Comprehensive Comparison of Classifiers for Metabolic Profiling Analysis

Yu Cao[1], Xirui Cui[1], Tianlu Chen[2*], Mingming Su[4]
[1]School of Life Science and Biotechnology,
Shanghai Jiao Tong University, Shanghai 200240, China;
[2] Shanghai Center for Systems Biomedicine, Key Laboratory
of Systems Biomedicine (Ministry of Education)
Shanghai Jiao Tong University, Shanghai 200240, China;

Aihua Zhao[3], Xiaoyan Wang[2], Yan Ni[4], Wei Jia[24*]
[3]School of Pharmacy,
Shanghai Jiao Tong University, Shanghai 200240, China;
[4]Department of Nutrition,
University of North Carolina, North Carolina 28081, USA

***Abstract*—Metabonomics is an emerging field providing insight into physiological processes and difference. Besides conventional PCA, PLS and OPLS approaches, more and more machine learning classifiers are likely to become the supplements for metabolic profiling data analysis. A comprehensive comparison of PLS, support vector machine (SVM, with linear and quadratic kernels), linear discriminant analysis (LDA), and random forest (RF) was reported applying on clinical metabonomics data. The accuracy of these classifiers was tested by 7-fold and holdout Cross Validation. Their stability and over fitting were evaluated by holdout Cross Validation and permutation (repeated 100 times). Their prediction ability was investigated by ROC curve, and their sensitivity on irrelevant variables was studied by variable ranking combining selection step by step. The overall performance of RF and SVM (linear kernel) is superior to the others. Some selected variables are of significance for further research on metabolic difference.**

Keywords- classification; metabolic profiling; random forest; support vector machine

I. INTRODUCTION *(HEADING 1)*

Metabonomics can be defined as a field of science that deals with the measurement of metabolites in an organism for the study of the physiological processes in a particular situation [1]. Through the analysis of one or several kinds of bio-fluids, serum, urine, tissue, etc., the changes in metabolites of living bodies can be tracked and compared. Thus, the reaction of metabolic network to various stimuli such as infection, disease, or drug use, would reflect in the fluctuation of physiological processes. Various high throughput analytical instruments combined with a wealth of bioinformatics methods make more and more regulation detectable and explicable. Therefore, parallel to genomics and proteomics, metabonomics has been increasingly used as a versatile tool in many areas such as diagnosing or prognosing clinical diseases [2], monitoring the chemical-induced toxicity in organs [3], exploring the potential mechanism of diverse diseases, assessing therapeutic effects of drugs [4] and so on.

So far, the most widespread and effective bioinformatics (also known as chemometrics) methods conducted in metabolic profiling processing mainly include Student T test, ANOVA(analysis of variance) test, principal component analysis (PCA), partial least squares analysis (PLS), and orthogonal partial least squares analysis (OPLS). These unitvariate and multivariate analysis methods share the objectives: classification of phynotype groups (e.g. normal volunteers or clinical patients, patients with benign or malign tumor, statuses of model rats in several time points) and selection of potential biomarkers while the former one gets more attention. In 2008, we tried using PCA, PLS, and OPLS to separate normal volunteers with patients with colorectal cancer as well as patients in different stages after GCTOF/MS (gas chromatography time-of-flight mass spectrometry) detection [2]. In 2009, we applied PCA and PLS to analyze metabolic profiling data of UPLC-QTOF/MS (ultra performance liquid chromatography-quadrupole time of flight mass spectrometry) instrument derived from rats' urine and discussed on the melamine-induced acute renal toxicity [3]. In 2010, we processed profiling data of GCTOF/MS instrument derived from serum and urine of patients with liver cancer by PLS. In these researches, the classical methods do not always have satisfactory performance when coping with phynotype groups of subtle difference or clinical profiling with large individual differences. To improve the effectiveness of group differentiation, many machine learning methods have been introduced into the field of metabonomics data analysis and some of them are noteworthy. Support vector machines (SVM) was used on binned NMR spectral data and was improved to be better than PLS [1]. Linear discriminant analysis (LDA) combined with PCA was used to analyze the data of the urine samples of renal cell carcinoma [5]. Besides these two classifiers, RF (random forest, expanded from decision tree) was also reported as an excellent classifier boasting following advantages: simple theory, capability of multiple group separation, automatic compensation mechanism on biased sample numbers of groups, and so on [6]. Although SVM, LDA and RF are likely to become the supplements of classical methods, no comprehensive and persuasive study of their performance on metabonomics data analysis has been reported. How to select the proper classifier according to specific samples and demands is still a question that needs to be answered.

In this paper, we compared and evaluated the typical classifier (PLS) and several potential ones (SVM, LDA, and RF) from various aspects attempting to provide some

*: Corresponding author

instruction and reference on the selection of metabolic profiling analysis method. Section two introduced the methodology and related parameters of the classifiers and evaluation approaches briefly. Section three showed the comparison results of five classifiers on "real world" metabolic profiling data sets. A k-fold cross-validation was used to provide a "nearly unbiased" estimate of the classification accuracy. A 15% (10%) holdout cross validation (applied 100 times) was used to test the accuracy and stability of the classifiers. Permutation was carried out to investigate the over fitting of the classifiers. An implementation of the ROC Analysis was realized to represent all the possible and optimal predictive performance of the classifiers. The influence of variable number (variable ranking and selection) was finally studied by stepwise reduction of the irrelevant ones. Section four is the conclusion.

II. METHODOLOGY

A. Methodology of Classifier

1) Partial Least Squares Analysis

The PLS is a projection method seeking the quantitative relationship between two data tables. Spectral or chromatographic data are usually comprised by a set of calibration samples and another set of concentrations of endogenous metabolites, which could be respectively represented by the response Y and the predictor X. The PLS classifier might be understood as simultaneously fitting two principal components 'like' classifiers, one in the X-space and one in the Y-space [7]. The blocks are decomposed as follows:

$$X = TP^T + E \quad (1)$$

$$Y = UC^T + F \quad (2)$$

Here, T and U are the score matrices and P and C are the loading matrices for X and Y, respectively, E and F are the residual matrices [8]. P is useful for interpreting which spectral variables are influential for the modeled responses, and which are not. In a similar way, the T is used to reveal similarities and differences among the measured responses.

The influence (contributions) on Y of every variables in the PLS classifier is called VIP (Variable importance) which is the sum of all the extracted components. The Sum of squares of all VIPs is equal to the number of terms in the classifier hence the average VIP would be equal to 1. One can compare the VIP of one term to the others. Terms with large VIP, larger than 1 especially, are the important ones that are relevant for explaining Y.

PLS is a typical representative of all the classical classifiers. It is an extension of PCA by including the supervising Y. OPLS is a revision of PLS by removing variation in X that is orthogonal to Y. Naturally, PLS is a good comparison reference for new classifiers.

2) Linear Discriminant Analysis

Linear discriminant analysis (LDA) projects the data onto a lower-dimensional vector space such that the ratio of between-class variance to the within-class variance is maximized, thus achieving maximum discrimination [9]. In general, if each class is tightly grouped, but well separated from the other class, the quality of the classifier is considered to be high. The discriminant function is acquired by regression and the response of the function is a linear combination of important variables. The coefficient of each variable indicates the importance of this variable thus can be used for variable ranking.

3) Support Vector Machines

Given a training set of instance-label pairs (x_i, y_i) , $i = 1,...,l$ where $x_i \in R^n$ and $y \in \{1,-1\}^l$, the support vector machines require the solution of the following optimization problem: $\min_{w,b,\xi} \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi_i$, Subject to $y_i(w^T \Phi(x_i) + b \geq 1 - \xi_i$, $\xi_i \geq 0$. Here training vector x_i are mapped into a higher (maybe infinite) dimensional space by the function Φ . The SVM finds a linear separating hyper plane with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term. Furthermore, $K(x_i, x_j) \equiv \Phi(x_i)^T \Phi(x_j)$ is called the kernel function. Though new kernels are being proposed by researchers, the following two basic kernels are most widely used.

Linear: $K(x_i, x_j) = x_i^T x_j$ (3)

Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$ (4)

We choose these two kernels (γ=1, r=1, d=2) for comparison. The importance of samples can be evaluated by the SVM RBF (Radial Basis Function) algorithm [10].

4) Random Forest

A random forest is a classifier consisting of a collection of tree structured classifiers and each tree casts a unit vote for the most popular class at input. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. The training uses a random selection of variables to split each node. Internal estimates monitor error, strength, and correlation are used to show the response to increasing the number of variables used in the splitting. If an individual variable in the out-of-bag (OOB) cases is randomly permuted before being put back into the tree, then the decrease in the estimated margins is an indication of how important that variable is [6].

B. Methodology of evaluation

1) K-fold Cross-Validation

For estimating the final accuracy of a classifier, we would like an estimation method with low bias and low variance but not the absolute accuracies [11]. In a k-fold cross-validation, we first divide the training set into k subsets (the folds) of approximately equal size. Sequentially one subset is tested using the classifier trained on the remaining k-1 subsets. Thus, each instance of the dataset is predicted once and averaging the test error over the k trials gives an estimate of the expected generalization error. If the induction algorithm is stable for a given dataset, the variance of the cross-validation estimates should be approximately the same, independent of the number of folds [11]. The cross-validation procedure is conducive to

evaluating the true accuracy of the classifiers [12]. Usually, K is set as 7.

2) Holdout cross validation

This method is similar to the K-fold cross-validation except for the repeated (100 times in this research) random selection. The holdout method, sometimes called test sample estimation, partitions the data randomly into two mutually exclusive subsets called a training set and a test set, or holdout set. The training set is given to the inducer, and the induced classifier is tested on the holdout set. After training, the classifier is evaluated on the holdout set and the errors give an estimate of the generalization error. The more instances we leave for the test set, the higher the bias of our estimate is. However, fewer test set instances means that the confidence interval for the accuracy will be wider. Besides accuracy, this method is able to test the stability of a classifier. 15% and 10% holdout set was used for the comparison in this research.

3) Permutation

The Y observations were randomly permuted, while the X-matrix was kept intact. It was a good way to test whether the data was over fit by checking whether it gives a comparable accuracy rate to that of the raw training data. If the training did not over fit the dataset, the error rates would be about 0.5.

4) Receiver Operating Characteristic

Receiver Operating Characteristic (ROC) analysis is a classic methodology from signal detection theory and is now common in medical diagnosis [13]. ROC of a classifier shows its performance as a tradeoff between specificity and sensitivity. Sensitivity is the proportion of patients with disease whose tests are positive (*TruePositive/(TruePositive+FalseNegative)*) and specificity is the proportion of patients without disease whose tests are negative (*TrueNegative/(TrueNegative+FalePoisitive)*). Typically, 1-specificity is plotted on the X axis and sensitivity is plotted on the Y axis. All the predictive behavior of a classifier can be represented by the points in the ROC curve independent of class distributions or error costs [13]. The area under the ROC curve (AUC) is a summary statistic of diagnostic performance and can be used to rank variable importance for class separation.

5) Variable selection

As is known to all, too many irrelevant variables are reliable to result in over fit decision while difference between groups could not be extracted and depicted completely if crucial variables are not concerned [12]. Variable ranking and selection is therefore an appropriate and necessary factor for classifier evaluation. To avoid the possibly negative influence of irrelative variables to useful ones, it is wise to rank and select important variables step by step. Carry out following steps iteratively until optimum variable set and classifier is achieved. The final top variables are of great potential to be biomarkers for further investigation.

a) Classify data into groups by using variable set (all variables are taken as the initial variable set) and record the performance of the classifier.

b) Rank the variables used in the classifier

TABLE I. SAMPLE INFORMATION

	Samples	
	HCC patients	***Healthy volunteers***
N	118	109
Age (Mean, range, yr)	55, 29–76	55, 42–65
Male/Female	55/27	39/32

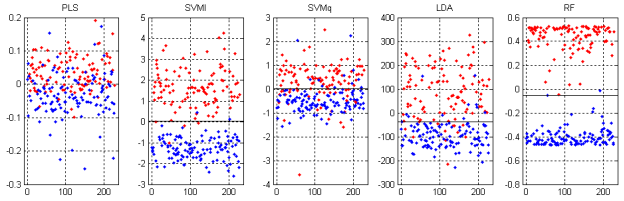

Figure 1. Classification plots of PLS, SVMl, SVMq, LDA, and RF classifiers based on the HCC dataset derived from GCTOF/MS spectral. Red dots represent the healthy volunteers while blue dots are HCC patients. X axis is the sample index and Y axis is the average of 7-fold cross-validation scores.

c) Update the variable set by only top x% variables (reducing the "irrelevant" variables).

X is taken as 100, 80, 56, 28, and 14 in this study.

III. RESULT AND DISCUSSION

1) Metabolic profiling data

Hepatocelluar carcinoma (HCC) is the fifth most common cancer and the third leading cause of cancer-related death with a five-year survival rate of less than 7%. A global metabolic profiling approach was used to identify metabolites differentially expressed in patients with HCC (n=118) compared to healthy volunteers (n=109). Clinical characteristics of all the 227 samples are detailed in Table I. These samples were analyzed by GCTOF/MS instrument and a total of 324 variables (areas of peaks denoting concentrations of metabolites) were obtained from the profiling spectral. Scaling (mean centered and followed by unit variance) was operated hoping to reduce the noise in the data, and to avoid variables in greater numeric ranges dominating those in smaller numeric ranges. Finally, a HCC dataset containing 227 samples and 324 variables was ready for further analysis.

2) Classification

Classifiers (PLS, LDA, SVM, and RF) were performed on the HCC dataset by using the same platform—MATLAB. Fig. 1 plots the classification results of these classifiers. Red and blue dots represent the healthy volunteers and HCC patients respectively. X axis is the sample index and Y axis is the corresponding "score" of the classifiers which was mentioned in the METHODOLOGY part. The scores were the average of the 7-fold cross-validation results. Comparatively, RF classifier gives the best separation (the accuracy is 96.9%) and the performance of SVMl (linear SVM) is the second with 96.8% accuracy. The LDA and SVMq (quadratic SVM) are similar in the classification ability with over 10% wrongly classified samples (accuracy slightly lower than 90%). Unfortunately, the performance of classical PLS is the worst with the accuracy of only 72.7%.

TABLE II. ERROR RATES OF HOLDOUT CV AND PERMUTATION

			PLS	SVM l	SVM q	LDA	RF
holdout CV	***15%***	***Mean***	0.140	0.017	0.177	0.128	0.031 0.000
		Sd.	0.049	0.021	0.053	0.062	
	10%	***Mean***	0.149	0.012	0.176	0.120	
		Sd.	0.075	0.023	0.080	0.065	
permutation	***15%***	***Mean***	0.512	0.529	0.510	0.444	0.506 0.010
		Sd.	0.070	0.077	0.081	0.080	
	10%	***Mean***	0.519	0.528	0.527	0.463	
		Sd.	0.098	0.113	0.104	0.103	

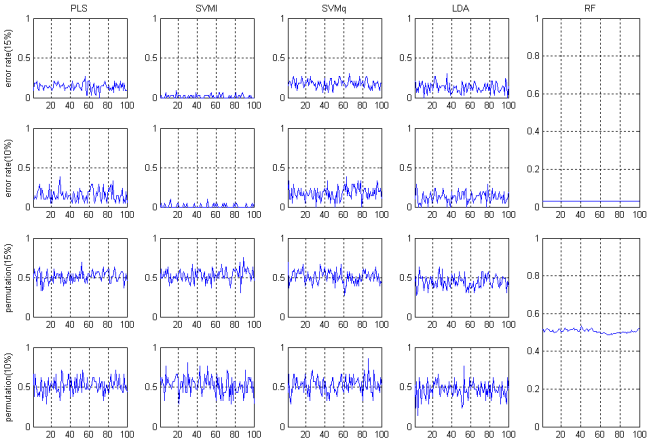

Figure 2. Error rate curves of the holdout CV and permutation (both are repeated 100 times). The upper part is the testing error rate curves for the 15% and 10% holdout cross-validation (30% holdout cross-validation for RF). The lower part shows the permutation error rate curves.

More reliable quantitative comparison is given by the error rates of the holdout cross-validation. The testing error rate curves of the 15% (or 10%) holdout cross-validation for each classifier (100 times) are presented on the upper two rows of Fig. 2 respectively. Each column is for one classifier.

The average (means) and standard deviation (Sds) of the accuracy of the 100 tests are listed in Table II (upper part). The estimation of error rate is related to the number of test samples as mentioned above. Because of the special process of the RF modeling (fixed holdout percent), we gave the back part of the learning curves of the RF classifiers for the comparison, that is to say, 30% of the whole dataset were used as the holdout test samples. Although the ratio of test set was higher (two or three fold) than other classifiers, the error rate of RF was still the second lowest (0.031). The classifier with lowest error rate was SVMl, either for the 15% holdout (0.017) or for the 10% holdout (0.012). This time, the PLS, LDA and SVMq showed similar performance with almost 15% samples being misjudged. On the other hand, since the holdout and training set was selected randomly and the whole process was repeated 100 times, the stability of the classifiers can be reflected as well. In general, the smaller the variation of the error rate curve (Sd), the more stable the classifier is. Therefore, the SVMl and the RF classifiers are more stable than the SVMq, PLS and LDA in this case.

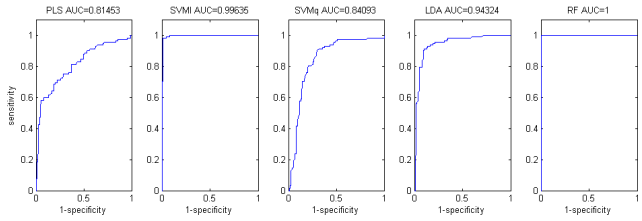

Figure 3. ROC curve of the five classifiers. The area under the curve (AUC) is smaller or equal to 1.

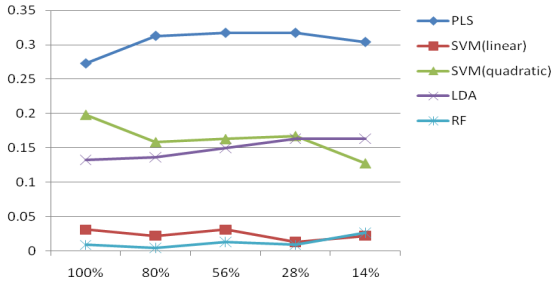

Figure 4. error rate for relevant variables

Over fitting is a common but undesirable problem in metabolic data processing especially the classification field. A classifier is over fitted if it is more accurate in fitting known data (training set) but less accurate in predicting new data (test set). Permutation is a typical method to measure over fitting.

The lower two rows of Fig. 2 is the 100 times permutation error rate curves of the 15% and 10% holdout cross validation respectively. The error rates of almost all the classifiers fluctuated around 0.5 except that the LDA was slightly lower than 0.5. This indicated that none of the classifies except LDA over fitted the data. For PLS, when the top two but not only the first components are used for classification, its error rates were lower than 0.5 dramatically. That is to say, the widely used PLS classifier (two components) was over fitting (figure not shown). Moreover, the fluctuation range of the 10% holdout permutation was generally larger than that of the 15%. The means and Sds of the error rates of the 100 tests are listed in Table II as well (lower part). As supposed, the error rate mean and Sd. of RF were the ones closest to 0.5 and 0 respectively. It was certainly the best one in terms of the permutation test.

The ROC curve is a popular way to estimate the prediction ability of classifiers. The area under the curve (AUC) is smaller than or equal to 1. The closer it is to 1 the more adaptable the classifier is, that is, the higher prediction ability the classifier gets. The ROC curves of all the classifiers are displayed in Fig. 3. As supposed, the RF performed perfectly (AUC=1) and the SVMl separated almost all the samples correctly (AUC>0.99). The result of LDA was satisfying possessing an AUC higher than 0.9. The SVM (q) and the PLS had relatively poor performance and PLS was the worst one with only 0.81 AUC here. This is consistent to the above results.

Besides group separation, potential biomarker selection (differential variable ranking) is a key objective of metabolic profiling analysis as well. Moreover, it is a convincing

criterion evaluating the performance of classifiers. Fig. 4 shows the classification error rate against the number of variables used for building the classifier. Agreeing with former result again, the error rates of RF and SVMl were always the lowest while those of the PLS were the highest. It is worthwhile to note that no significant decrease or increase was found in the classification ability of RF or SVMl, along with the of irrelevant variables. The ranks of important variables did not vary much with the variables included as well. The top four significant metabolites differentiating HCC patients from healthy volunteers are both bile acids, glycocholic acid, glycochenodeoxycholic acid and taurocholic acid. Consequently, these two classifiers seem to be more suitable to handle data sets directly no matter how many irrelevant variables are included. For the LDA (purple) and the SVMq (green) classifiers, the increasing or decreasing trends of their error rate curves are consistently indicating that variable selection might be helpful to the SVMq while unfavorable to the LDA. If the number of variables of a data set is not greatly more than that of samples, variable selection might not be necessary for LDA. If the data set is obviously nonlinear or the correlation of variables is not negligible, SVMq combined variable selection is highly recommended. Since the ambiguous trend of PLS error rate curve (blue), careful and appropriate variable selection should be considered for this classifier.

IV. CONCLUSION

In this study we compared five classification methods on a metabonomic dataset (GC-TOF-MS, clinical urine sample) from multiple aspects and the comprehensive comparison (ranking) of these classifiers are detailed in Table III. For this dataset, the overall performance of RF and SVMl are the best with good accuracy, stability, and prediction ability, almost no over fitting, and little sensitivity with the variable selection. The two classifiers are highly recommended for data set with uncertain characteristics. Since the performance of classical PLS with the first component is not optimistic and the one with two components is over fitted, special care should be taken in case it is selected to analyze metabonomic data. We have conducted the same comparison on two more metabonomic data sets (data of UPLCQTOF/MS derived from rat urine, data of GCTOF/MS derived from colorectal cancer patient serum) and achieved similar results. Although the best choices for these datasets are consistently the RF and SVMl, more metabonomic datasets and evaluation methods are required for further validation.

TABLE III. COMPREHENSIVE COMPARISON OF THESE CLASSIFIERS

Interested performance	Evaluation method	PLS	SVM l	SVM q	LDA	RF
Accuracy	7-fold CV and holdout CV	5	2	4	3	1
Stability	holdout CV and permutation	3	2	4	5	1
Over fit or not	permutation	N	N	N	Y	N
Prediction ability	ROC	5	2	4	3	1
Sensitivity of variable selection	variable selection	Y	N	Y(+)	Y(-)	N

+: variable selection might be helpful to the classifier

-: variable selection is not recommended to the classifier

ACKNOWLEDGMENT

This work was financially supported by the National Basic Research Program of China (2007CB914700), the National Science and Technology Major Project (2009ZX10005-020) and the Natural Science Foundation of Shanghai, China (10ZR1414800).

REFERENCES

[1] S. Mahadevan, S. L. Shah, T. J. Marrie, and C. M. Slupsky, "Analysis of metabolomic data using support vector machines", Anal. Chem. 80, 2008, pp. 7562-7570.

[2] Y. P. Qiu, G. X. Cai, M. M. Su, T. L. Chen, X. J. Zheng, and Y. Xu et al., "Serum Metabolite Profiling of Human Colorectal Cancer Using GC-TOFMS and UPLC-QTOFMS", J. Proteome Res. 8, 2009, pp. 4844-4850.

[3] G. X. Xie, X. J. Zheng, X. Qi, Y. Cao, Y. Chi, and M. M. Su et al., "Metabonomic Evaluation of Melamine-Induced Acute Renal Toxicity in Rats", J. Proteome Res. 9, 2010, pp. 125-133.

[4] Y. Q. Bao, T. Zhao, X. Y. Wang, Y. P. Qiu, M. M. Su, W. P. Jia, and W. Jia, "Metabonomic Variations in the Drug-Treated Type 2 Diabetes Mellitus Patients and Healthy Volunteers", J. Proteome Res. 8, 2009, pp. 1623-1630.

[5] K. Kim, P. Aronov, S. O. Zakharkin, D. Anderson, B. Perroud, I. M. Thompson, and R. H. Weiss, "Urine Metabolomics Analysis for Kidney Cancer Detection and Biomarker Discovery", Mol. Cell. Proteomics 8, 2009, pp. 558-570.

[6] D. Amaratunga, J. Cabrera , and Y. S. Lee, 'Enriched random forests', In Bioinformatics (2008), pp. 2010-2014.

[7] L. Eriksson, J. Trygg, E. Johansson, R. Bro, and S. Wold, "Orthogonal signal correction, wavelet analysis, and multivariate calibration of complicated process fluorescence data", Anal. Chim. Acta 420, 2000, pp. 181-195.

[8] J. Trygg, E. Holmes, and T. Lundstedt, "Chemometrics in metabonomics", J. Proteome Res. 6, 2007, pp. 469-479.

[9] K. Ueki, T. Hayashida, and T. Kobayashi, "Two-Dimensional Heteroscedastic Linear Discriminant Analysis for age-group classification", 18th ICPR, Vol 2, Proceedings 2006, pp. 585-588.

[10] Guyon, J. Weston, S. Barnhill , and V. Vapnik, "Gene selection for cancer classification using support vector machines", Mach. Learn. 46, 2002, pp. 389-422.

[11] J. H. Kim, "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap", Comput. Stat. Data. An. 53, 2009, pp. 3735-3745.

[12] K. Duan, S. S. Keerthi , and A. N. Poo, "Evaluation of simple performance measures for tuning SVM hyperparameters", Neurocomputing 51, 2003, pp. 41-59.

[13] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms", Pattern Recognition 30, 1997, pp. 1145-1159.

[14] X. B. Li, and D. O'Shaughnessy, "Clustering-based Two-Dimensional Linear Discriminant Analysis for Speech Recognition", Interspeech 2007: 8Th ICSLP-INTER SPEECH, Vols 1-4 2007, pp. 1949-1952.

[15] J. D. Katz, G. Mamyrova, O. Guzhva, and L. Furmark, "Random Forests Classification Analysis for the Assessment of Diagnostic Skill", American Journal of Medical Quality 25, 2010, pp. 149-153.