Journal of the Korean Data & Information Science Society
2015, **26**(2), 475–486

http://dx.doi.org/10.7465/jkdi.2015.26.2.475
한국데이터정보과학회지

Cross platform classification of microarrays by rank comparison[†]

Sunho Lee[1]

[1]Division of Mathematics and Statistics, Sejong University

Received 15 December 2014, revised 30 December 2014, accepted 25 January 2015

Abstract

Mining the microarray data accumulated in the public data repositories can save experimental cost and time and provide valuable biomedical information. Big data analysis pooling multiple data sets increases statistical power, improves the reliability of the results, and reduces the specific bias of the individual study. However, integrating several data sets from different studies is needed to deal with many problems. In this study, I limited the focus to the cross platform classification that the platform of a testing sample is different from the platform of a training set, and suggested a simple classification method based on rank. This method is compared with the diagonal linear discriminant analysis, k nearest neighbor method and support vector machine using the cross platform real example data sets of two cancers.

Keywords: Classification, cross platform, k nearest neighbor method, microarray, support vector machine.

1. Introduction

Microarray technique allows measuring large amounts of gene expression levels simultaneously and generates valuable information. Given an efficient statistical analysis, biomolecular information could become even more essential than traditional clinical factors. Hence, high-throughput gene expression has become one of the most important tools in functional genomic studies.

A recommendation that the data supporting some scientific results in the publication must be sharable with the research community made authors of the paper submit data to public repositories: Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo), Stanford Microarray Database (SMD, http://smd.stanford.edu), ArrayExpress (http://www.ebi.ac.uk/arrayexpress), etc. Several efforts to develop standards for microarray experiments and their analysis were made: The Minimum Information About a Microarray Experiment (MIAME) (Brazma *et al.*, 2001; http://www.mged.org/Workgroups/MIAME/miame.html) provided a

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2013-022457).
[1] Professor, Division of Mathematics and Statistics, Sejong University, Seoul 143-747, Korea.
E-mail: leesh@sejong.ac.kr

standard for recording and reporting microarray-based gene expression data, which facilitated the establishment of databases and public repositories. The Micro Array Quality Control (MAQC) (http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/) project has played an important role in improving the microarray and next-generation sequencing technologies and fostering their proper applications. Phase-I (Shi *et al.*, 2006) evaluated the microarray data quality and the reproducibility of results among laboratories and platforms, while Phase-II (Shi *et al.*, 2010) evaluated the performance of microarray-based classifiers for clinical use. Phase-III is ongoing for assessing the technical performance of next-generation sequencing platforms.

Recently, due to the obvious advantages of the economic feasibility from the high cost of arrays and the lack of samples, gene expression analysis utilizing the accumulated data sets from the public repositories has become active. Further, the results coming from several studies have the strength for generalization and reliability. Several issues, however, make dealing data from different studies together complicated. We have to control a possible source of variations such as protocols, environments of experiments, RNA preparation, array platform, data preprocessing, and other technical noises (Liu *et al.*, 2008; Liu *et al.*, 2013; Larsen *et al.*, 2014). It is imperative to represent all data sets by a common set of probes. Entrez Gene (Maglott *et al.*, 2005) is frequently used as an identifier matching common probes between different data sets and duplicated genes are summarized by their median.

In this research, we limited our interest to develop a simple classification procedure to predict a binary phenotype (e.g., normal vs. tumor) when the platform of a testing sample is different from that of a training set. In this cross platform situation, each data set measures the gene expressions using a different technology and a different scale. cDNA arrays and oligonucleotide arrays are two major forms.

This article is organized as follows. In Section 2, we give a brief review of several existing classification methods–diagonal linear discriminant analysis, k nearest neighbor method, and support vector machine–while describing our new classification procedure based on ranks. In Section 3, we explain and propose an algorithm on how to compare prediction accuracy among classifiers within platform and cross platform situations. In Section 4, we apply our approach using two real data sets. Section 5 contains concluding remarks.

2. Classification methods

Let $\boldsymbol{Y} = (y_1,\ y_2,\ \cdots, y_n)$ be a response vector of n samples where $y_i \in \{-1\ ,\ 1\}$ is a binary phenotype of ith sample, $i = 1, 2,\ \cdots, n$. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdot, \boldsymbol{x}_n)^{\boldsymbol{T}}$ and $\boldsymbol{x}_i \in R^g$ be expression values of g differentially expressed genes selected among c common genes of ith sample which are standardized in each gene. Suppose that the augmented matrix $[\boldsymbol{X} : \boldsymbol{Y}^T]$ is a training set. When a new gene expression features $\boldsymbol{x}^* \in R^g$, possibly from a different platform of a training set, is given, we want to assign a phenotype of $\boldsymbol{x}^*$.

2.1. Existing classification methods

The diagonal linear discriminant analysis (DLDA)

Suppose that the class conditional gene expression profile $\boldsymbol{X}_{|m}$, a subset of $\boldsymbol{X}$ satisfying a class phenotype $y = m$ $(m = -1, 1)$, is assumed to have a multivariate normal distribu-

tion $\boldsymbol{X}_{|m} \sim$ MVN $(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$. The maximum likelihood discriminant analysis function for a new sample $\boldsymbol{x}^*$ is $\delta_m(\boldsymbol{x}^*) = (\boldsymbol{x}^* - \boldsymbol{\mu}_m)\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}^* - \boldsymbol{\mu}_m)^T + log|\boldsymbol{\Sigma}_m|$ and its phenotype can be assigned to $\text{argmin}_l \delta_l(\boldsymbol{x}^*)$. In this research, DLDA assuming a same diagonal covariance matrix $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_g^2)$ is adopted. In spite of somewhat unrealistic assumption of independence among genes, this method has been found to work very well in microarray data analysis (Dudoit, Fridlyand and Speed, 2002; Diaz-Uriarte and Alvarez de Andres, 2006).

The phenotype of $\boldsymbol{x}^* = (x_1, x_2, \cdots, x_g)$ is assigned to $\widehat{y} = argmin_m \sum_{j=1}^{g}(x_j - \bar{x}_{j|m})^2/\widehat{\sigma}_j^2$ where $\widehat{\boldsymbol{\mu}}_m = \bar{\boldsymbol{X}}_{|m} = (\bar{x}_{1|m}, \bar{x}_{2|m}, \cdots, \bar{x}_{g|m})$ and $\widehat{\boldsymbol{\Sigma}} = diag(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \cdots, \widehat{\sigma}_g^2)$, the sample mean vectors and covariance matrices in the training set, respectively.

k nearest neighbor method (KNN)

KNN (Fix and Hodges, 1951) is a simple nonparametric classification rule. The classifier measures the distance between test sample $\boldsymbol{x}^*$ and each sample $\boldsymbol{x}_i \in R^g$ $(i = 1, 2, \cdots, n)$ in the training set $[\boldsymbol{X} : \boldsymbol{Y}^T]$. A phenotype of $\boldsymbol{x}^*$ is predicted by most common among its k nearest neighbors.

In the algorithm in KNN, there are several choices of distance metrics, and number of neighbors (k) can be tuned by cross validation to retain low error rate. Usually, the Euclidean distance, Manhattan distance, Minkowski distance, or one minus the correlation is used as a distance metric. Instead of tuning parameters, we choose the Euclidean distance, a typical metric for our continuous variables, and k=3 odd value to prevent ties for our classifier.

Support vector machine (SVM)

Define a hyperplane $f(x) = \beta_0 + \boldsymbol{x}\boldsymbol{\beta}^T$ where β_0 is an intercept and $\boldsymbol{\beta}$ is a g dimensional weight vector and this linear SVM can be extended to nonlinear machine by using $f(x) = \beta_0 + \boldsymbol{\Phi}(\boldsymbol{x})\boldsymbol{\beta}^T$, where $\boldsymbol{\Phi}(\boldsymbol{x})$ is a transform of $\boldsymbol{x}$. Assign one of two possible phenotypes to a new sample $\boldsymbol{x}^*$ through the classification rule $\widehat{y} = sign(f(\boldsymbol{x}^*))$.

($\boldsymbol{\beta}_0$,$\boldsymbol{\beta}$) is found by solving

$$min_{(\beta_0,\boldsymbol{\beta})} \frac{\|\boldsymbol{\beta}\|^2}{2} + C\sum_{i=1}^{n}\xi_i \text{ subject to } y_i(\beta_0 + \boldsymbol{\Phi}(\boldsymbol{x}_i)\boldsymbol{\beta}^T) \geq 1 - \xi_i,\ i = 1, 2, \cdots, n$$

where $\xi_i > 0$ is a slack variable and C is a tuning parameter which controls the tradeoff between loss and penalty (Cortes and Vapnik, 1995).

2.2. Newly suggested classification method based on a rank (RANK)

Previously mentioned standard classification methods are used when the distribution of observation values in a testing set is the same as that in a training set. However, in the cross platform situation, distributions of gene expressions from different platforms are different. For example, two colorectal carcinoma data sets from GEO on cDNA (GSE20970, 43 tumor samples, and 30 normal samples) and Affymetrix (GSE23878, 35 tumor samples, and 24 normal samples) platforms were downloaded. Figure 2.1 shows the histograms of the mean expression values of genes in each platform.

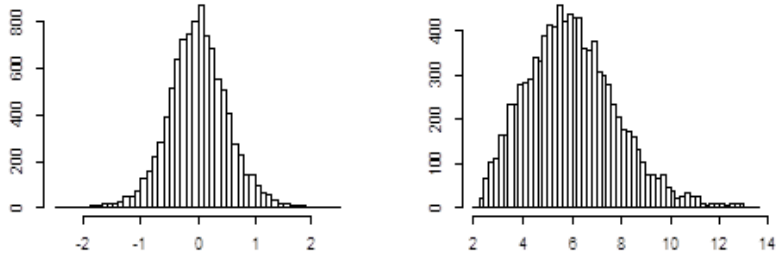

Figure 2.1 Histogram of mean expression value of each gene in different platforms (left: cDNA microarray, right: Oligonucleotide array)

For cross platform data analysis, comparability across platforms is an important issue. At the early stage of microarray experiment, Kuo *et al.* (2002) showed that matched mRNA measurements from two different platforms had poor correlation coefficients even using the same samples. As the microarray technology improved, this kind of problem was alleviated significantly, although it still remained.

There were several trials to transform data sets with different numerical formats to derive comparable measures to each other. Warnat *et al.* (2005) suggested a median rank scores method. For integrating data sets from different platforms, one of the data sets was declared as a reference set, a median expression value over all samples in the reference set was found for each gene, and the median values were arranged from lowest to highest. For each sample in the non-reference set, every gene expression value was replaced into a rank-corresponding median expression value of the reference set. Quantile discretization method was based on equal frequency binning (Liu *et al.*, 2002). For each sample, all of the sorted expression values were discretized into eight equal sized bins and every value in the same bin was changed to the same integer value from {-3, -2, -1, 0, 0, 1, 2, 3} depending on their quantiles. These transformations of expression values made all the samples follow the same distribution with a heavy information loss. Nilsson *et al.* (2006) and Chen *et al.* (2008) showed that z-score standardization method (mean centering and unit variance) was good in removing variation resulting from different platforms, but not enough to make the same distributions of gene expression values in a cross platform analysis.

There is no doubt that the distribution of gene expression values from each platform follows its own way depending on the technique for measuring gene expression. But there must be consistency of ranks between platforms based on the rationale that small numbers of differentially expressed genes in one platform also show extreme expression values far from the average expression values in another platform. Therefore, to assign a phenotype of a new sample which comes from a different platform of a training set, their ranks rather than the expression values were preferably compared.

From the training set, select g differentially expressed genes and find their ranks in each phenotype from lowest to highest among c common genes. For a new sample, rank genes and calculate the Manhattan distances based on ranks to each of the phenotype centroids in the training set using g predetermined differentially expressed genes. A phenotype of a new sample is also assigned to the nearest one.

3. Prediction performance of the classifiers

Selection of biomarkers and classification of samples are major tasks in gene expression studies, and many classifiers are usually compared for their predictive performance with the optimum number of genes. Since Parry *et al.* (2010) showed that their three feature ranking methods−significance analysis of microarrays (SAM) d-value, fold change ranking with P-value threshold of 0.05, and P-value ranking with fold change threshold of 1.5−performed similarly well in the research of KNN modeling strategy, and our main interest is to find a classifier which predicts well in the cross platform classification, only one method, SAM, is applied to rank the differentiability of genes.

Using the top ranked genes in the training set, we can build a classifier and examine how accurately the classifier predicts the phenotypes of testing samples. For assessing the performance appropriately, an independent validation, where the information of the test set should never be used in developing a classifier, has to be followed (Lee, 2008).

Each constructed classifier predicts a binary phenotype and its performance of prediction is measured by the correspondence between the true phenotype and the predicted one as displayed in Table 3.1.

Table 3.1 2×2 table for true and predicted phenotype

		true phenotype	
		positive	negative
predicted phenotype	positive	true positive (TP)	false positive (FP)
	negative	false negative (FN)	true negative (TN)

There are several metrics to measure the correspondence including accuracy, sensitivity, specificity, Matthews correlation coefficient, receiver operating characteristic curve, root mean squared error, etc. Among them, following Matthews correlation coefficient (MCC) is adopted which is used to report performance in MAQC-II (Shi *et al.*, 2010).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC is easy to calculate and is informative even when the proportion of binary phenotypes are not balanced. It is a correlation coefficient between the observed and predicted binary classifications, with +1 indicating perfect prediction, and 0 and -1 for random prediction and perfect inverse prediction, respectively.

In this research comparing several classifiers, MCC is calculated in two situations for each classifier with top-ranked g significant genes (g=2, 4,$\cdots$, 100): within platform and cross platform. Within platform performance measures the MCC of a classifier when the platform of a training set and that of a test set are the same, and the cross platform performance measures the MCC of a classifier when they are different.

Suppose there are two data sets, A and B, with different platforms. For set A with n samples, we can use $n-1$ samples to build a classifier and test its prediction accuracy through one remaining sample. We also have to separately calculate the MCC of the classifier testing all samples in Set B. These procedures for every sample in the set A are repeated to get a within platform performance as a result of leave one out cross validation (LOOCV) of set A

and a cross platform performance as an average of n times repeated MCCs for testing set B by following the algorithm in Table 3.2.

Table 3.2 Algorithm for calculating MCC values

Input Cross platform data sets A and B with common genes

Training set $A = [X_A : Y_A] = \cup_1^n A_i$ for n samples ($A_i = [\boldsymbol{x}_i : y_i] : i$th sample data)

Test set $B = [X_B : Y_B]$: for b samples

where X is a gene expression matrix of samples in a set, standardized in each gene
Y is a response vector of samples in a set

Return MCC of a classifier for within platform and cross platform classification

Begin
Set $TP_g^W = FP_g^W = TN_g^W = FN_g^W = 0$ for all g
$TP_{ig}^C = FP_{ig}^C = TN_{ig}^C = FN_{ig}^C = 0$ for all i, g
For i=1 to n do
Split A into $A_{-i} = A - A_i$ and A_i
Rank the differentially expressed genes in A_{-i}
For g=2 to 100 by 2 do
Build a classifier C_{ig} using top-ranked g differentially expressed genes in A_{-i}
Predict a phenotype of ith sample using C_{ig}
If prediction is truly positive, **then** $TP_g^W = TP_g^W + 1$
Elseif prediction is falsely positive, **then** $FP_g^W = FP_g^W + 1$
Elseif prediction is truly negative, **then** $TN_g^W = TN_g^W + 1$
Else $FN_g^W = FN_g^W + 1$
EndIf
For m=1 to do
Predict a phenotype of mth sample in B using C_{ig}
If prediction is truly positive, **then** $TP_{ig}^C = TP_{ig}^C + 1$
Elseif prediction is falsely positive, **then** $FP_{ig}^C = FP_{ig}^C + 1$
Elseif prediction is truly negative, **then** $TN_{ig}^C = TN_{ig}^C + 1$
Else $FN_{ig}^C = FN_{ig}^C + 1$
EndIf
EndFor

$$MCC_{ig}^C = \frac{TP_{ig}^C \times TN_{ig}^C - FP_{ig}^C \times FN_{ig}^C}{\sqrt{(TP_{ig}^C + FP_{ig}^C)(TP_{ig}^C + FN_{ig}^C)(TN_{ig}^C + FP_{ig}^C)(TN_{ig}^C + FN_{ig}^C)}}$$

EndFor
EndFor
For g=2 to 100 by 2 do
Within platform MCC of a classifier using top ranked genes

$$MCC_g^W = \frac{TP_g^W \times TN_g^W - FP_g^W \times FN_g^W}{\sqrt{(TP_g^W + FP_g^W)(TP_g^W + FN_g^W)(TN_g^W + FP_g^W)(TN_g^W + FN_g^W)}}$$

Cross platform MCC of a classifier using top ranked genes

$$MCC_g^C = \frac{\sum_i^n MCC_{ig}^C}{n}$$

EndFor

4. Real example data analysis

In order to check the performance of newly suggested classifier based on rank, RANK, we compare the MCC of RANK with three previously mentioned classifiers, DLDA, KNN, and SVM in the situation of within platform and cross platform classification

For conducting the cross platform data analysis, we searched publicly available data sets via the GEO Web site. After downloading two data sets with the same experimental purpose but different platforms from cDNA and oligonucleotide arrays, two data sets were linked using the Entrez Gene ID and common genes were selected. Afterwards, each sample was normalized separately and the gene expression values in each sample were standardized.

In this research, we downloaded the following cross platform data sets of two cancers: the diffuse large B-cell lymphoma and the non-small cell lung cancer.

Diffuse large B-cell lymphoma (DLBCL)

DLBCL, the most common type of non-Hodgkin's lymphoma among adults, showed a clinical heterogeneity. For a certain kind of the therapy, less than half of patients responded well and exhibited prolonged survival, whereas the others failed in the treatment. Using cDNA microarrays, Alizadeh *et al.* (2001) have conducted a systematic characterization of gene expression in B-cell malignancies and identified two molecularly distinct forms, germinal centre B-like DLBCL (GCB) and activated B-like DLBCL (ABC). Patients with GCB had a significantly better overall survival than those with ABC. For a cross platform analysis with Alizadeh *et al.* (2000), the oligonucleotide array data set of Williams *et al.* (2010), which proposed a new method of amplification of formalin-fixed paraffin-embedded tissue was adopted. Having all the microarrays of two data sets annotated the same way, we found an intersection of 2,556 genes.

Table 4.1 Details of DLBCL microarray studies

Name	Study reference	Platform	Samples	Probes	GEO
[O]	Williams *et al.* (2010)	Oligo	27 GCB, 21 ABC	54675	GSE19246
[C]	Alizadeh *et al.* (2000)	cDNA	14 GCB, 13 ABC	18432	GSE60

Figures 4.1 and 4.2 represent the MCCs of the four classifiers using DLBCL data under two different circumstances. Figure 4.1 is MCC plots obtained by LOOCV in the within platform classification over the number of differentially expressed genes. Even though classifiers had developed under the same purpose of distinguishing samples between GCB and ABC, they had different MCCs and different types of superiority depending on the data [O] and [C]. Plots generally showed a pattern that the MCC using [C] (data submitted in 2002) was lower and had larger performance variance than that of [O] (data submitted in 2009). This could have resulted partly from the fact that the quality control of the early stage microarray experiment was poor.

Figure 4.2 represents the means and standard deviations of MCCs as a result of repeating the number of sample size of a training set. In this cross platform classification, RANK shows an outstanding performance with the highest MCC and the smallest variance. On the contrary, SVM shows inferior performance with the largest variance.

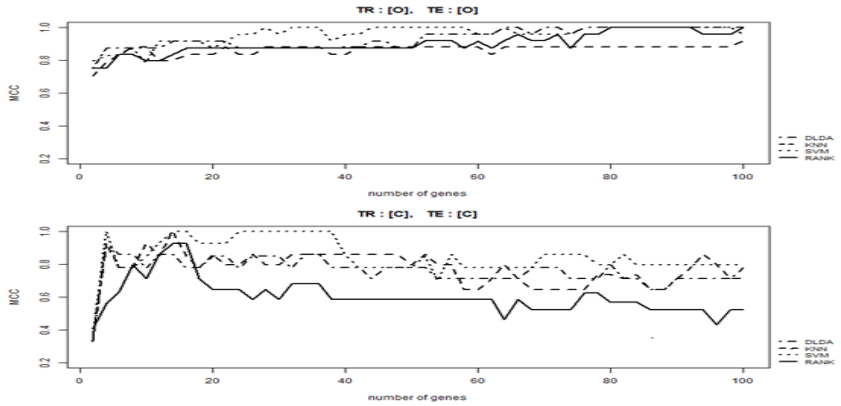

Figure 4.1 Plots of MCCs of the four classifiers DLDA, KNN, SVM, and RANK over the number of significant genes in DLBCL in within platform situation
(a training set [TR] and a test set [TE] have the same platform).

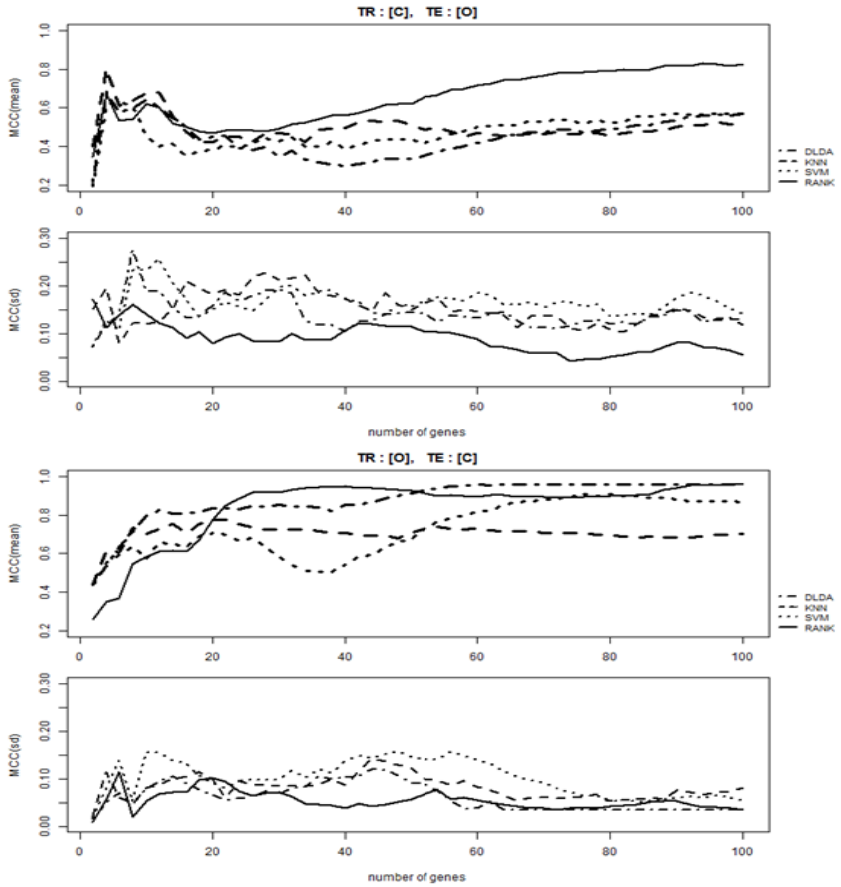

Figure 4.2 Plots of means and standard deviations (sd) of MCCs of the four classifiers DLDA, KNN, SVM, and RANK over the number of significant genes in DLBCL in cross platform situation
(a training set [TR] and a test set [TE] do not have the same platform).

Non-small cell lung cancer (NSCLC)

NSCLC can be classified into major subtypes adeno carcinoma (AC) and squamous cell carcinoma (SCC). Even though NSCLC is one of the most common causes of cancer death in Western communities and its explicit molecular and clinical characteristics have been reported, a specific therapy does not exist for it yet.

Using the global gene expression profiling of 58 human NSCLC specimens in oligonucleotide array, Kuner *et al.* (2009) showed large transcriptomic differences between AC and SCC which may help to understand the disease and find targets for therapies. For the same reason, Newnham *et al.* (2011) also identified critical genes involved in NSCLC pathogenesis using samples in 10,500 element cDNA microarray. We identified 3,672 common genes.

Table 4.2 Details of NSCLC microarray studies

Name	Study reference	Platform	Samples	Probes	GEO
[O]	Kuner *et al.* (2009)	Oligo	40 AC, 18 SCC	54675	GSE10245
[C]	Newnham *et al.* (2011)	cDNA	33 AC, 25 SCC	10500	GSE25326

Plots of Figure 4.3 show the MCCs of four classifiers obtained by LOOCV in the within platform classification using [O] and [C] of NSCLC, respectively. The performance shows different patterns depending on which data was analyzed. With [O], SVM shows a good performance compared to KNN with a difference of about 0.1. However, within platform classification result of [C], RANK showed an unexpected best performance and SVM got worse as the number of genes increased with more than 0.3 difference in MCC. It might be due to the quality of data sets or the biological influence depending on tumor types. In Figure 4.4, the means and standard deviations of MCCs of four classifiers are compared in the cross platform classification. RANK is rather a safe choice with the highest or next to highest MCC and a small variance. On the contrary, SVM shows inferior performance with the largest variance. For predicting the phenotype of [C], SVM is noticeably worse in the cross platform classification compared to the within platform classification.

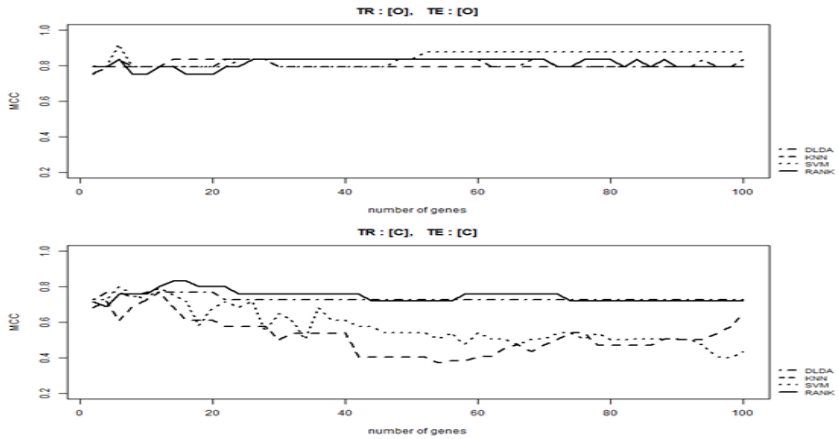

Figure 4.3 Plots of MCCs of the four classifiers, DLDA, KNN, SVM, and RANK over the number of significant genes in NSCLC in within platform situation
(a training set [TR] and a test set [TE] have the same platform).

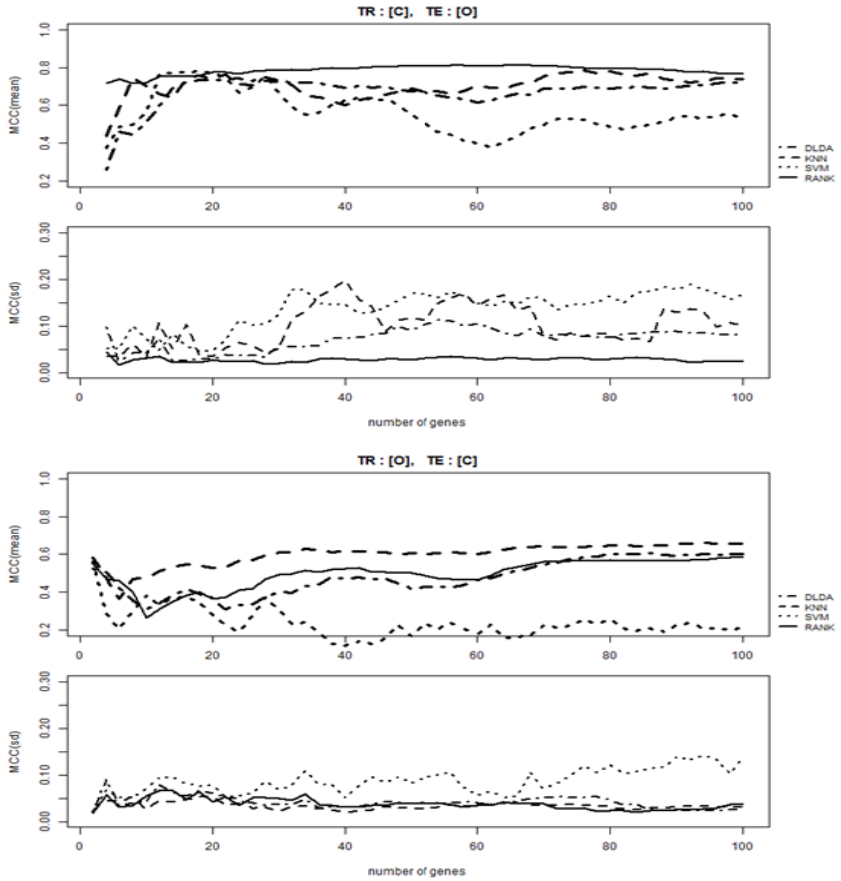

Figure 4.4 Plots of means and standard deviations (sd) of MCCs of the four classifiers DLDA, KNN, SVM, and RANK over the number of significant genes in NSCLC in cross platform situation (a training set [TR] and a test set [TE] do not have same platform).

5. Discussion

For testing samples, usually cross platform classification performance is worse than within platform performance and this tendency is greatly severe in SVM. Cross platform classification with a good quality of the training set sometimes shows a better performance, however, than within platform classification.

It is hard to conclude which method is superior to others in all conditions although RANK shows outstanding performance with a small variance in cross platform situation and it is easy to apply and interpret the result. Even though the good performance of RANK in cross platform classification is only validated through two examples, not by simulation, there is no need to argue the performance of RANK as a result of our previous pilot studies.

Applying KNN and SVM, we can tune some parameters for better performance by k-fold cross validation on the training data set. When we tune the parameters in real example data analysis (results are not shown), there is no effect in cross platform classification. Tuning parameters for KNN in within platform classification only shows some improvements in

MCC; however, considering the computation time, it is still recommendable to use default parameters.

For improving the performance of RANK, we may give weights in calculating Manhattan distance depending on the number of positively standardized genes among g differentially expressed genes, but its performance is not consistent.

How to decide the optimum number of differentially expressed genes in a classifier has to be studied further.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick. J.C., Sabet, H. *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W. *et al.* (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, **29**, 365-371.

Chen, Q. R., Song, Y. K., Wei, J. S., Bilke, S., Asgharzadeh, S., Seeger, R. and Khan, J. (2008). An integrated cross-platform prognosis study on neuroblastoma patients. *Genomics*, **92**, 195-203.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**, 273-297.

Diaz-Uriarte R. and Alvarez de Andres S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Dudoit, S., Fridlyand, J. and Speed, TP. (2002). Comparison of discriminant methods for the classification of tumors using gene expression data. *Journal of American Statistical Association*, **97**, 77-87.

Fix, E. and Hodges, J. L. (1951). *Discriminatory analysis, nonparametric discrimination: Consistency properties*,Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas.

Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L. and Kohane, I. S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405-412.

Kuner, R. Muley, T. Meister, M. Ruschhaupt, M. Buness, A. Xu, E., Schnabel, P., Warth, A. *et al.* (2009). Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer*, **63**, 32-38.

Larsen, M., Thomassen, M., Tan, Q., Sørensen, K. and Kruse, T. (2014). Microarray-based RNA profiling of breast cancer: Batch effect removal improves cross-platform consistency. *BioMed Research International*, Article ID 651751.

Lee, S. (2008). Mistakes in validating the accuracy of a prediction classifier in high-dimensional but small-sample microarray data. *Statistical Methods in Medical Research*, **17**, 635-642.

Liu, H., Hussain F., Tan C.L. and Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, **6**, 393-423.

Liu, H., Chen, C., Liu, Y., Chu, C., Liang, D., Shih, L. and Lin, C. (2008). Cross-generation and cross-laboratory predictions of Affymetrix microarrays by rank-based methods. *Journal of Biomedical Informatics*, **41**, 570-579.

Liu, H., Peng, P. C., Hsieh, T. C., Yeh, T., Lin, C., Chen, C. Hou, J., Shih, L. *et al.* (2014). Comparison of feature selection methods for cross laboratory microarray analysis. *BMC Bioinformatics*, **15**, 274.

Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, **33**, D54-58.

Newnham, G., Conron, M., McLachlan, S., Dobrovic, A., Do, H., Li, J., Opeskin, K., Thompson, N. *et al.* (2011). Integrated mutation, copy number and expression profiling in resectable non-small cell lung cancer. *BMC Cancer*, **7**, 11-93.

Nilsson, B., Andersson, A., Johansson, M. and Fioretos, T. (2006). Cross-platform classification in microarray-based leukemia diagnostics. *Haematologica*, **91**, 821-824.

Parry, R. M., Jones, W., Stokes, T. H., Phan, J. H., Moffitt, R. A., Fang, H., Shi, L., Oberthuer, A. *et al.* (2010). k-nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *Pharmacogenomics Journal*, **10**, 292-309

Shi L., Campbell, G., Jones, W. D., Campagne, F., Wen, Z., Walker, S. J., Su, Z., Chu, T. *et al.* (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, **28**, 827-838

Shi, L., Reid, L., Jones, W., Shippy, R., Warrington, Baker, S., Collins, P., Francoise de Longueville. *et al.* (2006). The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, **24**, 1151-1161.

Warnat, P., Eils, R. and Brors, B. (2005). Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, **6**, 265.

Williams, PM. Li, R., Johnson, NA., Wright, G., Heath, JD. and Gascoyne, RD. (2010). A novel method of amplification of FFPET-derived RNA enables accurate disease classification with microarrays. *Journal of Molecular Diagnosis*, **5**, 680-686.