Reducing False Positives in Molecular Pattern Recognition

Xijin Ge[1]
xge@genome.rcast.u-tokyo.ac.jp

Shuichi Tsutsumi[1]
shuich@genome.rcast.u-tokyo.ac.jp

Hiroyuki Aburatani[1]
haburata-tky@umin.ac.jp

Shuichi Iwata[2]
iwata@q.t.u-tokyo.ac.jp

[1] Genome Science Division, Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan
[2] Department of Quantum Engineering and Systems Science, School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Abstract

In the search for new cancer subtypes by gene expression profiling, it is essential to avoid misclassifying samples of unknown subtypes as known ones. In this paper, we evaluated the false positive error rates of several classification algorithms through a ‘null test’ by presenting classifiers a large collection of independent samples that do not belong to any of the tumor types in the training dataset. The benchmark dataset is available at `www2.genome.rcast.u-tokyo.ac.jp/pm/`. We found that k-nearest neighbor (KNN) and support vector machine (SVM) have very high false positive error rates when fewer genes (<100) are used in prediction. The error rate can be partially reduced by including more genes. On the other hand, prototype matching (PM) method has a much lower false positive error rate. Such robustness can be achieved without loss of sensitivity by introducing suitable measures of prediction confidence. We also proposed a cluster-and-select technique to select genes for classification. The nonparametric Kruskal-Wallis H test is employed to select genes differentially expressed in multiple tumor types. To reduce the redundancy, we then divided these genes into clusters with similar expression patterns and selected a given number of genes from each cluster. The reliability of the new algorithm is tested on three public datasets.

Keywords: prototype matching, support vector machine, pattern recognition, cancer diagnosis

1 Introduction

DNA microarrays are promising tools for accurate cancer diagnosis and the searching for new cancer subtypes [1, 2, 6, 12]. Nevertheless, expression data is often very noisy because only a small portion of the genes are correlated with the distinction of tumor subtypes. Even for these genes, variances in expression level can occur for various histological or technical reasons. Additionally, the number of replicates is often limited due to difficulty in collecting human samples. The great challenge is to develop reliable algorithms that fit the needs of current situation.

The first step in such algorithms is to select a set of genes that express differentially in distinct tumor types. In the terms of pattern recognition, this is a task of feature selection that should be distinguished from classification itself. For problems concerning two cancer subtypes [6], feature selection can be done by simply looking for genes that are activated in one patient group while suppressed in the other. In the multi-class problems involving three or more tumor types, a direct approach is to combine multiple pair-wise comparisons with the *all-vs-all* or *one-vs-all* strategy. Due to its simplicity the *one-vs-all* approach has been employed in several studies [9, 11, 12, 15], in which genes specifically expressed in one tumor type are selected. Selection of genes can also be performed in an iterative manner by observing the performance of a classifier [8].

In this paper we take a different approach for feature selection. Instead of searching for genes with predefined expression patterns, an unsupervised procedure is proposed to select all existing patterns of expression that could be useful in classification. This is made possible by introducing clustering analysis techniques, such as K-means clustering, to feature selection.

For the classification of tumors, many machine learning algorithms are available. Besides the simple methods like weighted voting scheme [6] and K nearest neighbor (KNN), support vector machine (SVM) has been widely used by many researchers. Khan *et al.* demonstrated the application of artificial neural networks for discriminating four subtypes of the small, round blue cell tumors (SRBCTs) of childhood [8]. Nevertheless, some comparative studies seem to suggest that simple algorithms tend to have a higher reliability than more complicated ones [4].

In the choice of classification algorithms, we find it important to ask the following questions. If a classifier is trained to discriminate, for example, two subtypes of leukemia, what kind of prediction will it produce for a sample of a newly discovered subtype it has never seen? What if the classifier is presented with normal tissues, or even tissues of stomach cancer? Ideally, these samples should not be classified as either of the two subtypes; otherwise, it would be counted as false positive. Therefore, the above questions lead to the test of false positives. Validation of classifiers in previous studies has been mainly focusing on the false negative cases as most samples for independent tests belong to one of the training subtypes. Despite the importance of avoiding false positives, especially in the process of defining new cancer subtypes and in the detection of metastatic cancers, extensive test of false positive error rates of various classification schemes have not been reported in the literature.

In this paper, we test the false positive rates of various classification schemes through a 'null test' in which a classifier is presented with a large number of samples that do not belong to any of the tumor types in the training dataset. To achieve a relatively large dataset, data from 239 microarray experiments performed in several laboratories are pooled together to test the false positive of one classifier. We compare both the false positive and false negative error rate of KNN, SVM, and prototype matching (PM), which is perhaps the simplest pattern recognition technique.

2 Method

The whole process of our approach is summarized in Fig. 1A. After pre-processing, a feature selection procedure is applied to select informative genes, which is used pattern classification by PM. Finally, the reliability is evaluated through a series of tests.

2.1 Statistical Feature Selection

Kruskal-Wallis H test. Kruskal-Wallis H test is the non-parametric counterpart of analysis of variance (ANOVA), which is a standard statistical tool for detecting differences in multi-group comparison. We choose the non-parametric test because it avoids making the assumption that the expression levels are normally distributed with equal variances within groups. It is believed that nonparametric statistical tests are nearly as powerful in detecting differences among populations as parametric methods when the data are normal. They are more powerful in situations where the data does not meet the underlying assumptions of parametric methods.

For each gene, a statistic H is calculated according to the ranks of its expression levels between multiple groups. The score is defined as: $H = \frac{12}{N(N+1)} \sum \frac{r_i^2}{n_i} - 3(N+1)$, where N is the number of tumor types in question, r_i is sum of ranks of tumor type i, which is represented by n_i samples. The higher H is, the higher the degree of association. The score tells us to what extent the gene expresses differently between *any* two groups. This score is directly related to P values because it follows a χ^2 distribution with $N-1$ degrees of freedom. For $N = 3$ case, if a gene's score is above the critical value of 9.21, we can tell with P<0.01 that this gene correlates significantly with group distinctions. Genes that fail to reach this level of significance are eliminated without further analysis. Note that

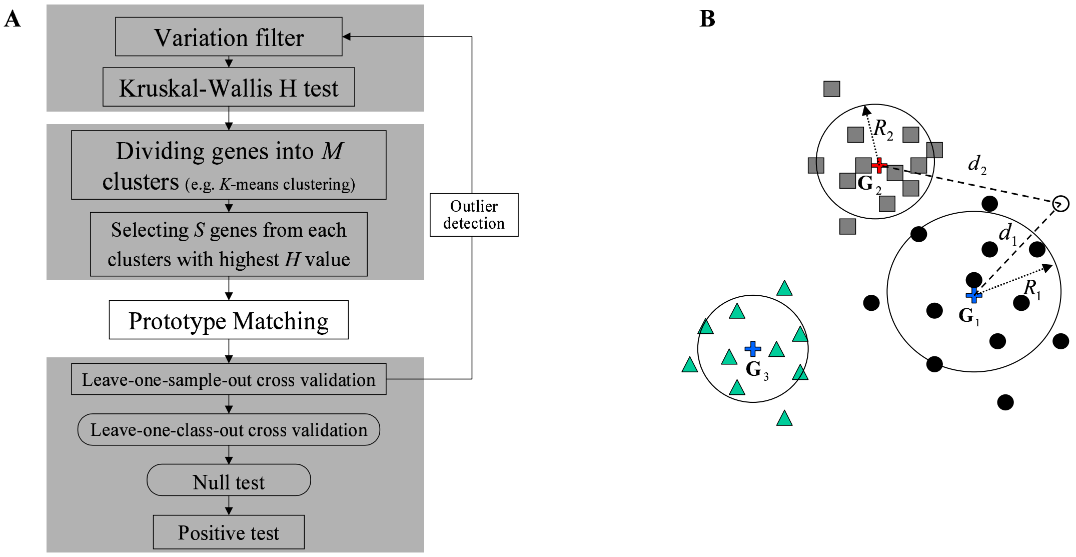

A
Variation filter
Kruskal-Wallis H test
Dividing genes into M clusters (e.g. K-means clustering)
Selecting S genes from each clusters with highest H value
Prototype Matching
Leave-one-sample-out cross validation
Leave-one-class-out cross validation
Null test
Positive test
Outlier detection
B
R2
d2
G2
d1
R1
G1
G3

Figure 1: **A,** Outline of cancer classification procedure. Based on non-parametric statistics, a cluster-and-select strategy is employed in the selection of informative genes. False positive errors are tested by null test and leave-one-class-out cross validation (LOCOCV). **B,** Prototype matching. A new sample (open circle) is compared with the existing prototypes.

the statistical significance does not decay with the increase of N. This could happen in *one-vs-all* and *all-vs-all* approaches, where the selection of informative genes is based on $O(N)$ and $O(N^2)$ statistical tests, respectively.

Redundancy reduction: classification of genes for the classification of samples. All those genes that passed the H test convey information that could be useful in classification. But still there are too many of them. We noted that many genes have very similar expression patterns. So it is possible to reduce the size of feature set without incurring classification accuracy. This is the so-called redundancy reduction problem in feature selection [7].

Another issue is that the H score does not tell us which pair-wise distinction a certain gene is associated with. It is possible that there are more genes associated with A-B distinction than those with B-C and C-A, when three subtypes A, B, and C are considered. To improve the overall accuracy of classification, the choice of genes should be made in balance.

We tackle these two problems at the same time through a cluster-and-select strategy. The idea is to select a relatively small number of representatives from each cluster of similarly expressed genes. Methods for clustering analysis have been the subject of extensive research in bioinformatics, and there exist many algorithms. Here we borrowed such techniques for the purpose of gene filtering. We used the simple K-means clustering method, which divides a set of genes into a predefined number of clusters by maximizing the between group variance. In the resultant grouping, some clusters may contain more genes than others. Nevertheless, we select a given number (S) of genes from all clusters to increase diversity of the feature set. More diversity in the feature set might contribute to robustness. Because the H score indicates the significance of association, the genes with higher scores are selected from each group.

2.2 Prototype Matching

Prototype matching is a simple method for pattern recognition. Basically, it stores prototypes and compares a new sample with them. As depicted in Fig. 1B, each tumor type is characterized by an expression prototype G_k and a radius of cluster R_k, where $k = 1, 2, 3$. Prototypes are simply calculated as the average of the expression pattern in training samples, while the radius R_k of each cluster is the average distance between samples and this prototype. Distances are calculated by $d_{ij} = 1 - P_{ij}$, where P_{ij} is the Pearson's correlation coefficients between two expression patterns i and j.

To demonstrate how PM algorithm works, we use the configuration in Fig. 1B as an example. For a new sample shown in the top right, we calculate its distance to all prototypes and find that d_1, the distance to prototype G_1, is the shortest. Therefore, it is temporarily assigned to type 1. The distance to the second nearest prototype G_2 is also calculated (d_2). The confidence of prediction can be measured by:

$$\begin{aligned} m &= (d_2 - d_1)/d_2, && (1) \\ d_r &= d_1/R_1, && (2) \\ C &= m/d_r. && (3) \end{aligned}$$

The parameter m characterize the margin of the winner prototype. For an ideal match, where $d_1 \ll d_2$, we have $m \approx 1$. By calculating the parameter d_r, we compare the distance d_1 with the radius of the prototype G_1. Ideally d_r should be about 1.0 or smaller, as R_1 is the average distance. Less typical samples will have a larger d_r. As a larger m and a smaller d_r indicates a confident prediction, it is convenient to define a confidence score as $C = m/d_r$. To make confident prediction, the C score must be larger than a certain threshold (0.08-0.15).

If C fails to reach the threshold value, a 'null' prediction is made. This may be caused by a small m, which indicates that the new sample is almost equally similar to the two best matches. This may also happen if d_r is much larger than 1. In this case, although the new sample is more similar to prototype G_1, it deviates significantly from most samples of this kind in the training set. In addition, it is also required that the Pearson's correlation coefficient between the new sample and prototype G_1 should be larger than 0.2. All together, these criteria help the algorithm avoid false positives. The simplicity of the PM algorithm makes it convenient to impose various common-sense based constraints for making predictions without significant loss of sensitivity.

These measures of prediction confidence are chosen empirically. More rigorously, one could assume that the distances to a prototype follow a Gaussian distribution. Thus the mean and standard deviation could be used to evaluate the likeliness that a new sample with distance d_i belongs to a group. Such P values can be calculated for each of the prototypes and the new sample is assigned to the one with the smallest P value. Again, both the absolute P value and the margin should be taken into account to avoid false positives. However, since the number of biological replicates within each cancer type is usually very limited, the mean and standard deviation may not be very reliable. Therefore, this approach is not used in the present study. Rather, we used the empirical formulae that are believed to be more robust for small sample size.

For comparison, we also applied KNN and SVM algorithms on the same datasets. KNN has many variations in the way that prototypes are chosen from training data and in the ways that votes are weighted [3]. Here a new sample is compared with all the samples in the training dataset. (Unlike PM, KNN uses all the samples in the training dataset as prototypes.) Then the 8-10 nearest neighbors vote with a weight of $1/r$, where r is the rank. The class that receives most votes wins. Confidence is simply characterized by the margin in the percentage of vote. The threshold to make diagnosis prediction is set to as high as 80%, which requires that most of the k neighbors should belong to one class. For SVM, we used an implementation of SVM-FU (`www.ai.mit.edu/projects/cbcl`) developed by Ryan Rifkin.

2.3 Validation: Sensitivity vs. Specificity

The false negative error rate is usually evaluated through two tests. In leave-one-sample-out cross validation (LOSOCV), each sample in the training set is withheld and used to test the performance of the classifier trained on the remaining samples. In the 'positive test', independent samples that belong to the training subtypes are presented to a classifier. These samples should be confidently assigned into one of the classes.

To evaluate the false positive error rate, we introduce a concept of 'null test'. In this test, we present a classifier samples that do not belong to any of the categories in the training dataset. Such samples can be, for example, normal tissues or those from other organs. For these samples, a reliable algorithm should produce a 'null' prediction because they should not be assigned to any of the subtypes known to the classifier. Otherwise, a false positive error is registered.

Sometimes, however, null test is impossible due to the lack of samples. An alternative procedure called leave-one-class-out cross validation (LOCOCV) is used. Withholding all the samples that belong to one tumor subtype, we train a classifier with the remaining samples. Then the classifier is tested against false positive error by presenting the samples that are left out. Basically it is a generalization of LOSOCV. The difference is that one withholds a cancer subtype instead of a sample. Note that LOCOCV is only applicable to larger datasets with more subtypes so that the elimination of one cancer subtype does not influence significantly the performance of the classifier. In this paper, we apply this procedure to a dataset consisted of 11 cancer types.

2.4 Outliers in Training Datasets

The training dataset sometimes contains outliers due to a variety of reasons such as sample preparation, array experiment, clinical diagnosis, etc. A small number of outliers in the training dataset could seriously degrade the performance of classifiers. As indicated in Fig. 1A, we eliminate such samples through a feedback loop. We reasoned that the training dataset should be consistent with itself. In our calculation, a sample is considered an outlier if (a) it is misclassified with a high C value in LOSOCV and (b) this single sample exerts un-proportionally large influence on the overall classifier. But one should be very careful because the effect of eliminating different samples might be inter-dependent. Additionally, the total number of samples to be eliminated should be kept small (less than 5%). For the detection of outliers, it would be helpful to examine the dataset with some outside programs such as hierarchical clustering and data visualization algorithms (principal component analysis or multidimensional scaling [5]).

3 Datasets and Results

Benchmark dataset uncovers hidden false positives. The first dataset is constructed to test the robustness of various algorithms against false positives. For the training dataset, we used the leukemia dataset [6], which contains expression patterns of samples for acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). As ALL samples can be further divided into two groups: T-cell lineage and B-cell lineage, we consider three subtypes in this dataset. There are 38 samples in the training dataset (11 AML, 19 B-ALL and 8 T-ALL), and 34 independent samples for positive test. Microarrays used in these experiment are Affymetrix HuFL which contains probes for 6817 transcripts.

To test the false positive rates, we incorporated HuFL datasets from several laboratories. The null test data includes an ovary dataset [14], a dataset of stomach and liver tissue samples, and a variety of other samples from our collaborators. These 239 samples are of various origins; they can be any types of human tissues except AML and ALL. See supplementary information for more information about all those samples. Expression scores on each array are normalized to have the same mean and standard deviation to make the data from different laboratories compatible.

In pre-processing, we first eliminate those genes that do not change significantly by requiring that the difference and ratio between the maximum and minimum be larger than 300 and 2, respectively [6]. We also require that the standard deviation and its ratio to mean be larger than 100 and 0.1. Those genes with a Kruskal-Wallis H score smaller than 9.21 are eliminated as their expression profiles do not correlate with tumor distinction with a statistical significance level P=0.01. The expression levels

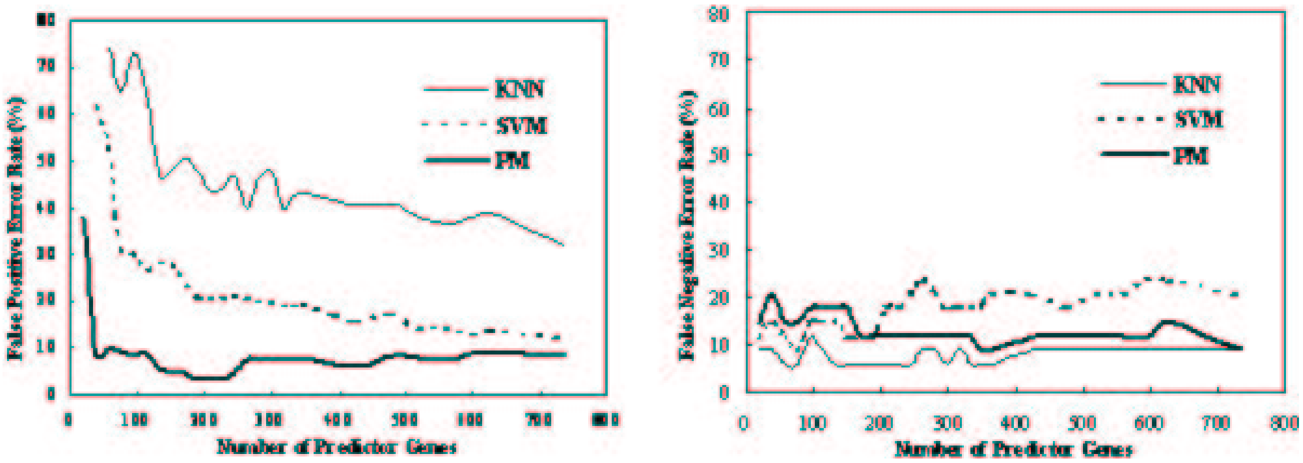


Figure 2: The change of false positive (left) and false negative (right) error rates with the number of genes used for prediction. KNN has the highest false positive error rates and the lowest false negative error rates. For SVM, although we raise the prediction threshold so high that its false negative error rate increases to about 20%, its false positive error rates are still higher than that of PM. The performance of PM are reasonable in both kinds of errors.

of the remaining 736 genes are log-transformed. Normalization is done simply by dividing the data by the length of a gene's expression vector so that each gene is characterized by a unitary vector. We found that the classification accuracy can be significantly degraded if we follow the popular way of normalization that makes all genes have the same variance.

According to their similarities, these genes are divided into 20 groups by K-means clustering. From each group we selected a small number of genes and construct a feature set. Based on these selected genes, prototype matching is used to make predictions on new samples. The threshold for the prediction confidence C is set to 0.15. By changing the number of genes selected from each cluster, the changes of false positive and false negative error rates are plotted in Fig. 2.

Surprisingly, a large difference is observed in the false positive rates for predictions made by different classification methods. When less than 100 genes are used, KNN could have a false positive error rate as high as 50%. SVM also has a relatively high error rate of about 20%. On the contrary, PM has an error rate smaller than 10%.

Even with as few as 19 predictor genes, most samples in the positive test can be correctly classified by any of the three methods. This could misleadingly suggest the use of small feature set. But null test indicates that the false positive rates could be as high as 92% for KNN, 89% for SVM and 38% for PM. When the number of predictor genes is increased we observed a quick decrease of false positive error rates. Change in false negative error rates are less sensitive except SVM, which produces slightly more false negatives when the number of genes is increased. Therefore, whatever the classification algorithms, it is important to include hundreds of genes in the feature set. This also demonstrates the importance of null test: those seemingly irrelevant datasets serve as a background based on which we can tell whether a feature set enables the unique definition of expression prototype for a certain cancer type.

From Fig. 2, we also observed that the error rate of PM is not sensitive to the number of genes used in prediction, as long as the number is not too small. We finally select 227 genes in our final feature set through optimizing the balance between two kinds of errors (see supplementary information for more details). In LOSOCV, most of the samples in the training set are correctly classified except 3 false negatives. In the positive and null test, PM has 4 false negatives (11.8%) and 8 false positives(3.3%).

Figure 3 gives these informative genes. The feature set contains all 6 possible alternative expression patterns in a 3-class problem. From top to bottom in the figure, there are genes of type (1,0,1),

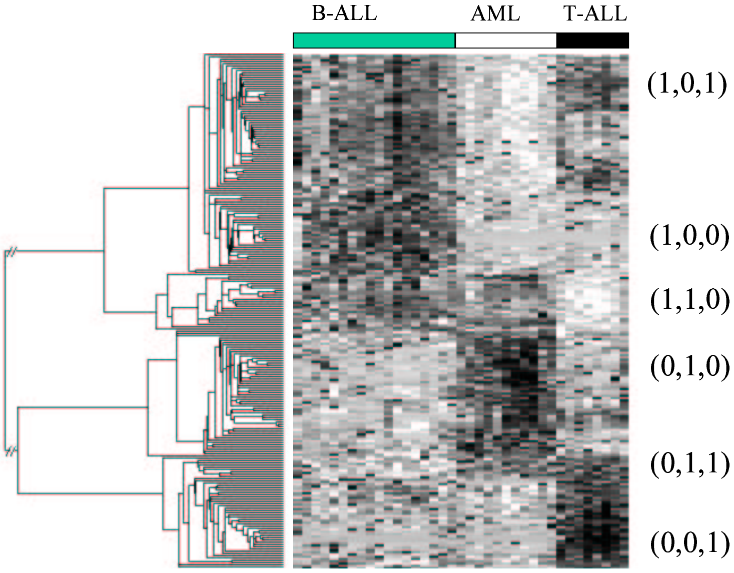


Figure 3: A set of 227 genes chosen for classification of B-cell acute lymphoblastic leukemia (B-ALL), T-cell acute lymphoblastic leukemia (T-ALL), and acute myeloid leukemia (AML)[6]. Instead of searching the whole gene list for predefined expression patterns, we try to find all existing patterns that could be helpful in cancer classification. Black indicates high expression.

(1,0,0), (1,1,0), (0,1,0), (0,1,1), (0,0,1), with 1 representing high expression and 0 low expression. Unsurprisingly, at the top region of the figure we find a large number of genes that are shared by B-ALL and T-ALL. Such genes are ignored in the one-vs-all gene selection method used by [11, 8, 9], as only genes of type (1,0,0), (0,1,0), (0,0,1) are selected. We include these genes because they reflect the true structure of similarity between tumor types and can help achieve a high classification accuracy.

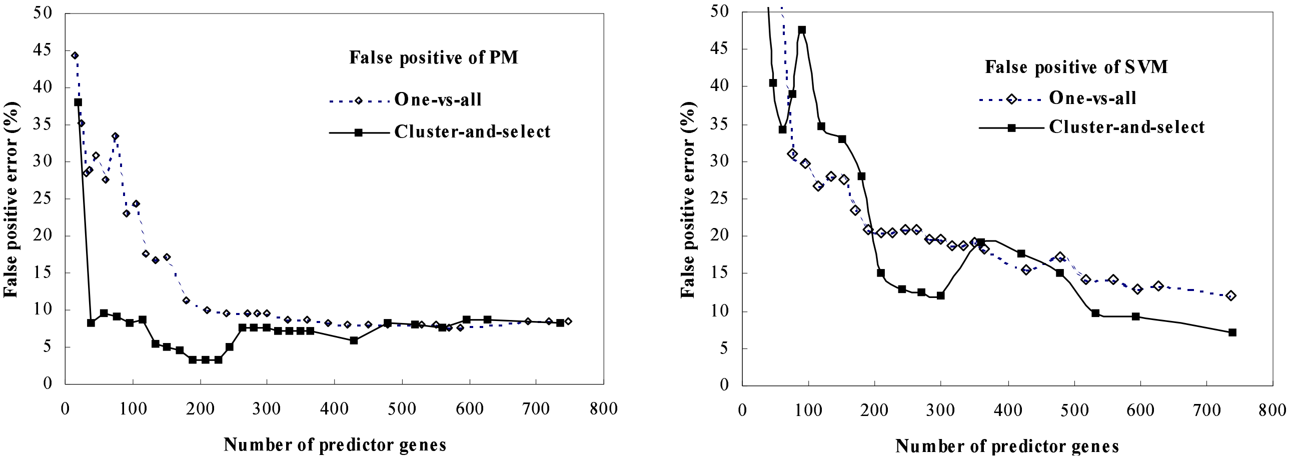


Figure 4: Comparison of feature selection methods. Prototype matching (PM) algorithm is found to be more reliable using a gene set produced by the cluster-and-select method.

To compare these two strategy of feature selection, we evaluated the false positive error rates in null test (Fig. 4). With PM, the cluster-and-select method yields more reliable predictions, especially when smaller feature sets are used. This might be attributed to the fact that the new feature selection method includes some informative genes that are useful in defining prototypes. When more than 500 genes are included in the feature set, error rates tend to be very close. When SVM is applied to these two feature sets, such tendency is not observed. SVM produces predictions with higher false positive

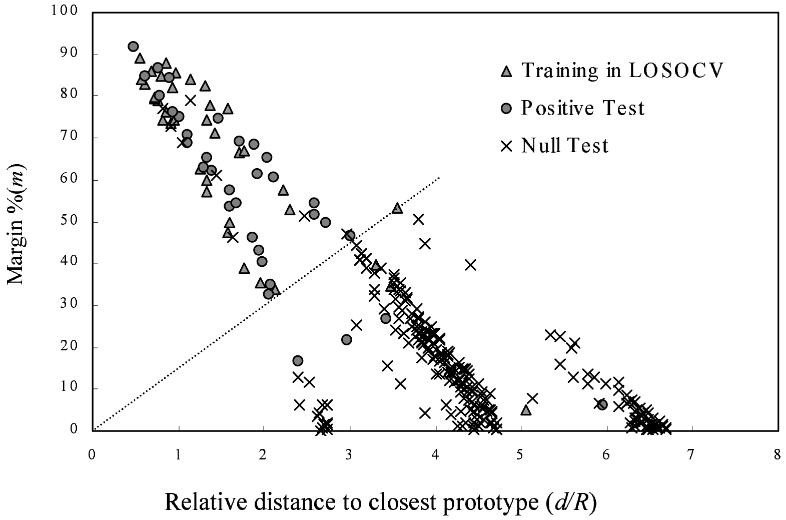


Figure 5: Distribution of samples plotted by two prediction parameters d_r and m. Parameter d_r is the relative distance to the nearest prototype and m is the margin over the second nearest. Positive predictions are made for the samples that lie above the dashed line.

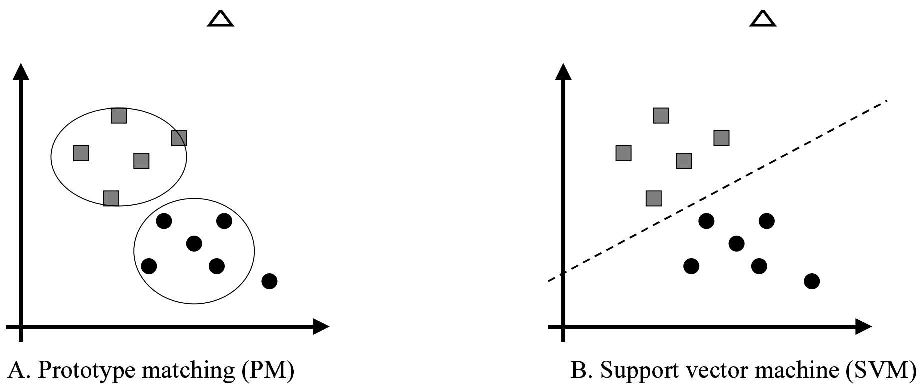
rates with both feature sets.

Finally, Fig. 5 shows the distribution of all samples with regard to prediction parameters m and d_r. These two parameters have intuitively simple meanings that could be more easily understood by biologists than parameters like vote percentage or the output of artificial neurons. The x axis is the relative distance to the nearest prototype, while the y axis represents the margin of this prototype over the second nearest one. Most samples in the training and the positive test dataset are located in regions with a large m and small d_r. On the contrary, samples in the null test dataset have a small m and a large d_r. The decision line for confident prediction $m = td_r$ is also drawn. Here $t = 0.15$ is the slope. Because both x and y axes relate to the distances of samples, we refer to such a plot as Distance-Distance (DD) plot. In general, the choice of the t can be made according to DD plots, which helps to balance false negative and false positive error.

Extensive testing on other datasets. The proposed approach for tumor classification (Fig. 1) is then tested on several other datasets including a lymphoma dataset [1], an SRBCT dataset [8], and the dataset of Su *et al.* [11]. Prediction results are summarized in Table 1. More information is available in the Supplementary Information. The proposed approach has a relatively lower rate of both false positive and false negative error in these datasets.

Table 1: Summary of four datasets and the performance of PM algorithm. Given in parentheses are the number of misclassified cases observed in leave-one-sample-out-cross-validation (LOSOCV) or independent test. Note that the false negative error rate is calculated according to both LOSOCV and positive test. For the dataset of Su *et al.*, which no independent data for null test is available, a leave-one-class-out-cross-validation (LOCOCV) procedure is employed.

	# Tumor	Samples size			False	False	# Genes
Dataset	types	Traning	Positive test	Null test	negative	positive	used
Leukemia [6]	3	37 (3)	34 (4)	239 (8)	11.8%	3.3%	227
Lymphoma [1]	3	40 (7)	26 (6)	27 (2)	19.7%	7.4%	328
SRBCT [8]	4	63 (6)	20 (2)	6 (0)	9.6%	0%	390
Su *et al.* [11]	11	97 (13)	74 (14)	—(12)	15.8%	12.4%	400



A. Prototype matching (PM)

B. Support vector machine (SVM)

Figure 6: Two paradigms of classification. Classifiers like PM tend to have more false negatives (the sample in the bottom right) while those like SVM and KNN may suffer seriously from a high false positive rates (the sample marked by a triangle at the top).

4 Discussion

There are a wealth of statistical and machine learning tools that could be useful for the classification of cancers. How to pick up the right tools and integrate them is an important issue. Such choice should be made based on knowledge of underlying computational principles. Classification algorithms like SVM and KNN define a hyperplane or hypersurface according to which a multidimensional space of samples are divided into two or more regions (Fig. 6). Implicitly, it is assumed that all samples presented to the classifier belong to at least one of the predefined tumor types. This might be true in some classification tasks such as metastatic vs. non-metastatic tumor [13], or curable vs. incurable DLBCL patients [10]. These are 'true' binary problems, in which SVM and KNN can make accurate predictions. But 'pseudo' binary classification tasks are more frequent: there could exit a third class missing in both training and test dataset. Clinically, the existence of new subtypes of cancers are always possible and it is very difficult to obtain a 'complete' training dataset as required by SVM and KNN. SVM and KNN may have high false positive rates when presented with samples of novel tumor types. This is confirmed in this study (Fig. 2).

Unlike SVM and KNN, PM defines a closely bounded region in the multidimensional space to represent each tumor subtypes(Fig. 6). There is a large rejection zone that a 'null' prediction will be outputted. The uniqueness of each subtypes is recognized by an expression prototype. Although PM can have a slightly higher false negative rate, false positive error is found to be much lower. Avoiding false positives is essential in the process of discovering new cancer subtypes. Therefore, we believe PM and methods alike might be more suitable for cancer classification.

Our results also give some hints to the question of how many predictive genes should be used in cancer classification. With the inclusion of more genes, we found that false positive errors decrease accordingly. But the opposite tendency is often observed for false negative error. Therefore, the optimal choice should be made by seeking a balance. This could be done by the minimization of the total error rate.

The searching of differentially expressed genes in two or more groups is one of the fundamentally important tasks in the bioinformatics of gene expression. Besides cancer classification, the cluster-and-select feature selection procedure proposed here might be useful in this context. The procedure is able to detect different patterns of gene expression in multiple groups.

To select features from the highly redundant measurements in expression profiling, we employed k-means clustering in addition to a statistics score. Unsupervised classification techniques themselves are used in the process of feature selection for the purpose of supervised classification. This strategy might be useful in other pattern recognition tasks where redundant measurements are involved.

Acknowledgments

We thank Shogo Yamamoto, Kunihiro Nishimura, Shumpei Ishikawa, Makoto Kano, Jerome Piat, Naohiro Shichijo, Yongxin Chen, Yingqiu Guo, Yoshitaka Hippo, and Shigeo Ihara for stimulating discussions. We are also greatly indebted to Todd R. Golub for helpful comments on our manuscript and Jiang Fu for critical proof-reading. Comments from the reviewers also helped to improve the presentation of this paper.

Supplementary Information. Supplementary information and the benchmark dataset for testing classification algorithms are availible at `www2.rcast.u-tokyo.ac.jp/pm/`.

References

[1] Alizadeh, A.A., *et al.*, The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes, *Nature*, 403:503–511, 2000.

[2] Bhattacharjee, A., *et al.*, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, *Proc. Natl. Acad. Sci. USA*, 98:13790–13795, 2001.

[3] Duda, R.O., Hart, P.E., and Stork, D.G., *Pattern classification*, (Wiley, New York, NY), 2001.

[4] Dudoit, S., Fridlyand, J., and Speed, T.P., Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Am. Stat. Assoc.*, 97:77–87, 2002.

[5] Ge, X.J., Yonamine, S., Mi, Y.M., Tsutsumi, S., Kobune, Y., Aburatani, H., and Iwata, S., A physics-inspired algorithm for information visualization with application to gene expression analysis, *J. of The School of Engineering, The University of Tokyo*, XLVII:89–103, 2000. (see also: `www.race.u-tokyo.ac.jp/~xge/CAMDA.html`)

[6] Golub, T.R., *et al.*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286:531–537, 1999.

[7] Heydorn, R.P., Redundancy in feature extraction, *IEEE Trans. Computers*, C20:1051–1054, 1971.

[8] Khan, J., *et al.*, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, 7:673–679, 2001.

[9] Ramaswamy, S., *et al.*, Multiclass cancer diagnosis using tumor gene expression signatures, *Proc. Natl. Acad. Sci. USA*, 98:15149–15154, 2001.

[10] Shipp, M.A., *et al.*, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, *Nature Medicine*, 8:68–74, 2002.

[11] Su, A.I., *et al.*, Molecular classification of human carcinomas by use of gene expression signatures, *Cancer Res.*, 61:7388–7393, 2001.

[12] Tsutsumi, S., *et al.*, Two distinct gene expression signatures in pediatric acute lymphoblastic leukemia with MLL rearrangements, *Cancer Res.*, 63:4882–4887, 2003.

[13] Van'T Veer, L.J., *et al.*, Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, 415:530–536, 2002.

[14] Welsh, J.B., *et al.*, Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer, *Proc. Natl. Acad. Sci. USA*, 98:1176–1181, 2001.

[15] Yeang, C.-H., *et al.*, Molecular classification of multiple tumor types, *Bioinformatics*, 17 Suppl.:S316–S322, 2001.