2008－BIO－13 (4)
2008／6／26

非計量多次元尺度構成法を用いたガン細胞判別

今野英[1] 田口善弘[2,3]

中央大学[1] 大学院理工学研究科物理学専攻,[2] 理工学部物理学科,[3] 理工学研究所

マイクロアレイの結果からガン細胞を判別するのは大切な研究目標である。通常、遺伝子数がサンプル数より圧倒的に多いため、判別に用いる遺伝子を少数に限定しないとover fittingが起きてしまうが、遺伝子の最適選択はサンプルが代わると代わってしまう欠点があった。本論文では非計量多次元尺度構成法を用いて遺伝子を選択せずに判別する方法を２つ提案する。１つは遺伝子発現プロファイル間のユークリッド距離を類似度として高次元にサンプルを埋め込んだ配置を用いて判別する場合、もうひとつは低次元への埋め込みを判別を行った後に再度繰り返す方法である。両方とも従来の研究と同じかそれ以上のaccuracyを達成できた。

Cancer discrimination based upon non-metric multidimensional scaling method

Masaru Konno and Y-h. Taguchi

Dept. Phys., Chuo. Univ., tag@granular.com

It is an important task to discrminate cancer using microarray experiments. Usually, since the number of samples is smaller than that of genes, it is unavoidable to select a part of genes for discrimination in order to avoid over fitting, but this choice may differ from samples to samples. In this paper, we propose two new methods for discrimination without selecting genes using non-metric multidimensional scaling (nMDS). One is the discrimination based upon embedding by employing Euclidean distance between gene expression profiles as dissimilarity and another is based upon iterative embedding after every discrimination. Both two achieve competitive performance with the conventional voting method.

1 Introduction

Discrimination of cancer from normal cell without the assist of human eye or *in vitro* test is important task. For example, microarray should have enough information for discrimination between cancer and normal cell, since the difference of gene expression profile should be fundamental difference between these two. The problem on the usage of microarray for discrimination of cancer is that number of samples is smaller than number of genes. This means, over fitting is unavoidable. In order to be free from this problem, usually genes used for discrimination is reduced so as to be smaller than the number of samples. Typically, genes which exhibits larger difference of expression between normal and cancer genes are selected as such representatives. Only difficult of this strategy is that choices can change from samples to samples.

In this study, we propose two methods to discriminate cancer using non-metric multidimensional scaling method (nMDS). One is the discrimination based upon embedding using the correlation coefficients between gene expression profiles as similarity measure. Another is based upon discrimination followed by embedding after every discrimination. Both two turn out to exhibit competitive performance with the conventional voting method.

2 Materials and Methods

We have employed Data set A[1] in Ref.[2]. It contains 64 primary (class 0, FALSE) adenocarcinomas and 12 metastatic (class 1, TRUE) adenocarcinomas (lung, breast, prostate, colon, ovary, and uterus) from unmatched patients prior to any treatment. Clinical stage of primary tumors and outcome unknown. Each sample consists of gene expression profiles of 12603 genes.

In order to discriminate these two classes, we propose two procedures using nMDS [3].

Method 1: We simply embed 76 samples into d dimensional space using nMDS with dissimilarity of Euclidean distance between gene expression profiles. Linear discriminant analysis (LDA)[4] is applied to embeddings to discriminate two classes for various d. Actual procedures are as follows. Suppose $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iM}), i = 1, ..., N$ is the gene expression vector for sample i. x_{ij} is gene expression profile of gene j in sample i. $N(= 76)$ is total number of samples. $i \leq 64$ stands for class 0 and $i \geq 65$ stands for class 1. $M(= 12603)$ is total number of genes. Dissimilarity $\delta_{ii'}$ is Euclidean distance

$$\delta_{ii'} = \sqrt{\sum_{j=1}^{M}(x_{ij} - x_{i'j})^2}.$$

If we would like to classify sample A, Then we get position vector $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{id}), i = 1, ..., N$, LDA has been done for the set of vectors $\mathbf{y}_i$ with leave one out cross validation.

Method 2: As method 1, we embed 76 samples into d dimensional space. Then, we check to which class one specific sample A is classified. When the sample A is classified into class 0(1), we pick up all samples classified into class 0(1). Then these sampled are embedded into d dimensional space again. Since usually these samples contain both classes, we can discriminate the sample A with other samples. Actual procedures are as follows. The procedure is the same as method 1 until we apply LDA to obtained embeddings. If sample A is classified into class 0(1) with leave one out cross validation, we pick up set of samples $Z_{0(1)}$ which are classified into class 0(1). When $Z_{0(1)}$ contains samples with only either of two real classes, this is the predicted class of sample A. If $Z_{0(1)}$ includes both of samples with real class 0 and class 1, we apply method 1 again to the set of samples $Z_{0(1)}$. The resulting class is the predicted class of sample A.

3 Results

In Table 1, we have presented performances by both method 1 and method 2 as a function of embedding dimension d. Those for both methods 1 and 2 are averaged values over 30 trials, since embeddings obtained differ from trials to trials due to initial configuration dependence. Error bars are 95 % conficence intervals. Those by voting are also presented. The definition of performance is as follows.

$$\begin{aligned} Sensitivity &= \frac{TP}{FN+TP} \\ Specificity &= \frac{TN}{FP+TN} \\ Accuracy &= \frac{TP+TN}{FN+TP+FP+TN}, \end{aligned}$$

	method 1			method 2				Voting[2]		
d	sens.	spec.	accu.	sens.	spec.	accu.	N	sens.	spec.	accu.
-	-	-	-	-	-	-	1	33	86	78
2	59±3	53±1	53±2	30±4	77±8	69±2	2	33	59	55
3	50±3	60±2	58±2	51±4	73±7	71±1	-	-	-	-
4	41±3	60±4	57±2	54±4	73±7	70±2	4	33	80	72
5	42±2	65±5	61±2	56±3	74±7	71±1	-	-	-	-
6	42±2	68±5	63±1	58±3	77±7	73±1	-	-	-	-
7	39±2	74±5	64±1	60±3	77±8	74±2	-	-	-	-
-	-	-	-	-	-	-	8	42	83	76
10	50±4	75±5	71±1	56±3	78±9	74±1	-	-	-	-
15	50±3	75±10	78±1	53±6	79±9	77±1	-	-	-	-
-	-	-	-	-	-	-	16	50	80	76
20	57±3	83±9	79±1	51±4	79±9	77±2	-	-	-	-
30	58±3	85±10	79±2	50±4	81±9	77±2	-	-	-	-
-	-	-	-	-	-	-	32	50	81	78
40	46±3	79±9	75±2	51±4	83±9	78±2	-	-	-	-
-	-	-	-	-	-	-	64	50	81	78
-	-	-	-	-	-	-	100	58	83	80
-	-	-	-	-	-	-	**128**	**58**	**83**	**80**
-	-	-	-	-	-	-	150	50	81	78
-	-	-	-	-	-	-	200	50	80	76
-	-	-	-	-	-	-	256	58	78	75
-	-	-	-	-	-	-	512	58	77	74
-	-	-	-	-	-	-	1024	58	75	72
-	-	-	-	-	-	-	2048	58	75	72
-	-	-	-	-	-	-	4096	58	73	71
-	-	-	-	-	-	-	8192	58	75	72
-	-	-	-	-	-	-	8716	58	75	72

Table 1: Comparison of performances (%) among several methods. N is number of features used in Ref.[2]. The performances for voting are taken/computed from Table D in Web Table A[2]. Sens., spec., and accu. stand for sensitivity, specificity, and accuracy respectively. Bold numbers are those employed as the best ones in Ref.[2].

where TP is the number of samples which belong to class 1 and are correctly predicted to be class 1, FN is the number of samples which belong to class 1 but are wrongly predicted to be class 0, TN is the number of samples which belong to class 0 and are correctly predicted to be class 0, and FP is the number of samples which belong to class 0 but are wrongly predicted to be class 1. As can be seen in Table 1, this is difficult discrimnation. By voting[2], specificity and accuracy are at most 58 % and 80 %. In addition to this, performances by voting do not depend upon the number of features used monotonically. For example, accuracy 78 % when only one feature is only 2 % less than the best 80 % when 128 features are considered. This fact enables us to doubt the consistency of their method. Either method 1 or 2 does not violate monotonicity upon embedding dimensions d within error bars. Thus, our method is more trusted then theirs. The most critical problem of voting is that it cannot consider all of genes since over fitting can occur. In contrast to this, method 1 with $d = 20$ achieves the same performance as the best by voting[2]. This means that our method can achieve the competitive performance even if we consider all of genes. If we regard the number of embedding dimension d as being analogous to the number of features N, our method is mostly better than voting with similar d and N. For example, for $d = 2$, both method 1 and 2 has better or competitive performances with those when voting is used with the number of feautures, N=2.

4 Conclusion

In this paper, we have proposed the usage of non-metric multidimensional scaling method for the discrimination of cancer. It turns out that nMDS can achieve competitive or better performance even if we consider all of genes, since we can control the effect of over fitting by changing embedding dimension d.

5 Acknowledge

This work has been partially supported by the Grant-in-Aid for Creative Scientific Research No.19500254 of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) from 2007 to 2008. We are grateful for their support.

References

[1] http://www.broad.mit.edu/mpr/publications/projects/Metastasis/DatasetA_Tum_vs_Met.res

[2] Ramaswamy S,Ross KN,Lander ES,Golub TR.(2003) A molecular signature of metastasis in primary solid tumor .Nat Genet.33:49-54.

[3] Y-h. Taguchi and Y. Oono, Relational patterns of gene expression via non-metric multidimensional scaling analysis, Bioinformatics 21 (2005) 730-740.

[4] lda module in R[5].

[5] R Development Core Team, R: A Language and Environment for Statistical Computing, (2007), ISBN 3-900051-07-0,http://www.R-project.org.