Kernel Methods를 이용한 Human Breast Cancer의 subtype의 분류 및 Feature space에서 Clinical Outcome의 pattern 분석

김혜진^{O} 최승진, 방승양
포항공과대학교 컴퓨터공학과 IM 연구실
{marisan0, seungjin, sybang}@postech.ac.kr

Subtype classification of Human Breast Cancer via Kernel methods and Pattern Analysis of Clinical Outcome over the feature space

Hey-Jin Kim0 Seungjin Choi Sung-Yang Bang
Dept. of C.S.E, Pohang Universit

Abstract

This paper addresses a problem of classifying human breast cancer into its subtypes. A main ingredient in our approach is *kernel machines* such as support vector machine (SVM), kernel principal component analysis (KPCA), and kernel partial least squares (KPLS). In the task of breast cancer classification, we employ both SVM and KPLS and compare their results. In addition to this classfication, we also analyze the patterns of clinical outcomes in the feature space. In order to visualize the clinical outcomes in low-dimensional space, both KPCA and KPLS are used. It turns out that these methods are useful to identify correlations between clinical outcomes and the nonlinearly projected expression profiles in low-dimensional feature space.

1. Introduction

Kernel methods have kernel feature space where data points are able to separate clearly. The more complicated the organization of data, the better the separability of data subtypes of kernel methods than other methods. The most popular technique among kernel methods is support vector machine (SVM)[1], which is good at the subtype classification and multiple regression. In the unsupervised manner, kenel principal component analysis (KPCA)[2] is also very attractive method for data visualization and dimension reduction. We also introduce kernel partial least sqaure (Kernel PLS). We improve its convergence ability and classification accuracy by modifying the algorithm devised by Rosipal et. al.[3] The ability of kernel PLS is compared to SVM for the sepability and to kernel PCA and PLS for the data visualization using one of microarray data.

Microarray is the highly developed biotechnology and provide thousands of gene expression information simultaneously. The data is Sørlie's Breast Tumor Clinical Implication microarray data in the SMD database[4]. The data set have 85 tissue samples and four subtypes, one of which consist of 3 small subtypes. The cellular heterogeneity of breast tumors and the large number of genes potentially involved in controlling cell growth, death and differentiation emphasize the importance of studying multiple genetic alterations in concert. Systematic investigation of gene expression patterns based on kernel methods reveals that the clinical values of the subtypes are correlated with clinical parameters such as the separation of estrogen receptor (ER) - positive/negative tumors and the state of TP53 gene, wild type or mutant type. Moreover, we report that the classifiication of tumors can be used as a prognostic marker with respect to overall survival on the feature space.

2. Materials and Methods

2.1 Materials

We considered a gene expression data set with several clinical outcome information, a breast carcinomas microarray data set, published by Sørlie et al.(2001)[4] . A total of 78 breast carcinomas(71 ductal, five lobular, and two ductal carcinomas in situ obtained from 77 different indivisuals; two independent tumors from one individual diagnosed at different times) and three fibroadenomas were analyzed in this study. Sørlie et al. identified six clusters of gene expression profiles corresponding to basal-like, ERBB2+ (overexpression of ERBB2 oncoprotein), normal breast-like, luminal subtype A, B and C. They selected two group f gene sets from total 8,102 genes. One of them is named "intrinsic gene set" that have significantly greater variation in expression between different tumors than between paired samples from the same tumor. Another is named SAM264 genes gained by SAM software[5].

2.2 Data normalization

The original data set consists of 85 xls files, each of them represent different tumor tissues including various normalization factors. We chose data with flag equal to zero, which means that the data had been obtained by the credible experiment process, with the regression correlation more than 0.6 varing by at least 4-fold from the median red/green ratio. The quality control ScanAlyze parameter of "%pixels > background of at least 0.55 in both the red and green channels. Before being rejoined into a single table, the data table was then split into tissues tables and the subtables were separately median polished.

Finally we selected HJE400 gene set with higher score of Pearson correlation values over 50 percentile among 1632 genes resulted in the previous normalization step.

2.3 Methods

2.3.1 Partial Least Square (PLS)

Partial Least Square (PLS) is the name of a set of algorithm developed by Wold for use in econometrics. The PLS approach has been used in chemometrics for extracting chemical information from complex spectra which contain interference effects from other facotrs (noise) than those of primary interest. The PLS approach in microarray has been applied to binary tumor classification[6] as well as multi-class discrimination[7] for the purpose of dimension reduction.

PLS regression is a technique for modelling a linear relationship between a set of output variables and a set of input variables with the model y=XB. PLS is similar to the well known method of principal component analysis (PCA). In a PCA, maximizing the variance of the linear combination of the predictor variables (genes), namely var(Xv) is not directly related to the response variables. Instead, PLS is to sequentially maximize the covariance between the response variable (y) and a linear combination of the genes (X). Thus, we find the weight vector w satisfying the following objective criterion,

$$\mathbf{w}_k = \underset{\mathbf{w'w=1}}{\operatorname{argmax}} \operatorname{cov}^2(\mathbf{Xw},\mathbf{y})$$

subject to the orthogonality constraint

$$\mathbf{w'}_k \mathbf{Sw}_j = 0$$

for all 1 ≤ j ≤ k where S=X'X. The ιth PLS score vector is a linear combinations of the original predictive variables

$$\mathbf{t}_i = \mathbf{Xw}_i$$

but the weights are non-linear functions of both y and X. These variabes are obtained by NIPALS algorithm[8].

2.3.2. Kernel Partial Least Square (Kernel PLS)

The kernel PLS algorithm was created by Roman *et al.*[9] The model of kernel PLS is

$$\hat{\mathbf{y}} = \mathbf{\Phi B}$$

where ϕ denote an matrix of regressors whose vector is nonlinearly transformed on the feature space. Let kernel K be $\phi\phi$T. Based on NIPALS algorithm, they replaced data X to kernel K. Different from PLS, kernel PLS is a nonlinear regression method depending on the choice of kernel. In a reproducing kernel hilbert space(RKHS), reproducing property helps to overcome the common problem, the poor generalization properties of existing nonlinear regression techniques.

The kernel PLS algorithm by Roman et al. turns out to have poor convergency and the data set was under-fitted regressed. Therefore we modified the previous algorithm to support better iteration convergence. As a result, it was showed that the data visualization as well as the convergency got better.

In addition to the data projection, kernel PLS is very good at the subtype classification. The predictive response variables are obtained by in the case of PLS,

$$\hat{\mathbf{y}} = \mathbf{\Phi B} \qquad \text{where } \mathbf{B} = \mathbf{W}(\mathbf{P'W})^{-1}\mathbf{C}^T$$

instead in kernel PLS,,

$$\hat{\mathbf{y}} = \mathbf{\Phi B} = \mathbf{KU}(\mathbf{T}^T\mathbf{KU})^{-1}\mathbf{T}^T\mathbf{Y}$$

3. Results and Discussion

Recently kernel methods attract people because those generate kernel feature space where data points can be linearly separable. We address three kernel methods such as SVM, KPCA and KPLS. SVM and KPLS comparing to PLS were used for the subtype classification. KPCA and

KPLS as well as PLS were employed for the purpose of the correlation between clinical outcome and the projected data points (tissues). The KPLS algorithm designed by Roman *et al* was converged so poorly in this data set that the classification result could much less than 50%. Therefore we modified both the normalization part and the regression coefficient calculation part. Figure 1 showed the convergence of our improved KPLS algorithm. The subtype classification results listed in Figure 2 for KPLS, Table 1 for SVM and Table 2 and Table 3 for PLS. In the lieu of the classification accuracy for training data, SVM recorded the best algorithm but for test data, KPLS and PLS were better than SVM. Moreover, when we obtained those data from 10-fold randomly sampled data, the accuracy variance of SVM was much worse than others.

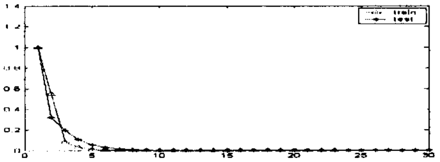
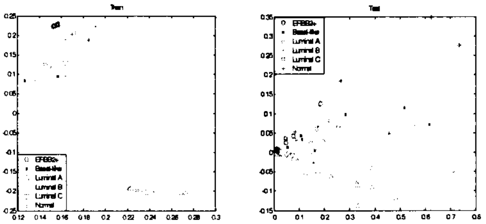


Figure 1. The Kernel convergence



(a) Train (b) Test

Figure 2. The result of 6 subtypes via KPLS

Table 1. SVM classification results

	Intrinsic set	SAM set	HJE set
train	99.54 ±1.04	98.68±2.07	99.11±2.35
test	75.67±12.82	70.75±12.99	72.98±12..86

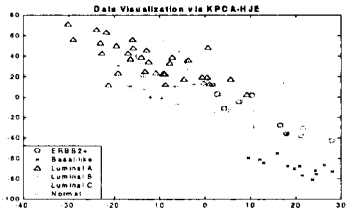


Figure 3. Data projection via KPCA

Table 2. PLS classification for Luminal subtypes

	Intrinsic	SAM	HJE
train	99.00±0.0211	96.00±0.0459	96.5±0.0337
test	81.11±0.0564	84.44±0.0340	86.3±0.0429

Table 3. PLS classification for 4 classes

	Intrinsic	SAM	HJE
train	84.18±0.0264	83.50±0.0396	88.00±0.0396
test	77.85±0.0679	76.00±0.0596	80.00±0.0596

Another interesting result was the correlation between clinical outcomes and the projected data on the feature space. We projected data using KPCA, KPLS and PLS. As seen in the figure 3, the data are projected as their subtypes without prior knowledge. We found that some data projected coincidentally had common clinical outcomes. Particularly, the useful marker for tumor, ER +/- and tp53 are very similar between them.

Finally, we explored that the survival month has close relationship with subtypes. Using PLS and KPLS, we trained data with the response variable, survival month. Then the projected data made groups which are clearly matched to the tumor subtypes.

4. References

[1] Nello Cristianini and John Shawe-Taylor, "*An Introduction to Support Vector Mechines,*" pp93-124 Cambridge university press, 2000

[2] B. Scholkopf and A. Smola, "Nonlinear component analysis as a kernel eigenvalue problem",*Neural Computation*, 1996.

[3] Roman Rosipal and Lenard J. Trejo, "Kernel Partial Least Squares Regression in Reproducion Kernel Hilbert Space," *Journal of Machine Learning Research* vol.2 97-12

[4] Charles M. Perou, Therese SrØlie et al, et al., "Molecular portraits of human breast tumours," *Nature*, vol. 406 747-75

[5] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu,"Significance analysis of microarrays applied to the ionizing radiation response," *Proc.Natl.Acad.Sci.USA* vol. 98 u no. 9 pp5116-5121 April 24, 2001

[6] Nguyen, D.V. and Rocke, D.M. "Classification of acute leukemia based on DNA microarray gene expressions using partial least squares," Lin, S.M and Jonhnson, K.F. edition, *Methods of Microarray Data Analysis*, Kluwer, Dordrecht, pp109-124

[7] Nguyen, D.V. and Rocke, D.M."Multi-class cancer classification via partial least squares with gene expression profiles," *Bioinformatics*, vol 18, pp1216-1226

[8] Paul G. and Bruce R.K., "Partial Least Square regression: a tutorial," *Analytica Chimica Acta* vol 185 (1986) pp 1-17

[9] Roman Rosipal and Leonard J. Trejo, "Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space," *Journal of Machine Learning Research* vol 2 pp97-124 2001