Active Learning with Support Vector Machine Applied to Gene Expression Data for Cancer Classification

Ying Liu*

Georgia Institute of Technology, College of Computing, Atlanta, Georgia 30322

Received June 10, 2004

There is growing interest in the application of machine learning techniques in bioinformatics. The supervised machine learning approach has been widely applied to bioinformatics and gained a lot of success in this research area. With this learning approach researchers first develop a large training set, which is a time-consuming and costly process. Moreover, the proportion of the positive examples and negative examples in the training set may not represent the real-world data distribution, which causes concept drift. Active learning avoids these problems. Unlike most conventional learning methods where the training set used to derive the model remains static, the classifier can actively choose the training data and the size of training set increases. We introduced an algorithm for performing active learning with support vector machine and applied the algorithm to gene expression profiles of colon cancer, lung cancer, and prostate cancer samples. We compared the classification performance of active learning with that of passive learning. The results showed that employing the active learning method can achieve high accuracy and significantly reduce the need for labeled training instances. For lung cancer classification, to achieve 96% of the total positives, only 31 labeled examples were needed in active learning whereas in passive learning 174 labeled examples were required. That meant over 82% reduction was realized by active learning. In active learning the areas under the receiver operating characteristic (ROC) curves were over 0.81, while in passive learning the areas under the ROC curves were below 0.50

1. INTRODUCTION

Machine learning is an automatic and intelligent learning technique, which has been widely used to solve many real-world and complex problems. Since their introduction to the bioinformatics community, machine learning approaches helped to accelerate several major researches, such as biomolecular structure prediction, gene finding, genomics and proteomics. Because machine learning techniques are efficient and inexpensive in solving bioinformatics problems, the applications of these approaches in bioinformatics are becoming popular and continue to develop.[1] Shavlik et al. (1995)[2] described the field of molecular biology as tailor-made for machine learning.

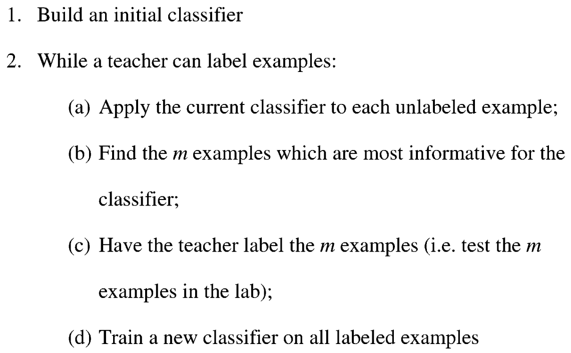
Generally, there are two types of learning schemes in machine learning: supervised learning where the output has been given a priori labeled or the learner has some prior knowledge of the data; and unsupervised learning where no prior information is given to the learner regarding the data or the output. The overall tasks for the learner are to classify, characterize, and cluster the input data. Supervised learning, such as classification, is the most common task in biological problem where given two different sets of examples, namely positive E^+ and negative E_- examples ($E^+ \cdot E^- = \O$), the learner needs to construct a classifier to distinguish between the positive examples and the negative ones. This classifier can then be used as the basis for classifying as yet unseen data in the future.[1]

The weakness of supervised learning approach is that it requires a training set, the size of which is reasonably large. Even if we have a good supervised-learning method, we cannot get high-performance without a good training set. However, labeling instances to create a training set is labor intensive and very expensive. Furthermore, when selecting training set, a bias will be introduced if the proportions of positive and negative examples do not represent the real world data distribution. The result of this bias is the concept shift between the training set and the test set. While such situations are usually avoided by machine learning researchers, it is not uncommon in the real world, where it can be difficult to obtain exactly the right training set for the intended usage of the classifier. Such concept drift can severely impair the performance of a deployed classifier.[3]

Therefore, finding ways to minimize the number of labeled instances and the difference between distribution of the training set and the real-world data is beneficial. A promising approach is active learning. Unlike most conventional learning methods where the training set used to derive the model remains static, the classifier can actively choose the training data and the size of training set increases. The classifier selects examples to be labeled, and then requests a teacher to label them and the model is recomputed based on all the examples labeled so far. It is hoped that allowing the learner this extra flexibility will reduce the learner's need for large quantities of labeled data. Pool-based active learning was introduced by Lewis and Gale (1994).[4] The learner has access to a pool of unlabeled data and can request the true class label for a certain number of instances in the pool. Active learning has been extensively studied in economic

* Corresponding author phone: (404) 385-6380; fax: (404) 894-9442; e-mail: yingliu@cc.gatech.edu.

10.1021/ci049810a CCC: $27.50
Published on Web 08/28/2004



```
1. Build an initial classifier
2. While a teacher can label examples:
    (a) Apply the current classifier to each unlabeled example;
    (b) Find the m examples which are most informative for the
        classifier;
    (c) Have the teacher label the m examples (i.e. test the m
        examples in the lab);
    (d) Train a new classifier on all labeled examples
```

Figure 1. Algorithm of pool-based active learning.

theory and statistics,[5] text categorization,[4,6,7] and drug discovery.[3,8]

Cancer classification has been the central topic of research in cancer treatment. The conventional approach for cancer classification is primarily based on the morphological appearance of the tumor. The limitations for this approach are the strong bias in identifying the tumor by experts and also the difficulties in differentiating between cancer subtypes. This is due to most cancers being highly related to the specific biological insights such as responses to different clinical treatments. It therefore makes biological sense to perform cancer classification at the genotype level compared to the phenotypic observation. Due to the large amount of gene expression data available on various cancerous samples, it is important to construct classifiers that have high predictive accuracy in classifying cancerous samples based on their gene expression profiles.[9]

We describe here the use of active learning with support vector machine to classify cancers based on gene expression profiles. We analyze data from gene expression profiles of colon cancer, lung cancer, and prostate cancer samples.

2. METHODS

2.1. Active Learning. *2.1.1. Learning Algorithm.* In pool-based active learning we have a pool of unlabeled examples. It is assumed that the examples x are independently and identically distributed according to some underlying distribution $F(x)$ and the labels are distributed according to some conditional distribution $P(y|x)$. Given an unlabeled pool U, an active learner l has three components (f, q, X). The first component is a classifier, $f\colon\ \chi \rightarrow \{-1, 1\}$, trained on the current set of labeled data X. The second component $q(X)$ is the selecting function that, given a current labeled set X, decides which examples in U to select next. The active learner can return a classifier f after each selection or after some fixed number of selections.[7]

Initially, when the training set is empty, candidates may be chosen randomly until one positive example and one negative example have been confirmed. Then the positive example and negative example are used to build an initial classifier. Figure 1 shows an algorithm of pool-based active learning.[4]

2.1.2. Batch Size m. The obvious batch size m is 1, which means the example with the largest predicted value was chosen. It has been shown that this is a good strategy to find many positive examples in a few iterations.[3,8] However, if the lab can efficiently handle a batch of $m > 1$ at a time, then the m strongest predictions would be taken to label. To evaluate the effect of the batch size (the number of examples selected to be labeled in each iteration) on the classification, in this paper, we chose $m = 1$, which meant that the most informative example was chosen, and $m = 5$.

2.2. Passive Leaning. In passive learning the next example was randomly selected to be labeled. This strategy does not make use of the examples obtained in previous iterations.[6]

2.3. Data Sets. *2.3.1. Colon Cancer.* This is a collection of gene expression profiles of 40 colon cancer and 22 normal colon tissue samples.[10] These profiles were obtained by hybridization on the Affymatrix microarray containing probes from more than 6500 genes.

2.3.2. Lung Cancer. The second group of data was a set of lung cancer samples, which was first reported by Gordon et al.[11] The authors used gene expression ratios and rationally thresholds to classify between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There were 181 tissue samples (31 MPM and 150 ADCA). The training set contained 32 of them, 16 MPM and 16 ADCA. The rest of the 149 samples were used for testing. Each sample was described by 12 533 genes. In this study, MPM was treated as positive and ADCA as negative examples.

2.3.3. Prostate Cancer. The third data set was prostate cancer samples.[12] Tumor versus Normal classification: training set contained 52 prostate tumor samples and 50 nontumor (labeled as "Normal") prostate samples with around 12 600 genes. An independent set of testing samples was from a different experiment[13] and had a nearly 10-fold difference in overall microarray intensity from the training data.

All the data in this paper were from http://sdmc.lit.org.sg/GEDatasets/Datasets. In this study, for the datasets of lung cancer and prostate cancer, the training set and testing set were combined as single datasets.

2.4. Classifier. After the training set was built either actively (the m most informative examples were chosen to label) or passively (examples were randomly chosen to label), one classifier derived a model from the training set and the model was used to classify the cancer examples. In this study, support vector machine (SVM) was use for cancer classification. SVM was found to be very suitable for the active learning setup.[6−8] Furthermore, SVM has been successfully applied to cancer classification using gene expression data.[13−16]

SVM is a kind of blend of linear modeling and instance-based learning. An SVM selects a small number of critical boundary samples, called *support vectors*, from each category and builds a linear discriminate function that separates them as widely as possible. In the case that no linear separation is possible, the technique of "*kernel*" will be used to automatically inject the training samples into a higher-dimensional space and to learn a separator in that space.[17,18] In linearly separable cases, SVM constructs a hyperplane which separates two different categories of feature vectors with a maximum margin, i.e., the distance between the separating hyperplane and the nearest training vector. The training instances that lie closest to the hyperplane are *support vectors*. The hyperplane was constructed by finding another vector w and a parameter b that minimizes $||w||^2$ and satisfies the following conditions

$$w \cdot x_i + b \geq +1, \text{ for } y_i = +1 \text{ Category 1 (positive)}$$

$$w \cdot x_i + b \leq -1, \text{ for } y_i = -1 \text{ Category 2 (negative)}$$

where y_i is the category index (i.e., active, inactive), w is a vector normal to the hyperplane, $|b|/||w||$ is the perpendicular distance from the hyperplane to the origin, and $||w||^2$ is the Euclidean norm of w. After the determination of w and b, a given vector x can be classified by sign[($w \cdot x$) + b].[19]

In this paper SVMLight v.3.5 was used.[17] Linear kernel was applied.

2.5. Performance Evaluation. In this study receiver operating characteristic (ROC) was used to evaluate the performance of active learning.

ROC analysis has long been used in clinical applications to evaluate the usefulness of diagnostic tests.[20,21] Recently, ROC analysis has been increasingly recognized as an important tool for evaluation and comparison of classifiers.[22] Research has shown that ROC analysis offers more robust evaluation of the relative prediction performance of alternative models than traditional comparison of relative errors.[23-28]

ROC includes elements of both sensitivity and specificity.[29,30] The ROC is evaluated by means of a plot of the true positive rate (sensitivity) vs the true negative rate (1 − specificity). The true positive and false positive rates are defined as follows

$$\text{true positive rate} = (\text{true positives predicted})/(\text{total positives in the data set})$$

$$\text{false positives rate} = (\text{false positives predicted})/(\text{total negatives in the data set})$$

The area under the ROC curve (AUC) measures the probability of correct classification, so its values lie between 0 (worst) and 1 (best). The closer AUC is to 1, the better the overall classification performance of the test, and a test with an AUC value of 1 is one that is perfectly accurate. An area of 0.9, for instance, indicates that an example chosen from the positive group has a probability of 0.9 of predicted value higher than an example chosen from the negative group.

3. RESULTS

3.1. Active Learning vs Passive Learning in Cancer Classification. Figure 2 shows the number of positives found after each round of learning. Either active learning or passive learning carried to the end would finish with all positives found. To provide an upper bound, we also plotted the number of positives of the unrealistic optimal selection strategy, which chose purely positives in the pool until all positives were selected. In this ideal case the learner identified all the positives at the beginning, yielding a 45° slope until the positives were exhausted. In passive learning the example was randomly selected to be labeled. The number of positives grew only linearly with the number of iterations. For each of the three data sets tested, active learning outperformed passive learning. For lung cancer classification, before all 31 positives were found, only 2 false

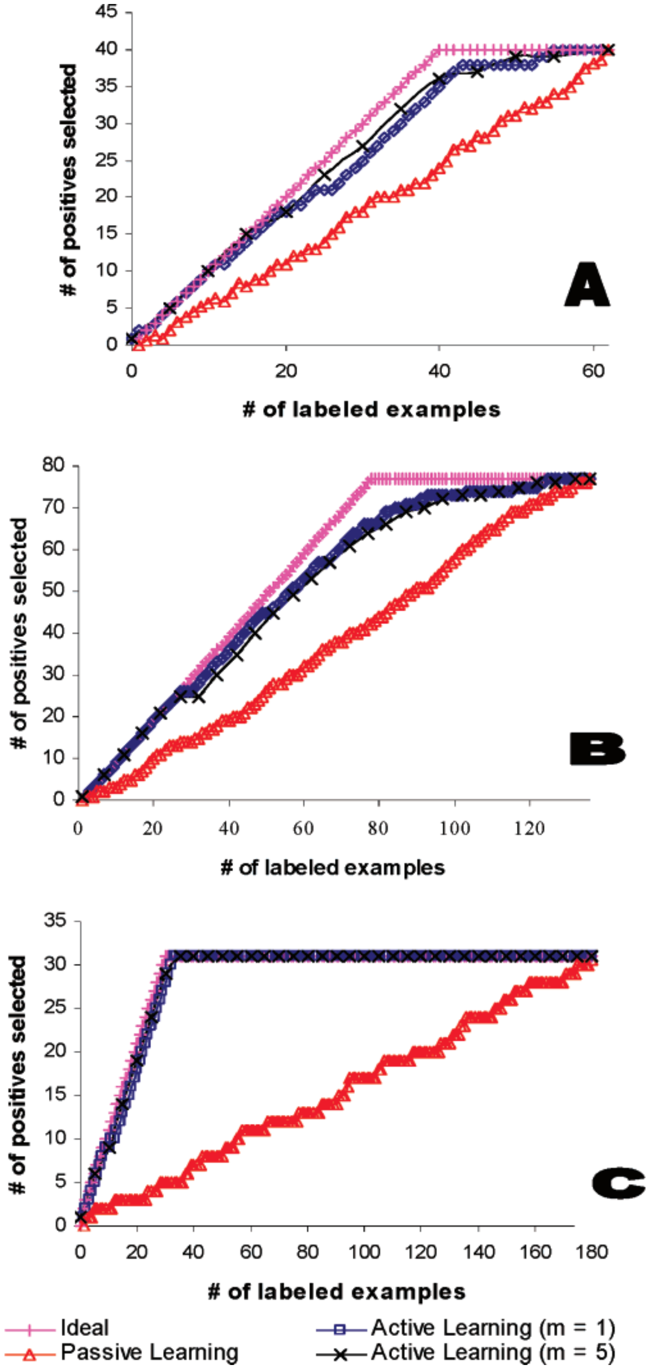


Figure 2. Number of positive examples of colon cancer (A), prostate cancer (B), and lung cancer (C) found after each iteration. Initially a positive example and a negative example were chosen randomly to form the training set, which was used to build an initial classifier. The classifier was applied to each unlabeled example, m examples that were most informative for the classifier were selected to add to the training set, and a new classifier was derived from all labeled examples. To evaluate the effect of the batch size (the number of examples selected to be labeled in each iteration) on the classification, in this paper, we chose $m = 1$, which meant that the most informative example was chosen, and $m = 5$. To provide an upper bound, we also plotted the number of positives of the unrealistic optimal selection strategy, which chose purely positives in the pool until all positives were selected. In this ideal case, the learner identified all the positives at the beginning, yielding a 45° slope until the positives were exhausted. In passive learning the example was randomly selected to be labeled. The number of positives grew only linearly with the number of iterations. For each of the three data sets tested, active learning outperformed passive learning.

Table 1. Number of Labeled Examples Required to Find a Majority of the Positives

	positives found	active learning ($m = 1$)	active learning ($m = 5$)	passive learning
colon cancer	50%	23	25	31
	90%	41	40	57
	96%	43	45	62
prostate cancer	50%	42	47	68
	90%	84	92	119
	96%	101	112	127
lung cancer	50%	16	15	90
	90%	29	30	157
	96%	31	31	174

positives were selected. Therefore, the curve was almost identical to that of the ideal case (Figure 2C).

3.2. Active Learning Reduced Cost. Active learning significantly reduced the cost (number of examples labeled) to obtain a majority of the positives (Table 1). For lung cancer classification, to achieve 96% of the total positives only 31 labeled examples were needed in active learning with $m = 1$ or 5, whereas in passive learning 174 labeled examples were required. That meant over 82% reduction was realized by active learning. For the other two data sets (prostate cancer and colon cancer), active learning also reduced the cost.

3.3. Receiver Operating Characteristic (ROC) Curve. To present these results in more familiar terms, we expressed them as ROC curves (Figure 3). The areas under the ROC curves (AUC) are shown in Table 2. In active learning the areas under the ROC curves were over 0.81, while in passive learning the areas under the ROC curves were below 0.50. For lung cancer classification, the area under the ROC curve was 0.99 when active learning was applied with $m = 5$, whereas when passive learning was used the area under the ROC curve was 0.49.

4. DISCUSSION

Machine learning is the subfield of artificial intelligence which focuses on methods to construct computer programs that learn from experience with respect to some class of tasks and a performance measure.[31] Machine learning methods are suitable for molecular biology data due to the learning algorithm's ability to construct classifiers/hypotheses that can explain complex relationships in the data. Therefore, machine learning has increasingly gained attention in bioinformatics research. Cancer classification based on gene expression data remains a challenging task in identifying potential points for therapeutics intervention, understanding tumor behavior, and also facilitating drug development.[9]

4.1. **Active Learning Outperformed Passive Learning.** Traditional supervised learning, such as classification, requires a large training set. Labeling examples to create a training set is time-consuming and costly. Furthermore, the training set is chosen to be a random sampling of examples, which is similar to passive learning in this paper.[7] However, in many cases active learning can be employed. In this report we presented the experimental results showing that active learning consistently performed better than passive learning over all three tested data sets. By applying active learning, much fewer labeled examples were required to achieve similar accuracy (Table 1). Moreover, the areas under the ROC curves for active learning were significantly greater than the area under the ROC curve for passive learning (Table 2). Therefore, active learning considerably reduces

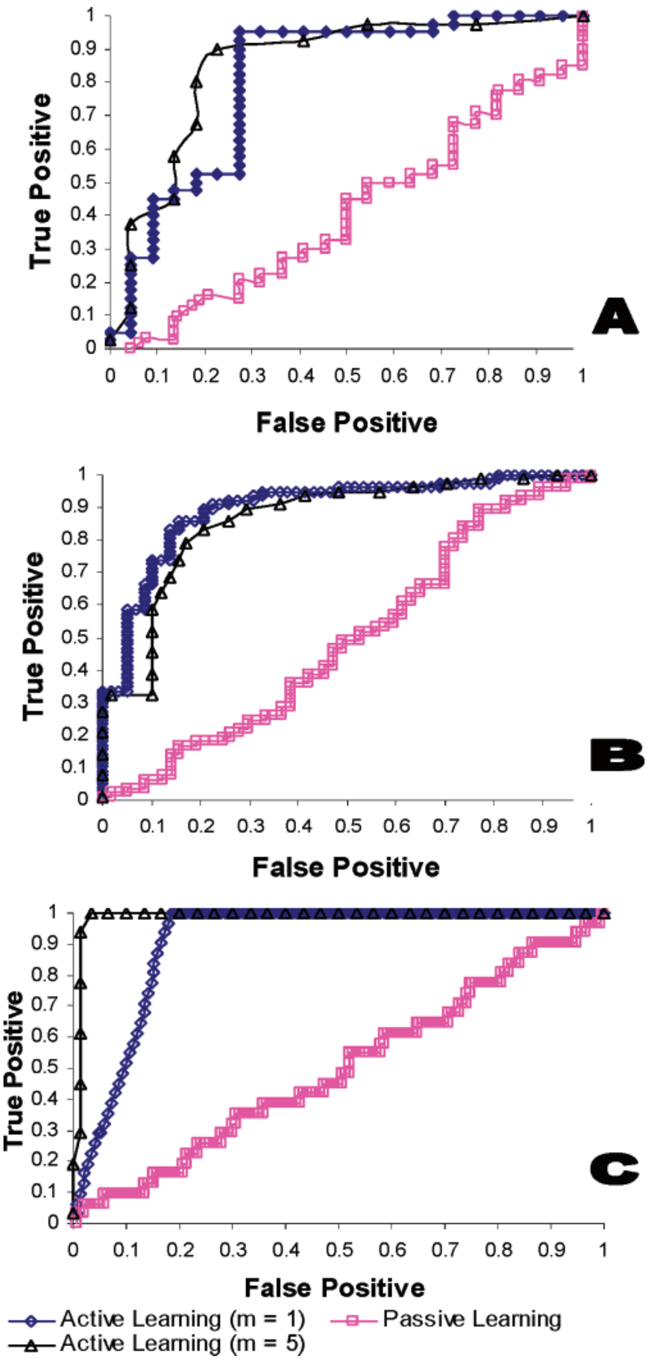


Figure 3. Receiver operating characteristic (ROC) curves for colon cancer (A), prostate cancer (B), and lung cancer (C). In active learning the areas under the ROC curves were over 0.81, while in passive learning the areas under the ROC curves were below 0.50. For lung cancer classification the area under ROC curve was 0.99 when active learning was applied with $m = 5$, whereas when passive learning was used the area under ROC curve was 0.49.

Table 2. Areas under the ROC Curves

	active learning (m=1)	active learning (m=5)	passive learning
colon cancer	0.85	0.81	0.41
prostate cancer	0.90	0.86	0.49
lung cancer	0.91	0.99	0.49

manual labeling cost while keeping and even improving performance.

4.2. **Concept Drift.** In traditional classifiers the training set is often drawn from an earlier point in time or from a restricted subset of geographical samples, having a somewhat different distribution than encountered in actual use. The sample distribution in the training set can significantly affect the derived model and eventually the quality of the classification. Concept drift between training set and test set is an important factor which can impair the performance of a classifier. Complete immunity from concept drift is impossible; however, active learning has significantly less exposure to this risk since it is continuously learning on the very examples it is being used to classify and the pool of available examples is changed over time and takes on a different character from the earlier examples.[3] Furthermore, active learning is a general framework and does not depend on tasks or domain.[6]

4.3. **Effect of Batch Size on Classification.** There are different strategies to select the example to be labeled. One obvious strategy is to select the example with the largest predicted value (largest positive strategy). Warmuth et al.[8] tested different selection strategies and showed that the largest positive strategy is a good strategy to find many positive examples in a few iterations. To evaluate the effect of the batch size (the number of examples selected to be labeled in each iteration) on the classification, in this paper, we tested $m = 1$, which meant that the most informative example was chosen, and $m = 5$. The results showed that the classification with $m = 5$ had similar performance as that with $m = 1$ but had a much smaller number of iterations. Similar results were reported by Forman[3] when active learning was applied for drug discovery. Therefore, larger batch size can be used to achieve good performance with fewer iterations.

5. CONCLUSION

Active learning has been an active research topic in machine learning but is still relatively new to the bioinformatics community. Most of the machine-learning-oriented bioinformatics literature still largely concentrates on traditional learning approaches. Active learning is suitable for bioinformatics applications due to the fact that the classifiers have the freedom to choose the training data and reduce the risk of concept drift. We demonstrated that active learning with support vector machines can accurately classify cancers based on expression data from DNA microarray hybridization experiments and presented some theoretical explanations on the performance of active learning. We believe this approach has significant potential and should be considered for the task of classifying gene expression data for cancerous samples.

Abbreviations: SVM: support vector machine. ROC: receiver operating characteristic. AUC: area under ROC curve.

ACKNOWLEDGMENT

The author is grateful to the editor and the two anonymous reviewers for a number of suggestions for improvement.

REFERENCES AND NOTES

(1) Tan, A. C.; Gilbert, D. An empirical comparison of supervised machine learning techniques in bioinformatics. *Proceedings of First Asia Pacific Bioinformatics Conference (APBC 2003)*, 2003; pp 219−222.

(2) Shavlik, J.; Hunter, L.; Searls, D. Introduction. *Machine Learning* **1995**, *21*, 5−10.

(3) Forman, G. Incremental machine learning to reduce biochemistry lab costs in the search for drug discovery. BIOKDD02. **2002**, 33−36.

(4) Lewis, D.; Gale, W. A sequential algorithm for training text classifiers. *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*; 1994; pp 3−12.

(5) Federov, V. V. *Theory of optimal experiments*; Academic Press: New York, 1972.

(6) Sassano, M. An empirical study of active learning with support vector machines for Japanese word segmentation. *Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA; 2002; pp 505−512.

(7) Tong, S.; Koller, D. Support vector machine active learning with application to text categorization. *J. Machine Learning Res.* **2001**, *2*, 45−66.

(8) Warmuth, M. K.; Ratsch, G.; Mathieson, M.; Liao, J.; Lemmen, C. Support vector machines for active elarning in the drug discovery process. *Adv. Neural Inf. Proc. Sys.* **2002**, *14*, 1449−1456.

(9) Tan, A. C.; Gilbert, D. Ensemble machine learning on gene expression data for cancer classification. *Appl. Bioinformatics* **2003**, *2*, S75−S83.

(10) Alon, U.; Barkai, N.; Notterman, D. A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A. J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci., U.S.A.* **1991**, *96*, 6745−6750.

(11) Gordon, G. J.; Jensen, R. V.; Hsiao, L.-L.; Gullans, S. R.; Blumenstock, J. E.; Ramaswamy, S.; Richards, W. G.; Sugarbaker, D. J.; Bueno, R. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression rations in lung cancer and mesothelioma. *Cancer Res.* **2002**, *62*, 4963−4967.

(12) Singh, D.; Febbo, P. G.; Ross, K.; Jackson, D. G.; Manola, J.; Ladd, C.; Tamayo, P.; Renshaw, A. A.; D'Amico, A. V.; Richie, J. P.; Lander, E. S.; Loda, M.; Kantoff, P. W.; Golub, T. R.; Sellers, W. R. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **2002**, *1*, 203−209.

(13) Lee, Y.; Lee, C. K. Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics* **2003**, *19*, 1132−1139.

(14) Liu, H. X.; Zhang, R. S.; Luan, F.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Diagnosing breast cancer based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 900−907.

(15) Chang, R. F.; Wu, W. J.; Moon, W. K.; Chou, Y. H.; Chen, D. R. Support vector machines for diagnosis of breast tumors on US images. *Acad Radiol.* **2003**, *10*, 189−197.

(16) Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **2000**, *16*, 906−914.

(17) Joachims, T. Transductive inference for text classification using support vector machine. *International Conference on Machine Learning (ICML'99)*; 1999; pp 200−209.

(18) Liu, H.; Li, J.; Wong, L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genomic Informatics* **2002**, *13*, 51−60.

(19) Liu, Y. A Comparative Study on Feature Selection Methods for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2004**, accepted for publication.

(20) Zweig, M. H.; Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **1993**, *39*, 561−577.

(21) Park, S. H.; Goo, J. M.; Jo, C.-H. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Kor. J. Radiol.* **2004**, *5*, 11−18.

(22) Ferri, C.; Flach, P.; Hernandez-Orallo, J. Learning decision trees using the area under the ROC curve. *Proceedings of the 19th International Conference on Machine Learning*; 2002; pp 139−146.

(23) Webb, G. I.; Ting, K. M. On the Application of ROC Analysis to Predict Classification Performance Under Varying Class Distributions. *Machine Learning* **2004**, in press.

(24) Weinstein, M. C.; Fineberg, H. V. *Clinical Decision Analysis*; Saunders: Philadelphia, PA, 1980.

(25) Duda, O. R.; Hart, P. E.; Stork, D. G. *Pattern Classification*; John Wiley: New York, 2001.

(26) Adams, N. M.; Hand, D. J. Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition* **1999**, *32*, 1139−1147.

(27) Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* **1997**, *30*, 1145−1159.

(28) Provost, F.; Fawcett, T.; Kohavi, R. The case against accuracy estimation for comparing induction algorithms. *Proceedings of The Fifteenth International Conference on Machine Learning*; 1998; pp 43−48.

(29) Gribskov, M.; Robinson, N. L. Use of receiver operative characteristic (ROC) analysis to evaluate sequence maching. *Comput. Chem.* **1996**, *20*, 25−33.

(30) Schaffer, A. A.; Aravind, L.; Madden, T. L.; Shavirin, S.; Spouge, J. L.; Wolf, Y. I.; Koonin, E. V.; Altschul, S. F. Improving the accuracy of PSI−BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **2001**, *29*, 2994−3005.

(31) Mitchell T. *Machine learning*; McGraw-Hill: New York, 1997.

CI049810A