2005－BIO－1（2）
2005／7／25

二値分類器集合による遺伝子発現プロファイルからの癌サブクラス識別法

行縄 直人 †, 大羽 成征 †, 加藤 菊也 ‡, 石井 信 †

† 奈良先端科学技術大学院大学情報科学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
‡ 大阪府立成人病センター研究所
〒 537-8511 大阪府大阪市東成区中道 1-3-2

概要—遺伝子発現プロファイルによる癌の病理診断を想定対象として，新たな多クラス識別法を提案する．本手法では，多クラス識別問題を一対一ペアや一対残りペアなどのラベルの任意の組み合わせから成る 2 値分類問題群に分解し，各問題での判別結果を統合することによって最適な識別結果を得る．各 2 値分類問題における真の分類確率がクラス所属確率をパラメータとした確率モデルによって生成されると考え，これを 2 値分類器によって得られた分類確率の推定値から推定する方法を導いた．本手法を人工データおよび甲状腺がん分類問題に適用し，従来のヒューリスティクスによる投票法と並ぶ性能を達成することを示した．さらに，この分野で提案されてきたいくつかの多クラス識別法との比較を行い，本手法の優位性を示した．

Classification of Tumor Subclass from Gene Expression Profiles based on a Probabilistic Model of Combining Binary Classifiers

Naoto Yukinawa†, Shigeyuki Oba†, Kikuya Kato‡ and Shin Ishii†

†Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama-cho, Ikoma, Nara, Japan
‡Research Institute, Osaka Medical Center for Cancer and Cardiovascular Diseases,
1-3-2 Nakamichi, Higashinari-ku, Osaka 537-8511, Japan

Abstract—In cancer classification problems based on gene expression profiling, which are important for pathological diagnosis, a stable classification algorithm is required. We propose a novel probabilistic model for constructing a multi-class pattern classifier by weighted aggregation of general binary classifiers including one-versus-the-rest, one-versus-one, and others. Our model has a latent variable that represents class membership probabilities, and it is estimated by fitting it to probability estimate outputs of binary classifiers. We apply our method to classification problems of synthetic datasets and a real world dataset of gene expression profiles. We show that our method achieves comparable performance to conventional voting heuristics. In addition, our method exhibits superior performance to some other multi-class classification algorithms used in this domain.

1. Introduction

Histopathological diagnosis has been playing the decisive role in cancer diagnostics. It is usually the final answer in discrimination of malignant and benign tissues. In spite of its long history and establishment as the routine medical procedure, there are still unsolved problems. For example, some histological diagnosis depends on individual pathologists, and their diagnostic results do not necessarily match [1][2]. In addition, in some cases, diagnosis of malignant and benign tissues is difficult. An idea to complement the diagnosis by microscopic observations is application of molecular markers, usually expressed genes. The emerging approach is gene expression profiling, genome-wide analysis of gene expression. In gene expression profiling, expression levels of thousands of genes are measured by DNA microarrays or alternative techniques, and diagnostic systems using multiple genes are constructed [3].

Clinical status such as prognosis or drug resistance is the popular target, and the supervised learning theory is often applied.

Supervised pattern classification methods can be categorized into two. Those of one type have applicability to multi-class classification problems as well as binary classification problems – such as K nearest neighbours method [4], parametric mixture models [5], and Naive Bayes method. On the other hand, those of the other type have been developed in particular for binary classification problems. The most popular one is Support Vector Machine (SVM) [6] which tries to find the optimal hyperplane that separates samples of two classes with a maximum margin.

When applying a method belonging to the latter type to multi-class (M classes) classification problems, we need some devices; the following voting heuristics are frequently used: 1) prepare a set of M binary classifiers, each of which separates one class from the other classes (one-versus-the-rest, 1R), then a class is decided by voting the outputs of probability estimates derived by M binary classifiers [7]; and 2) prepare a set of $M(M-1)/2$ binary classifiers, each of which separates one from another class (one-versus-one, 11), then a class is decided by a vote done by them [8][9].

Although the voting methods have weak theoretical background, they demonstrate fairly good performance for problems in which a binary classification subproblem is well performed by a binary classifier like SVM. However, which is better to use 11 or 1R is still an unknown problem. A previous study [10] evaluated various methods for multi-class classification problems by using several published datasets of gene expression pattern vectors. They found that SVM-based methods showed overwhelming performance in most cases, but also how to choose a set of binary classifiers was problem-specific.

In this study, we propose a statistical framework for obtaining optimal decision with aggregation of binary classifiers including not only 1R and 11 but also others such as 12 and 22, in classification problems for more than two classes. Especially when the number of classes is not large, a simple voting procedure like 1R or 11 has no plausibility, and any combination can be considered. To deal with this problem, we propose a probabilistic model for aggregating binary classifiers, in which we assume class membership probabilities of each data point are consistent with set of class membership probabilities of arbitrary binary classifications. This model exhibits a natural voting mechanism by the classifiers.

2. Probabilistic model of combining binary classifiers

There are N observations $\boldsymbol{L} = \{\boldsymbol{x}^{(n)}, t^{(n)}\}_{1:N}, \boldsymbol{x}^{(n)} \in \mathcal{R}^D, t^{(n)} \in C$, where $\boldsymbol{x}^{(n)}$ and $t^{(n)}$ are the pattern vector and the true class label, respectively, of the n-th sample. $C \equiv \{1, 2, \ldots, M\}$ is a set of M class labels. The objective of a multi-class pattern classification is to predict the class label of an unknown test pattern vector based on the training dataset $\boldsymbol{L}$.

2.1. Unit binary classifiers and class probability estimates

We decompose an M-class classification problem into all possible binary classification problems by drawing two subsets of class labels, l and m ($l, m \in \tilde{2}^C, l \cap m = \emptyset$), without overlapping, from the label's power set:

$$\tilde{2}^C \equiv 2^C - \{\emptyset, C\} = \{\{1\}, \{2\}, \ldots, \{1,2\}, \ldots, \{1,2,3\}, \ldots, \}.$$

We call a pair of label subsets a "target", represented by

$$\begin{aligned} [l|m] \qquad \in B^{AA} \equiv \{&[1|2], [1|3], \ldots, [1|M], \ldots, [M-1|M], \\ &[12|3], [12|4], \ldots, [12|M], \cdots, \\ &[1 \ldots M-1|M], \cdots\}, \end{aligned}$$

where B^{AA} is the set of all possible targets. We call this set of targets "type AA". We also consider some types for the set of targets. Those are:

Type 11 One to one, i.e., $B^{11} = B_{1,1}$

Type 1R One to the rest, i.e., $B^{1\mathrm{R}} = B_{1,M-1}$

Type 1A One to a subset in the rest, i.e., $B^{1\mathrm{A}} = \bigcup_{i=1}^{M-1} B_{1,i}$

where

$$B_{j,i} \equiv \{[l|m] \,|\, l, m \in \tilde{2}^C, l \cap m = \emptyset, \#l = j, \#m = i\},$$

and $\#l$ and $\#m$ are the numbers of class labels in the subsets l and m, respectively, and $\lfloor \cdot \rfloor$ denotes the floor integer.

Provided that we have a discriminant function $f^{\boldsymbol{L}}_{[l|m]}(\boldsymbol{x}) \in \mathcal{R}$ on a target $[l|m]$, of a binary classification algorithm trained with training dataset $\boldsymbol{L}$. Let $q_{[l|m]}(\boldsymbol{x}) = \Pr\left(\boldsymbol{t} \in l | f^{\boldsymbol{L}}_{[l|m]}(\boldsymbol{x}), \boldsymbol{t} \in l \cup m\right)$ be the class membership probability after applying a specific method to the discriminant function value; in this study, we use 1-D logistic regression [11] for this conversion.

2.2. Class probabilities

Let $p_i(\boldsymbol{x}) = \Pr(\boldsymbol{t} = i|\boldsymbol{x}) \in [0, 1], \sum_{i \in C} p_i(\boldsymbol{x}) = 1$ describe the membership probability of a pattern vector $\boldsymbol{x}$ to a class label i. We also define the membership probability vector of whole class labels as $\boldsymbol{p}(\boldsymbol{x}) = \{p_i(\boldsymbol{x})\}_{i \in C}$. The membership probability of $\boldsymbol{x}$ to an arbitrary set of class labels $l \in \tilde{2}^C$, $p_l(\boldsymbol{x})$, is given by $p_l(\boldsymbol{x}) = \sum_{i \in l} p_i(\boldsymbol{x})$. If we know $\boldsymbol{p}(\boldsymbol{x})$, the $\boldsymbol{x}$'s class label is decided as $\hat{t} = \mathrm{argmax}_{i \in C}\, p_i(\boldsymbol{x})$; this decision is Bayes optimal if $p_i(\boldsymbol{x})$ gives the posterior probability of the class label i.

2.3. Probabilistic model of binary classifications

In reality, we do not know the true posterior probability $\boldsymbol{p}(\boldsymbol{x})$, but have a set of class membership probabilities $\boldsymbol{q}(\boldsymbol{x}) = \{q_{[l|m]}(\boldsymbol{x})\}_{[l|m] \in B^+}$, corresponding to the arbitrary set of binary classifiers, B^+. The problem here is to set $\boldsymbol{p}(\boldsymbol{x})$ so as to show the best fit to $\boldsymbol{q}(\boldsymbol{x})$. Given a true $\boldsymbol{p}(\boldsymbol{x})$, the true class probability of binary classification on target $[l|m]$, $\pi_{[l|m]}(\boldsymbol{x}) = \Pr(\boldsymbol{t} \in l | \boldsymbol{x}, \boldsymbol{t} \in l \cup m)$ is given by

$$\pi_{[l|m]} = \frac{p_l}{p_l + p_m}. \tag{1}$$

For simplicity, we omit the argument $(\boldsymbol{x})$ in the followings. Then, our objective is to obtain a $\boldsymbol{p}$ so that $\hat{\boldsymbol{\pi}} \equiv \{\hat{\pi}_{[l|m]}\}_{[l|m] \in B^+}$ shows the best correspondence with $\boldsymbol{q}$. To do this, we use the Kullback-Leibler divergence as similarity measure between $\boldsymbol{q}$ and $\boldsymbol{\pi}(\boldsymbol{p})$, and maximize

$$\begin{aligned} L_0(\boldsymbol{p}) &\equiv -KL(\boldsymbol{q}; \boldsymbol{\pi}(\boldsymbol{p})) \\ &= -\sum_{[l|m] \in B^+} \left\{ q_{[l|m]} \ln \frac{q_{[l|m]}}{\pi_{[l|m]}} + (1 - q_{[l|m]}) \ln \frac{1 - q_{[l|m]}}{1 - \pi_{[l|m]}} \right\}, \end{aligned} \tag{2}$$

with respect to $\boldsymbol{p}$. In addition, we introduce a Dirichlet prior $\boldsymbol{p} \sim \mathrm{Dir}(\gamma_0)$ to (2) for regularization where γ_0 is the hyper parameter, then we have a modified objective function:

$$\begin{aligned} L_1(\boldsymbol{p}) &= \sum_{[l|m] \in B^+} \left\{ q_{[l|m]} \ln p_l + q_{[m|l]} \ln p_m - \ln(p_l + p_m) \right\} + \sum_{i \in C} \gamma_0 \ln p_i + R \\ &= \sum_{k \in B^+} a_k \log p_k + \sum_{i \in C} \gamma_0 \log p_i + R, \end{aligned} \tag{3}$$

where R is a constant depending only on $\boldsymbol{q}$. We define a_k as

$$a_k \equiv \sum_{[l|m] \in B^+, k = l} q_{[l|m]} + \sum_{[l|m] \in B^+, k = m} (1 - q_{[l|m]}) - \sum_{[l|m] \in B^+, k = l \cup m} 1.$$

In the experiments in the later section, we set $\gamma_0 = 1$. The objective function (3) is maximized with respect to $\boldsymbol{p}$ under the condition $\sum_{i \in C} p_i = 1$; this optimization can be performed by the Lagrange method. This model and method are an extension of the Bradley-Terry model for paired (one to one) comparisons [12] and the multi-class classification method by pairwised coupling of probability estimates [8] to such to incorporate any possible pairs. We call this new probabilistic approach to optimize the membership probability vector $\boldsymbol{p}$ the maximum *a posteriori* (MAP) method.

3. Estimate class probabilities based on heuristics

To evaluate our proposed method, we use two ways to calculate class probabilities, an existing simple voting heuristic method and a modification of it.

3.1. Single Summation method

In the simplest heuristics called Single Summation (SIS) method [7][13], the class probability is given as

$$p_i = \Big\{ \sum_{[l|m]\in B \text{ s.t. } l=i} q_{[l|m]} + \sum_{[l|m]\in B \text{ s.t. } m=i} q_{[m|l]}) \Big\} / Z, \tag{4}$$

where Z is a normalization term:

$$Z = \sum_{i \in C} p_i.$$

This heuristics sums up probability estimate of binary classification on each target whose subclass has a single class label: $q_{[l|m]}$, where $\#l = 1$ and $\#m = 1$.

3.2. Shared summation method

We here propose another heuristics called Shared Summation (SHS). In SHS, if a target subclass consists of multiple class labels, the probability estimate output is distributed equally to every class label in the subclass. This allows sets of targets such as B_{22} to join in voting. In contrast, such a target cannot be used in SIS. The class membership probability by SHS is given by

$$p_i = \Big\{ \sum_{[l|m]\in B \text{ s.t. } i\in l} \frac{q_{[l|m]}}{\#l} + \sum_{[l|m]\in B \text{ s.t. } i\in m} \frac{q_{[m|l]}}{\#m} \Big\} / Z, \tag{5}$$

where Z is a normalization term similar to the above.

4. Experiments

4.1. Application to synthesized dataset

In order to examine the performance of our MAP method, we first prepared a synthesized dataset; the true distribution was a mixture of four 2-D Gaussians (Fig. 1), consisting of four classes, from which $20 \times 4 = 80$ training data and $40 \times 4 = 160$ test data were generated.

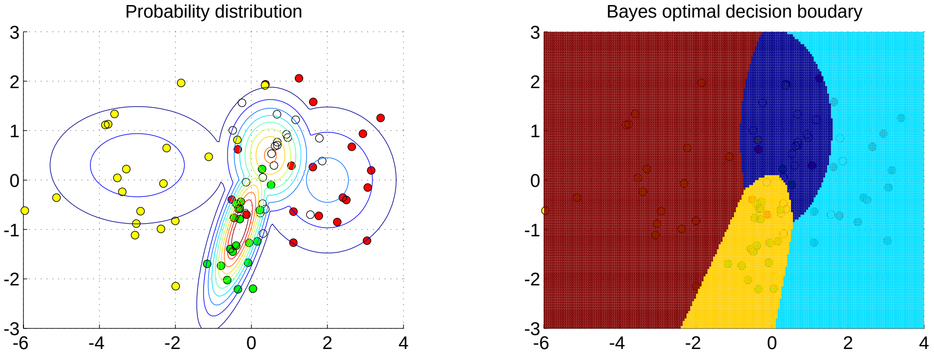

Figure 1: Synthesized dataset. The left panel represents training data and the true probability distribution. The right panel represents the Bayes optimal decision boundaries of the four classes based on the true distribution.

We compared the classification performances of each combination of the three voting procedures (MAP, SHS, and SIS) and four types of targets (11, 1R, 1A and AA). In this experiment, we used an

Table 1: Classification accuracy of the synthesized data.

	MAP	SHS	SIS
1R	0.794	0.787	0.806
11	0.819	0.800	0.825
1A	0.812	0.812	0.812
AA	0.812	0.812	N/A

SVM with a linear kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T\boldsymbol{x}'$ as a binary classifier. Table 1 shows the classification result for the test dataset. In each entry, the value represents the classification accuracy.

The binary results show that our MAP method has comparable or even better performance than SHS and SIS, with an exceptional case for 1R; in 1R, the high accuracies by SHS and SIS are due to the large number of indeterminate samples. In addition, we can see that the performance of the combination of either of MAP and SHS, and AA is fairly good; although the AA combines a lot of classifiers, its performance is better than the conventionally-used 1R and 11, and such a higher-order combination is naturally done by our MAP method.

4.2. Application to a gene expression dataset

As an application to a realistic problem, our method was applied to a tumor classification problem using gene expression profiling data. We used a dataset of gene expressions from four classes (FA, FC, N and PC) of thyroid cancer: 119 training samples consisted of 41 FA, 20 FC, 28 N, and 30 PC, and 49 test samples consisted of 17 FA, 8 FC, 12 N, and 12 PC. Each gene expression was a vector of log-expression ratios of 2,000 genes. We used weighted voting [14][3], a kind of linear discriminator after the gene selection based on the statistical test using signal-to-noise ratio, as a binary classifier, because it has often been used in the field of gene expression analyses. We set the significant level p in the gene selection to 0.001.

Table 2 shows the accuracy by each combination (left part: leave-one-out (LOO) accuracy, right part: test dataset accuracy). From these results, we can see that the accuracies are higher by using a higher-order combination (1A and AA) than those by the conventionally-used combination (1R and 11). If we can use classifiers of higher-order combination, the performance does not depend on the combination way, i.e., MAP and SHS show comparable results. SIS may be inferior to MAP or SHS when the AA combination shows the best performance. As for the reason why the results by 1A and AA are the same in the independent test case, we consider the classification accuracy has been saturated with 1A, and additional classifiers in AA are no more necessary.

Table 2: Classification accuracy of the gene expression dataset.

	LOO accuracy			Test dataset accuracy		
	MAP	SHS	SIS	MAP	SHS	SIS
1R	0.765	0.765	0.765	0.816	0.816	0.816
11	0.782	0.782	0.782	0.816	0.816	0.816
1A	0.798	0.798	0.798	0.857	0.857	0.857
AA	0.807	0.807	N/A	0.857	0.857	N/A

As a consequence, we have found that there are cases in which a higher-order combination like AA shows the best performance, and our MAP method shows comparable classification accuracies with the heuristic voting methods, SHS and SIS. SIS is not good, because it cannot deal with higher-order combination (AA).

Then, we examined the performance of two multi-class classification algorithms which have been used in gene expression analyses to validate our method. One is the shrunken centroid algorithm which is an improvement of the linear discriminant classifier [15] and the other is the multi-class support vector machine (MC-SVM) [16] which extends the SVM objective function to a multi-class one and optimizes it. Note that these methods differ from our method on the point that they construct inherent multi-class discriminant function. We applied these methods to the gene expression dataset and evaluated their classification performances.

In the shrunken centroids algorithm, the best value for the shrinkage parameter Δ was searched over from 0.5 to 6 at intervals of 0.5. In the MC-SVM, a linear kernel and a 2-D polynomial kernel were used. Table 3 shows the prediction performances of these algorithms. The prediction accuracies of these methods were substantially lower than that obtained by our MAP-AA.

Table 3: Performance comparison of multi-class classification algorithms

Classifier	LOO accuracy	Test dataset accuracy
Shrunken centroids ($\Delta = 1$)	0.748	0.796
MC-SVM (linear kernel)	0.765	0.674
MC-SVM (polynomial kernel, d=2)	0.714	0.735
MAP-AA	0.807	0.857

5. Conclusion

In this article, we proposed a probabilistic model of binary classifiers for constructing a multi-class classifier, and its estimation method. We showed that our method achieves comparable performance to the heuristics voting methods and superior performance to two other existing methods the shrunken centroids algorithm and the multi-class SVM. We found that there are cases in which higher-order combination of binary classifiers shows the best performance, and in such cases, our probabilistic approach to constructing a multi-class classifier exhibits a natural model for the vote by the classifiers. Although the eligibility of each classifier is fixed in this study, its tuning can be done in the framework of probabilistic inference, which is our near future work.

References

[1] Saxen, E., Franssila, K., O. Bjarnason, T.N., Ringertz, N.: Observer variation in histologic classification of thyroid cancer. Acta Pathol Microbiol Scand [A] **86A** (1978) 483–486

[2] Fassina, A.S., Montesco, M.C., Ninfo, V., Denti, P., Masarotto, G.: Histological evaluation of thyroid carcinomas: reproducibility of the "WHO" classification. Tumori **79** (1993) 314–320

[3] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286** (1999) 531–537

[4] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley Interscience (2000)

[5] Oba, S., Sato, M., Ishii, S.: Variational Bayes method for mixture of principal component analyzers. In: Proceeding for 7th International Conference on Neural Information Processing (ICONIP2000). Volume 2. (2000) 1416–1421

[6] V. Vapnik: Statistical Learning Theory. Wiley, NY (1998)

[7] B. Schölkopf and C. Burges and V. Vapnik: Extracting support data for a given task. In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining. (1995) 252–257

[8] Hastie, T., Tibshirani, R.: Classification by pairwise coupling. In Jordan, M.I., Kearns, M.J., Solla, S.A., eds.: Advances in Neural Information Processing Systems. Volume 10., The MIT Press (1998)

[9] B. Schölkopf and C. Burges and A. Smola: Advances in Kernel Methods *Support Vector Learning*. The MIT Press (1999)

[10] Li, T., Zhang, C., Ogihara, M.: A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics **20** (2004) 2429–2437

[11] Anderson, J.: Logistic discrimination. Biometrika **59** (1972) 19–35

[12] Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs, I. The method of paired comparisons. Biometrika **41** (1952) 324–345

[13] Tax, D., Duin, R.P.W.: Using two-class classifiers for multi-class classification. In: Proceedings 16th International Conference on Pattern Recognition (ICPR). Volume 2. (2002) 124–127

[14] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., Golub, T.R.: Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A **98** (2001) 15149–15154

[15] R. Tibshirani and T. Hastie and B. Narasimhan and G. Chu: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A **99** (2002) 6567–6572

[16] Weston, J., Watkins, C.: Multi-class support vector machine. Technical report, University of London (1998)