)^{-1}e, \tag{8}$$

and

$$G = D(K - e). \tag{9}$$

The other variables can also be solved,

$$u = DK^T Dv \tag{10}$$

$$\gamma = -e^T Dv \tag{11}$$

$$y = v/v. \tag{12}$$

The optimization problem of (6) is equivalent to that defined by (1). Both theoretic proof and geometric interpretation are given in [7]. PSVM is much faster to solve than ordinary SVMs requiring quadratic programming.

Naive Bayes Classifier

The naive Bayes classifier assumes independence of the features of a dataset and uses Bayes theorem to estimate the posterior probabilities. A class with the highest posterior probability is chosen as the label of a given data point. Suppose data point x has features $x_1,x_2,...,x_n$, and there are C classes in the label set. The probability of data point x given a class c is,

$$P(x \mid c) = \prod_i^n P(x_i \mid c), \tag{13}$$

where $P(x_i|c)$ is the probability of feature x_i given class label c. The decision rule of the naive Bayes classifier is,

$$\hat{c} = \arg\max_{c \in C} P(c) \prod_i^n P(x_i \mid c), \tag{14}$$

where $P(c)$ is the prior probability of class c. One advantage of the naive Bayes classifier is that it can easily handle classification problems having multiple classes.

Decision Tree Classifier

A decision tree classifier consists of a set of rules describing the conditions based on which a given data point should be classified. The rules are trained using the features of a dataset to maximize the information gain. An internal node of a decision tree is labeled with the name of the feature and there is one branch for each range of the feature value. The leaf nodes specify class categories. The classification process for a data point traverses the tree from the root down to one of the leaf nodes.

Nearest Neighbor Classifier

The nearest neighbor classifier assigns the label of a given data point based on the majority of the labels of a fixed number (k) of data points in its neighborhood. It is also called the k-nearest neighbor classifier (abbreviated as kNN). The classification rule for the kNN algorithm summarized as follows:

1. Identify the k training data points that lie nearest to the test data point x.
2. Assign x to the class that is most frequently represented in the neighborhood.

COMPOUND ACTIVITY PREDICTION

In this section we describe the data used for our experiments and analyze classification accuracies for the different classifiers. We also analyze the false positive ratios and plot the ROC curves for the SVM classifier.

Data

The datasets we use are the pyrimidines and triazines QSAR datasets, which are obtained from the UCI machine learning repository. The classification functions are used to predict the inhibition of dihydrofolate reductase by pyrimidines and triazines [2]. Each dataset contains active and inactive compounds and is divided into a training set and a test set. We need to use effective machine learning algorithm to discriminate active compounds from inactive compounds.

In the pyrimidines dataset, each drug has 3 positions of possible substitution. There are 9 attributes for each substitution position, including polarity, size, flex, h_doner, h_acceptor, pi_doner, pi_acceptor, polarisable, and the sigma property. Therefore every compound has 54 attributes. In the triazines dataset, each drug has 6 positions of possible substitution. There are 10 attributes for each substitution position, including polarity, size, flex, h_doner, h_acceptor, pi_doner, pi_acceptor, polarisable, sigma, and branch. As a result, each compound has a total of 120 attributes. There are two drugs in each compound's record.

Classification Qualities

We first use 10-fold cross-validation accuracy to measure classification quality for each of the classifiers in the two datasets. The results are shown in Table 1, which indicates that SVM has achieved the highest accuracy.

To investigate the generalization capabilities of the classifiers, each classifier is also used to predict the test set after being trained using the training set. Generalization tests are performed on both datasets and the test results are displayed in Table 2. SVM also has achieved the highest generalization accuracies, as shown in the table.

Since SVM has obtained superior classification qualities, we study its behavior in more detail. If an active compound is classified as inactive, we miss a valuable chance of discovering a drug candidate. This case is referred to as false negative. On the other hand, if an inactive compound is classified as active, we get a wrong candidate. This case is referred to as false positive.

Both false positive and false negative cases decrease classification accuracy. However, they do not have equal consequences in real applications. If we were concerned more of discovering drug candidates than finding a few wrong candidates, we might focus more on eliminating false negative rates. Therefore we need to investigate the false negative and false positive ratios in detail, in addition to measuring the classification accuracies of the classifiers. We analyze the false positive and false negative ratios by using the ROC curves. Figure 1 shows the ROC curves of the SVM classifier for the two datasets in the generalization test. It is the plot of true positive ratio against the false positive ratio. We can evaluate the false positive ratio corresponding to any given true positive ratio from this figure. SVM has performed better on the pyrimidine dataset because there is a larger area under its ROC curve. For the pyrimidine dataset, the false positive ratio is 16.2%, and the false negative ratio is 12.3%. Therefore, slightly more inactive compounds have been classified as active. For the triazine dataset, however, more active compounds have been classified as inactive, as indicated by a false positive ratio of 24.5%, and the false negative ratio of 34%.

Table 1. Ten Fold Cross Validation Accuracies of the Classifiers on the Pyrimidine and Triazine Datasets (NB stands for naive Bayes; DT, decision tree; 10NN, 10 nearest neighbor)

Classifier	NB	DT	10-NN	SVM
Pyrimidines	82.1%	92.56%	91.54%	95.91%
Triazines	65.83%	75.5%	69.17%	78.33%

Table 2. Generalization Accuracies of the Classifiers for the Pyrimidine and Triazine Datasets (NB stands for naive Bayes; DT, decision tree; 10NN, 10 nearest neighbor)

Classifier	NB	DT	10-NN	SVM
Pyrimidines	81.18%	85.77%	83.72%	86.35%
Triazines	65.91%	65.25%	67.19%	70.83%

Figure 1. ROC Curves of the SVM Classifier (the curve marked with "pyrim" is the ROC curve for the pyrimidine dataset, and "triaz" for the triazine dataset)

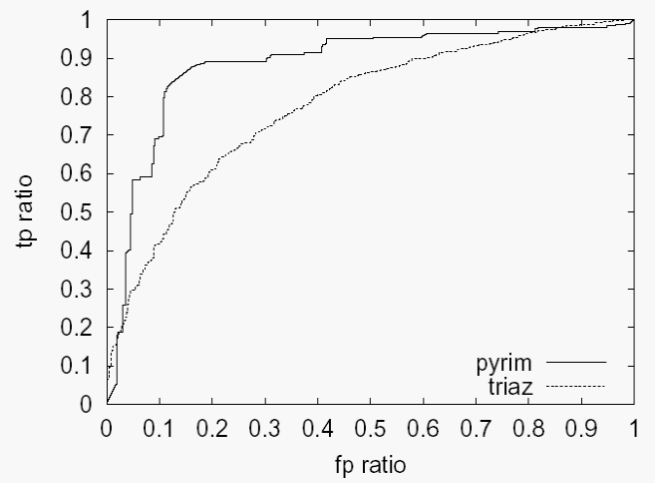
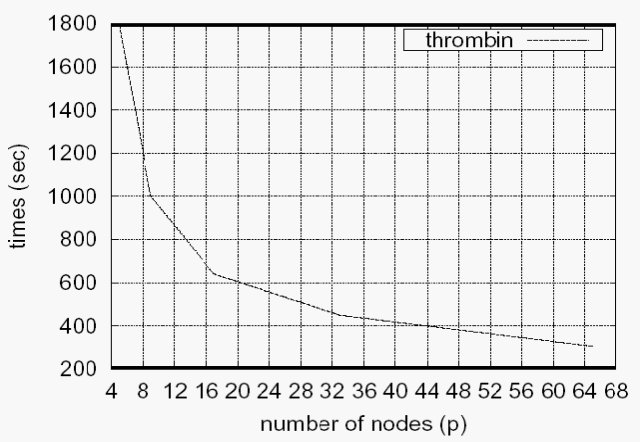


Figure 2. Parallel Classification Times for Thrombin Dataset



Parallel Computation for High Dimensional Datasets

Drug design datasets are often large and high dimensional. The thrombin drug design dataset [1], which was also used for the 2001 KDD cup, has 139,351 features and 2,543 data points. Computing an kernel matrix on this dataset requires a considerable amount of processing time. The CDK2 drug design dataset has 14,223 compounds, each of which has 35,926,557 descriptors [14]. This dataset requires a few Tera bytes of memory space, which is beyond the capacity of most single processor computers. However, training time is an even significant barrier. The time complexity of computing a kernel matrix is $O(m^2n)$, where m is the number of data points and n is the dimension of the input space. A computer with sufficient memory and the capacity of 4 GFlops per second, the common speed of today's high end computer, would need 21 days to compute the kernel matrix for this dataset, and much longer

if there is not enough memory. Though feature selection can reduce the dimensionality, it also consumes preprocessing time. The solution to these large datasets is parallel computing.

Linear PSVM has been implemented in parallel [12]. In this paper, we implemented a parallel nonlinear PSVM using a Gaussian kernel. We used the dimension-wise partition [11] for parallel kernel computation. Figure 2 shows that parallel computing times for the thrombin drug dataset. It took 5 hours to train a PSVM classifier and classify the test set using a single processor computer. It only takes 4 minutes on a computer cluster of 65 nodes. Dramatic speedup has been achieved in this experiment. Therefore parallel computing is a proper solution for larger and high dimensional datasets.

FEATURE SELECTION AND RANKING

Drug datasets often have a large number of noisy features. Feature selection can reduce the dimensionality in the input space and improve computational performance. It can also help gain insights into the descriptors of the compounds and explain which features are more important. In addition, smaller number of features can sometimes improve generalization capability of a classifier. We use the SVM based recursive feature elimination (RFE) algorithm to rank the features, and select the most important features to predict the testing data.

SVM-based RFE is an application of the recursive feature elimination [8] algorithm by using weight magnitude from an SVM as the ranking criterion. The algorithm is described in the following list.

1. Initialize using a subset of surviving features $S = [1,2,\ldots,d]$, feature ranked list $r = \varnothing$;
2. Restrict training examples to good feature indices $X = X_0(:,s)$;
3. Use SVM to get weight vector ω;
4. Find feature with smallest weight: $f = argmin(|\omega|)$;
5. Update feature ranked list: $r = [s(f),r]$;
6. Eliminate the feature with the smallest weight: $s = s(1 : f - 1, f+1 : length(s))$;
7. repeat steps 2-6 until $s = \varnothing$.

The algorithm repeatedly eliminates features with the smallest weights and updates the ranked list until all features have been ranked. While most research uses linear SVM to perform feature elimination, we use a nonlinear SVM with a Gaussian kernel. We rank the features in the training set of the pyrimidine dataset. Table 3 shows the ranks for the first position of the substitution corresponding to the first drug in the dataset.

Based on the ranks in Table 3, the polarisable property is the most important feature in the first position of the substitution for the pyrimidines dataset. Poloarisable indicates the polarisability of the molecular orbitals. Sigma is the second important feature, which is the Ã-*property*. Size ranks the third, which is a measure of the extended volume of the group. Flex is the fourth important feature, which measures flexibility and is assigned to the number of rotatable bonds. Polar is the fifth. It indicates polarity and the amount of residual charges

Table 3. Feature Ranks for the Pyrimidines Dataset

Rank	Feature
1	Polarisable
2	Sigma
3	Size
4	Flex
5	Polar
6	h doner
7	pi acceptor
8	pi doner
9	h acceptor

Table 4. SVM generalization accuracy versus the number of features used for the pyrimidines dataset. N is the number of features and r is the classification accuracy using the corresponding number of features.

N	9	18	27	36
ρ	76.41%	81.87%	84.11%	82.45%

on the a and b atoms of the substituent. h_doner ranks the sixth in the feature list. And h acceptor is ninth important feature and indicates the presence and strength of hydrogen-bonding acceptors and donors. pi_acceptor and pi_doner indicate the presence and strength of p-*acceptors* and p-*donors* [9]. The important features contribute more to the activity of a compound. Though acceptor properties are less important than other features based on the SVM RFE algorithm, they are important structural characteristics in drug screening.

Table 4 shows the relationship of SVM generalization accuracy and the number of features used for the pyrimidine dataset. In the experiment, we include the important features with higher priority. Table 4 demonstrates that if we only use the 9 high rank features, the accuracy is not high enough. When we use more features we get better accuracy. We get the best accuracy with 27 features and the accuracy begins to decrease when we use more than 36 features.

CONCLUSIONS

One important problem in modern drug design is to predict the activity of a compound as active or inactive to a binding target using its descriptors, which can be accomplished using machine learning approaches. Computationally, we must use efficient algorithms in the implementation, since drug datasets are large and high dimensional. We have implemented the proximal support vector machine since it can be efficiently trained. For the larger thrombin dataset, we used parallel computing and made the training time acceptable. To achieve high classification quality, we have implemented and compared the performances of the PSVM, naive Bayes, decision tree and the k-nearest neighbor classifiers. The classifiers have achieved varied qualities based on 10 fold cross validation and generalization accuracies. The average cross validation accuracy was 90.25% for the pyrimidine dataset, and 72.25% for the triazine dataset. We found that SVM often performs better. We generated the ROC curves for the SVM classifier to investigate the false positive and false negative ratios to solve the asymmetry issues of the miss classifications. Descriptors of the compounds have different contributions to their activities and class memberships. We used a nonlinear feature selection algorithm and selected the most important features according to their importance. Based on the feature ranking, we tested with different number of features and obtained the optimal number of features for SVM classification.

In the future, we will post the algorithms on a web server and make them available through a web interface.

REFERENCES

[1] The thrombin dataset. http://www.cs.wisc.edu/˜dpage/kddcup2001/

[2] UCI machine learning datasets. http://www.ics.uci.edu/ ˜mlearn / MLRepository.html.

[3] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *JMLR*, 3:1229–1243, March 2003.

[4] R. Burbidge, M. Trotter, B.Buxton, and S.Holden. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers and Chemistry*, 26(1):4–15, December 2001.

[5] J. Devillers. *Genetic Algorithms in Molecular Modelling*. Academic Press, 1999.

[6] J. Devillers. *Neural Networks and Drug Design*. Academic Press, 1999.

[7] G. Fung and O. L. Mangasarian. Proximal support vector machine classifiers. In *Proceedings of the the Second SIAM International Conference on Data Mining* , pages 77–86, 2001.

[8] I. Guyon, J.Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[9] D. M. Hawkins, S. S. Young, and A. Rusinko. Analysis of a large structure-activity data set using recursive partitioning. *Quantitative Structure Activity Relationships*, 16:296–302, 1997.

[10] R. D. King, S. Muggleton, R. A. Lewis, and M. J. Sternberg. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. In *Proceedings of the Natl. Acad. Sci. USA*, 1992.

[11] S. Qiu and T. Lane. *Parallel kernel computation for high dimensional data and its application to fMRI image classification*. Technical Report, TR-CS-2004-12, Computer Science Dept., University of New Mexico, 2004.

[12] A. Tveit and H. Engum. *Parallelization of the incremental proximal support vector machine classifier using a heapbased tree topology*. Technical Report, IDI, NTNU, Trondheim, Norway, August 2003.

[13] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, second edition, 2000.

[14] M. K. Warmuth, J. Liao, G. Ratsch, M. Mathieson, S. Putta, and C. Lemmen. Support vector machines for active learning in drug discovery process. *Journal of Chemical Information Science*, 43(2):667–673, 2003.