Prediction of Molecular Bioactivity for Drug Design using a Decision Tree Algorithm

Sanghoon Lee, Jihoon Yang*, and Kyung-whan Oh

Department of Computer Science, Sogang University
1 Shinsoo-Dong Mapo-Ku Seoul 121-742, Korea
sadclan@ailab.sogang.ac.kr, {jhyang, kwoh}@ccs.sogang.ac.kr

Abstract. A machine learning-based approach to the prediction of molecular bioactivity in new drugs is proposed. Two important aspects are considered for the task: *feature subset selection* and *cost-sensitive classification.* These are to cope with the huge number of features and unbalanced samples in a dataset of drug candidates. We designed a pattern classifier with such capabilities based on information theory and re-sampling techniques. Experimental results demonstrate the feasibility of the proposed approach. In particular, the classification accuracy of our approach was higher than that of the winner of KDD Cup 2001 competition.

1 Introduction

Drugs consist of small organic molecules that achieve their desired activity by binding to a target site on a receptor. The first step in the discovery of a new drug usually involves identifying and isolating the receptor to which it should bind, followed by testing many small molecules for their ability to bind to the target site. This leaves researchers with the task of determining what separates the active (binding) compounds from the inactive (non-binding) ones [1]. Machine learning can thus be an appropriate choice for the classification task.

In general, the problem of analyzing structure, function, and localization of biological data can be solved by classifying feature patterns of the data [2]. We can understand and identify key characteristics of data by classifying feature vectors. However there might be several issues we need to consider when using classification algorithms for biological data (including the dataset for drug design used in this paper). First, a dataset can contain a number of irrelevant, redundant features. In this case, including inappropriate feature can make the classification result less accurate. Second, the examples in the training set might not be drawn from the same distribution where the test examples are drawn. Furthermore, the class distribution of patterns (i.e. number of patterns in each class) can be quite biased. Third, the number of patterns in a dataset is relatively much smaller than the number of features, which incurs high chance of over-fitting.

* This research was partially supported by Korea Research Foundation Grant (KRF-2002-003-D00133) to Jihoon Yang.

We aim to produce an efficient classifier for biological data using a decision tree learning algorithm. Our classifier is designed considering the three issues mentioned above: *feature subset selection, cost-sensitive classification*, and *over-fitting avoidance*. By feature selection, irrelevant, redundant features are eliminated to produce a subset with relevant features only. Since most of biological data consist of very large number of features, feature selection is important. It is also necessary to consider non-uniform costs for misclassification. For instance, predicting a good drug target as mediocre will be more expensive than predicting a mediocre as good. In addition, it is also important to check how well a training set reflects the distribution of real world data, especially when the training set is relatively small compared to the instance space. Preventing over-fitting is thus also of interest. Against this background, we introduce a decision-tree based classifier using entropy-based feature selection, re-sampling-based cost-sensitive classification, and cross-validation-based stopping criterion, and verify its outstanding performance with real-world biological data (from KDD Cup 2001 competition) which will be described in detail in Section 2.3.

2 Related Work

This section briefly introduces related techniques for feature subset selection and cross-validation, and summarizes proposed approaches in KDD Cup 2001.

2.1 Feature Subset Selection

A number of approaches to feature subset selection have been proposed in the literature [3][4][5]. These approaches involve searching for an optimal subset of features based on some criteria of interest. Feature selection algorithms can be broadly classified into the following three categories according to the characteristics of the search strategy employed: *exhaustive search, heuristic search*, and *randomized search* [6]. Exhaustive search strategy is most appropriate when the number of features is sufficiently small, since it finds the optimal feature subset. However, statistical heuristics [7][8][9] or randomized heuristics [6][10][11][12] are used commonly since there are too many features in most cases. Each strategy has advantages as well as disadvantages in a specific domain. In many cases, however, as large as search space becomes, statistical heuristics may become more reasonable than randomized heuristics because of the relatively low computational cost.

2.2 Cross-Validation for Accuracy Estimation

There are several methods to validate the learning model. One of the most widely used techniques is k-fold cross-validation. In k-fold cross validation, training data are partitioned into disjoint k folds of the same size. Then the classification accuracy for each fold is computed as follows: At each time, one of the k folds is used as a validation set and the others as a training set. The average accuracy of

k-run is called k-fold cross-validation accuracy. It is known that cross-validation can make reliable prediction on unknown test set [13]. Generally, 10-fold cross-validation yields the best performance in accuracy estimation [14].

2.3 KDD Cup 2001

KDD Cup 2001 is focused on data from genomics and drug design [1]. Among the tasks, Task 1 is about the prediction of molecular bioactivity for designing a hemostatic. The dataset used here is thrombin dataset (also used in our experiments), and it has various representative characteristics of biological data.

The training set consists of 1,909 compounds (i.e. samples or patterns) tested for their ability to bind to a target site on thrombins - a key receptor in blood clotting. Among the compounds, 42 are active (i.e. binds well to the target site) and the others are inactive. Each compound was described by a single feature vector comprised of a class value (A for active, I for inactive) and 139,351 binary features that describe three-dimensional properties of the molecule [1]. The test set contains 636 additional compounds that were in fact generated based on the assay results recorded for the training set. Therefore the test set has a different class distribution from the training set.

In KDD Cup 2001 competition, a total of 114 groups submitted predictions for Task 1, the thrombin binding problem. In evaluating the accuracy, an average cost model was used, since the data set contains much less active classes than inactive ones. In other words, the average of true positive and negative accuracies (i.e. weighted average accuracy) is used for assessing the performance of classifier [1]. The winner of task 1 achieved 71.1% of test accuracy, and 68.4% of weighted average accuracy [1]. And the second place winner achieved 72% of test accuracy and 64.3% of weighted average accuracy.

3 Our Approaches

3.1 Feature Subset Selection using GINI Index

Although there are many sophisticated feature subset selection techniques, we just employed a simple statistical heuristic to reduce the computational cost. The proposed technique is composed of the following steps: 1) Information gains are computed to measure the amount of information that each feature contains. 2) A feature subset which consists of features that have information gain above specific threshold δ is created. The 10-fold cross-validation accuracy is then computed using C4.5 algorithm for the dataset considering selected features only. (details on optimality check is explained in Section 3.3.) The process terminates if the feature subset satisfies stopping criterion; otherwise step 2) is repeated with a decreased value of δ.

Generally, the information gain means Shannon's information gain [15], but this can be formulated with other information measures (e.g. chi-square, GINI index, etc. [15]). Since the computational cost of GINI index is less than that of

Shannon's entropy, we used GINI index to calculate the information gain. GINI index is defined as

$$G(S) = 1 - \sum_{i=1}^{c} p_i^2$$

where, c is the number of class, and p_i is the proportion of S belonging to class i. The information gain is defined as

$$IG(S, A) = G(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} G(S_v)$$

where, $Values(A)$ is the set of all possible values for feature A, and S_v is the subset of S for which feature A has value v.

A feature with larger information gain contains more information that needed to predict classes. Therefore, we can organize features according to their information gain in a decreasing order and form subsets from the top. In the case of biological data with large number of features, most of the features are often meaningless. Thus, it is possible to obtain an optimum feature subset without irrelevant features by selecting features with high information gain.

3.2 Bootstrap

In the case of thrombin data, relatively small active examples are included in the training set. Therefore, a feature with high discrimination capability on active examples can have small information gain. Since misclassification cost of the active example is higher than the inactive one, an appropriate technique —*re-sampling technique* in this paper— should be developed.

We can define a cost-weighted information gain by re-sampling the active examples of original data. To be precise, we just duplicate the active examples at a certain rate. When we compute the information gain (i.e. in the process of the feature subset selection), we use the re-sampled data of active examples. Even though we manage to decrease the information gain of a feature that is important in predicting the inactive class, the information gain of the features with high discrimination capability on the active class have increased. This means that we can obtain a feature subset which consists of features that can predict the class with high misclassification cost.

3.3 Cross-Validation for Deciding Optimal Feature Subset

In Section 3.1 we mentioned that in order to find out whether a feature subset, which was produced by the value of δ, is an optimal one, we use 10-fold cross-validation. If we decide a point that has highest cross-validation accuracy as an optimum, then there would be high chance for biological data to be over-fitted [13]. This is because the training set occupies only a small portion of the instance space and therefore not sufficient enough to represent the instance space. So the result that is optimized on the training set might have high accuracy of

itself but it can produce bad result on the real test set. This implies that we need an alternative criterion to avoid over-fitting.

When the number of features included in the feature subset increases, due to decrease of threshold δ, we can see that the cross-validation accuracy also increases. But since most of the features of biological data are often irrelevant ones, the steep improvement of cross-validation accuracy only takes place until relevant features are included in the subset at the beginning. At later times, the improvement of cross-validation accuracy becomes steady even though features with smaller information gain are included. Therefore, the increase rate of subset's cross-validation accuracy becomes close to 0 or begins to fluctuate. From these facts, we can say that the point, where nothing changes or begin fluctuating, is the time when irrelevant features are included. For this reason, we use first optimum point of cross-validation to decide optimal feature subset. If the accuracy holds up in a similar level, choosing small features may yield more general prediction without over-fitting.

4 Experiments

To demonstrate the feasibility of proposed approaches, we conducted two experiments. The first one is to find out how well the proposed feature subset selection technique performs and whether the optimality check of the feature subset is suitable for biological data. In the second experiment, we focused on how well we can consider the misclassification cost by using example re-sampling technique. In each experiment, we obtained the accuracy by using C4.5 Algorithm and we had used weighted average accuracy (see Section 2.3) to consider the misclassification cost. The dataset used was thrombin training (introduced in Section 2.3) set and test set respectively.

4.1 Cross-Validation and Test Accuracies (without Re-sampling)

We first, using thrombin training set, computed each information gain of the features and organized a feature subset by selecting features with highest information gain. We set up the threshold to have value between 0.014 and 0.0055 and set the interval to decrease by 0.0005. The initial threshold value and the interval are determined by the characteristic of the data so we can say the value is arbitrary. The number of feature subset generated was 17 and the number of features in each feature subset was between 5 and 2932. Next, we computed the 10-fold cross-validation accuracy of the feature subset and as the cross-validation accuracy changed, we determined the optimal feature subset by using the criterion we presented in Section 3.3. The cross-validation accuracy of the training set is depicted in Fig. 1 (bold line).

In the figure, we can see that the accuracy of the feature subset monotonically increases until point 49. But at point 83, the accuracy becomes smaller compare to point 49 and after point 83 we can see that the accuracy fluctuates with similar values. Therefore, if we use the previous criterion, we can make the

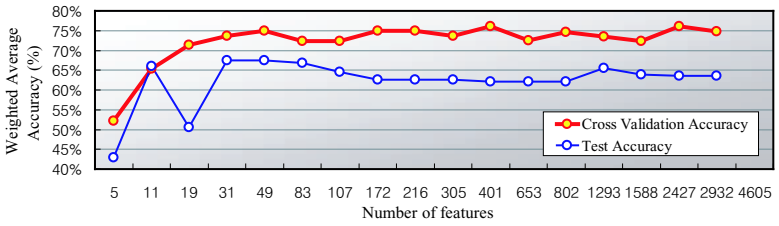

Fig. 1. Cross-validation and Test Accuracies (without Re-sampling)

optimal feature subset to be the subset with 49 features. Next, to verify whether the determined feature subset is an optimal one, we measured the test accuracy of the feature subset. The result is also depicted in Fig. 1 (thin line).

We can see that the change of test accuracy is similar to the one of cross-validation accuracy. But feature subset with highest cross-validation accuracy does not produce the highest test accuracy. As we can see, feature subset with 49 features produces the highest test accuracy compared to the one with 31 features. The feature subset shows that it is identical with the previously determined optimal subset and this implies that the criterion we used to determine the optimal subset was appropriate. By using the feature subset, we acquired test accuracy of 75.39% (unweighted) and weighted average accuracy of 67.55%.

4.2 Accuracies of Feature Subset with Re-sampling

In the second experiment, we tried to verify whether we could improve the accuracy when using re-sampling technique. First, to determine a suitable ratio of re-sampling, we generated feature subset by using various value of re-sampling ratio. The value used here is between 2 and 8. (i.e. we duplicated active examples by multiples of 2 through 8). Also, we computed the cross-validation accuracy of the subset generated by the method described. From the result, we discovered that when the re-sampling ratio was 3 or higher, there was a steep decline of cross-validation accuracy. Therefore, we had omitted the result and instead, in Table 1, we have shown the cross-validation accuracy of the subset that was generated by using re-sampling ratio 2 and 3.

In the table, we can see that the approach that did not use re-sampling tends to be better when the number of features considered are small. But as the

Table 1. Cross-Validation Accuracy according to Re-sampling Ratio

Number of Features	without Re-sampling	Re-sampling P*2	Re-sampling P*3
31	73.68 %	55.95 %	55.95 %
49	74.89 %	55.95 %	55.95 %
83	72.40 %	72.19 %	70.50 %
216	74.92 %	78.20 %	68.50 %

number of features being considered increase (i.e. more than 83 features), the cross-validation accuracy when re-sampling ratio is 2 tends to be higher than others. Generally, the cross-validation accuracy is in the highest when the ratio is 2, so we generated several feature subsets fixing the ratio to 2. We computed the cross-validation accuracy of each feature subset and depicted the result in Fig. 2 (bold line).

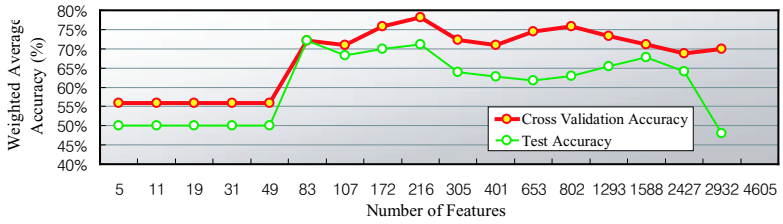

Fig. 2. Cross-Validation and Test Accuracies (Re-sampling Ratio 2)

In Fig. 2, we can see that the fluctuation of cross-validation accuracy is larger than the one without using re-sampling. However, as in the case that did not use re-sampling, the improvement of cross-validation accuracy occurs only until a certain level. From the result, we can determine the feature subset with 83 features or 216 features as an optimal feature subset. To verify whether the determined feature subset is optimal, we measured the test accuracy of each subset. The result is also shown in Fig. 2 (thin line).

The result shows that we acquired the highest accuracy when the subset has 83 features and the second highest accuracy when the subset has 216 features, as we had expected. It can be confirmed that both two subsets yield good test accuracy, although there were difficulties to decide optimal subset clearly, because of the fluctuation of cross-validation accuracy. When the number of features in the subset is 83 and 216, the test accuracy (unweighted) is 80.28% and 79.02%. Also weighted average accuracy is 72.13% and 71.08%, respectively. There is significant improvements comparing this result with KDD Cup 2001 winner, in both unweighted accuracy and weighted average accuracy.

5 Conclusion

In this paper, we proposed approaches to handle the issues of *feature subset selection, cost-sensitive classification*, and *over-fitting avoidance* to solve the problems of classifying large dimensioned, imbalanced, and non-representational data, which are the characteristics of most biological data. These techniques are already in use in different fields, but the significance of this paper is that we have shown how to effectively classify complex biological data only using the simple techniques. Experiments with thrombin data are conducted and produced outstanding results.

The performance of the winner in KDD Cup 2001 (Task 1) is an unweighted accuracy of 71.1% and weighted average accuracy of 68.4%. Through the proposed methods in this paper, we can achieved significant improvement in both of unweighted accuracy of 80.28% and weighted average accuracy of 72.13%. In many problems of classifying large dimensional, biased, and non-representative data, this entropy-based feature subset selection and accuracy estimation using cross-validation techniques can be used for good data-preprocessing as well as accurate classification.

References

[1] Christos Hatzis, David Page(2001), KDD-2001 Cup The Genomics Challenge.

[2] Cynthia Gibas, Per Jambeck : Developing Bioinformatics Computer Skills, O'Reilly (2001).

[3] Siedlecki, W and Sklansky, J. (1988). On automatic feature selection. International Journal of Pattern Recognition, 2:197-220.

[4] Langley, P. (1994). Selection of relevant features in machine learning. In Proceedings of the AAAI Fall Symposium on Relevance, pages 1-5, New Orleans, LA. AAAI Press.

[5] Dash, M. and Liu, H. (1997). Feature selection for classification. Intelligent Data Analysis, 1(3).

[6] Yang, J. and Honavar, V. (1997). Feature Subset Selection Using A Genetic Algorithm. In: Proceedings of the GP-97. Stanford, CA. pp. 380-385.

[7] Nucciardi, A. and Gose, E. (1971). A comparison of seven techniques for choosing subsets of pattern recognition. IEEE Transactions on Computers, 20:1023-1031.

[8] Roberto Battiti. (1994). Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Transaction on Neural Networks, Vol. 5, No. 4, July, pages 537-550.

[9] Al-Ani, A., Deriche, M.(2002). Feature selection using a mutual information based measure. Pattern Recognition, 2002. Proceedings. 16th International Conference on, Volume: 4 , 2002. Pages 82-85.

[10] Siedlecki, W. and Sklansky, J. (1989). A note on genetic algorithms for large-scale feature selection. IEEE Transactions on Computers, 10:335-347.

[11] Brill, F., Brown, D., and Martin, W. (1992). Fast Genetic selection of features for neural network classifiers. IEEE Transactions on Neural Networks, 3(2):324-328.

[12] Richeldi, M. and Lanzi, P. (1996). Performing effective feature selection by investigating the deep structure of the data. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pages 379-383. AAAI Press.

[13] Ng, A. Y. (1997). Preventing "over-fitting" of cross-validation data. In Proceedings of the 14th International Conference on Machine Learning (ICML), pp.245-253, Nashvilli, TN.

[14] Ron Kohavi. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Conference on Artificial Intelligence(IJCAI).

[15] Richard O. Duda, Peter E. Hart, David G. Stork : Pattern Classification, Wiley Interscience (2001).