New Bandwidth Selection Criterion for Kernel PCA: Approach to Dimensionality Reduction and Classification Problems

Thomas Minta*[1], De Brabanter Kris[1]and De Moor Bart[1]

[1]ESAT-SCD / iMinds Future Health Department Katholieke Universiteit Leuven,3001, Leuven,Belgium.

Email: minta.thomas@esat.kuleuven.be;

*Corresponding author

Abstract

Background: DNA microarrays are potentially powerful technology for improving diagnostic classification, treatment selection, and prognostic assessment. The use of this technology to predict cancer outcome has a history of almost a decade. Disease class predictors can be designed for known disease cases and provide diagnostic confirmation or clarify abnormal cases. The main input to these class predictors are high dimensional data with many variables and few observations. Dimensionality reduction of these features set significantly speeds up the prediction task. Feature selection and feature transformation methods are well known preprocessing steps in the field of bioinformatics. Several prediction tools are available based on these techniques.

Results: Studies show that a well tuned Kernel PCA (KPCA) is an efficient preprocessing step for dimensionality reduction, but the available bandwidth selection method for KPCA was computationally expensive. In this paper, we propose a new data-driven bandwidth selection criterion for KPCA, which is related to least squares cross-validation for kernel density estimation. We propose a new prediction model with a well tuned KPCA and Least Squares Support Vector Machine (LS-SVM). We estimate the accuracy of the newly proposed model based on 9 case studies. Then, we compare its performances (in terms of test set Area Under the ROC Curve (AUC) and computational time) with other well known techniques such as whole data set + LS-SVM, PCA + LS-SVM, t-test + LS-SVM, Prediction Analysis of Microarrays (PAM) and Least Absolute Shrinkage and Selection Operator (Lasso). Finally, we assess the performance of the proposed strategy with an existing KPCA parameter tuning algorithm by means of two additional case studies.

Conclusion: We propose, evaluate, and compare several mathematical/statistical techniques, which apply feature transformation/selection for subsequent classification, and consider its application in medical diagnostics. Both feature

selection and feature transformation perform well on classification tasks. Due to the dynamic selection property of feature selection, it is hard to define significant features for the classifier, which predicts classes of future samples. Moreover, the proposed strategy enjoys a distinctive advantage with its relatively lesser time complexity.

Introduction

Biomarker discovery and prognosis prediction are essential for improved personalized cancer treatment. Microarray technology is a significant tool for gene expression analysis and cancer diagnosis. Typically, microarray data sets are used for class discovery [1,2] and prediction [3,4]. The high dimensionality of the input feature space in comparison with the relatively small number of subjects is a widespread concern; hence some form of dimensionality reduction is often applied. Feature selection and feature transformation are two commonly used dimensionality reduction techniques. The key difference between feature selection and feature transformation is that, in the former only a subset of original features is selected while the latter is based on generation of new features.

In this genomic era, several classification and dimensionality reduction methods are available for analyzing and classifying microarray data. Prediction Analysis of Microarray (PAM) [5] is a statistical technique for class prediction from gene expression data using Nearest Shrunken Centroids (NSC). PAM identifies subsets of genes that best characterize each class. LS-SVM is a promising method for classification, because of its solid mathematical foundations which convey several salient properties that other methods hardly provide. A commonly used technique for feature selection, t-test, assumes that the feature values from two different classes follow normal distributions. Several studies, especially microarray analysis, have used t-test and LS-SVM together to improve the prediction performance by selecting key features [6, 7]. The Least Absolute Shrinkage and Selection Operator (Lasso) [8] is often used for gene selection and parameter estimation in high-dimensional microarray data [9]. The Lasso shrinks some of the coefficients to zero, and extend of shrinkage is determined by the tuning parameter, often obtained from cross validation.

Inductive learning systems were successfully applied in a number of medical domains, e.g. in localization of primary tumors, prognostic of recurring breast cancer, diagnosis of thyroid diseases, and rheumatology [10]. An induction algorithm is used to learn a classifier, which maps the space of feature values into the set of class values. This classifier is later used to classify new instances, with the unknown classifications (class labels). Researchers and practitioners realize that the effective use of these inductive learning systems requires data preprocessing, before a learning algorithm could be applied [11]. Due to the instability of feature selection techniques, it might be difficult or even impossible to remove irrelevant and/or redundant features from a data set. Feature transformation

techniques, such as KPCA, discover a new feature space having fewer dimensions through a functional mapping, while keeping as much information, as possible in the data set.

KPCA, which is a generalization of PCA, a nonlinear dimensionality reduction technique that has proven to be a powerful pre-processing step for classification algorithms. It has been studied intensively in the last several years in the field of machine learning and has claimed success in many applications [12]. An algorithm for classification using KPCA was developed by Liu *et al.* [13]. KPCA was proposed by Schölkopf and Smola [14], by mapping features sets to a high-dimensional feature space (possibly infinite) and applying Mercer's theorem. Suykens *et al.* [15, 16] proposed a simple and straightforward primal-dual support vector machine formulation to the PCA problem.

To perform KPCA, the user first transforms the input data x from the original input space F_0 into a higher-dimensional feature space F_1 with a nonlinear transform $x \rightarrow \Phi(x)$ where Φ is a nonlinear function. Then a kernel matrix K is formed using the inner products of new feature vectors. Finally, a PCA is performed on the centralized K, which is an estimate of the covariance matrix of the new feature vectors in F_1. One of the commonly used kernel function is radial basis function (RBF) kernel: $K(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2h^2})$ (RBF kernel with bandwidth h). Traditionally the optimal parameters (bandwidth and number of principal components) of RBF kernel function are selected in a trial and error fashion.

Pochet *et al.* [17] proposed an optimization algorithm for KPCA with RBF kernel followed by Fisher Discriminant Analysis (FDA) to find the parameters of KPCA. In this case, the parameter selection is coupled with the corresponding classifier. This means that the performance of the final procedure depends on the chosen classifier. Such a procedure could produce possible inaccurate results in the case of weak classifiers. In addition, this appears to be a time consuming procedure, while tuning the parameters of KPCA.

Most classification methods have inherent problem with high dimensionality of microarray data and hence require dimensionality reduction. The ultimate goal of our work is to design a powerful preprocessing step, decoupled from the classification method, for large dimensional data sets. In this paper, initially we explain an LS-SVM approach to KPCA. Next, by following the idea of least squares cross-validation in kernel density estimation, we propose a new data-driven bandwidth selection criterion for KPCA. The tuned LS-SVM formulation to KPCA is applied to several data sets and serves as a dimensionality reduction technique for a final classification task. In addition, we compared the proposed strategy with an existing optimization algorithm for KPCA, as well as with other preprocessing steps. Finally, for the sake of comparison, we applied LS-SVM on whole data sets, PCA+LS-SVM, t-test + LS-SVM, PAM and Lasso. Randomization on all data sets are carried out in order to get a more reliable idea of the expected performance.

Data sets

In our analysis, we collected 11 publicly available binary class data sets (diseased vs. normal). The data sets are: colon cancer data [18, 19], breast cancer data [20], pancreatic cancer premalignant data [21, 22], cervical cancer data [23], acute myeloid leukemia data [24], ovarian cancer data [21], head & neck squamous cell carcinoma data [25], early-early stage duchenne muscular dystrophy(EDMD) data [26], HIV encephalitis data [27], high grade glioma data [28], and breast cancer data [29]. In breast cancer data [29] and high grade glioma data, all data samples have already been assigned to a training set or test set. The breast cancer data in [29] contains missing values; those values have been imputed based on the nearest neighbor method.

An overview of the characteristics of all the data sets can be found in Table 1. In all the cases, 2/3rd of the data samples of each class are assigned randomly to the training and the rest to the test set. These randomizations are the same for all numerical experiments on all data sets. This split was performed stratified to ensure that the relative proportion of outcomes sampled in both training and test set was similar to the original proportion in the full data set. In all these cases, the data were standardized to zero mean and unit variance.

Methods

The methods used to set up the case studies can be subdivided into two categories: dimensionality reduction using the proposed criterion and subsequent classification.

LS-SVM approach to KPCA

The PCA analysis problem is interpreted as a one-class modeling problem with a target value equal to zero around which the variance is maximized. This results into a sum of squared error cost function with regularization. The score variables are taken as additional error variables. We now follow the usual SVM methodology of mapping the d-dimensional data from the input space to a high-dimensional feature space $\phi : \mathbb{R}^d \to \mathbb{R}^{n_h}$, where n_h can be infinite, and apply Mercer's theorem [30].

Our objective is the following

$$\max_{v} \sum_{k=1}^{N} [0 - v^T(\phi(x_k) - \hat{\mu}_\phi))]^2 \tag{1}$$

with $\hat{\mu}_\phi = (1/N) \sum_{k=1}^{N} \phi(x_k)$ and v is the eigenvector in the primal space with maximum variance. This formulation states that one considers the difference between $v^T(\phi(x_k) - \hat{\mu}_\phi)$ (the projected data points to the target space) and the value 0 as error variables. The projected variables correspond to what is called *score* variables. These error variables are maximized for the given N data points. Next, by adding a regularization term we also want to keep

the norm of v small. The following optimization problem is formulated now in the primal weight space

$$\max_{v,e} J_P(v,e) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} v^T v \tag{2}$$

such that

$$e_k = v^T(\phi(x_k) - \mu_\phi), k = 1, ..., N.$$

The Lagrangian yields

$$\mathcal{L}(w, e; \alpha) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} v^T v - \sum_{k=1}^{N} \alpha_k (e_k - v^T(\phi(x_k) - \hat{\mu}_\phi))$$

with conditions for optimality

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial v} = 0 \rightarrow v = \displaystyle\sum_{k=1}^{N} \alpha_k (\phi(x_k) - \hat{\mu}_\phi) \\ \dfrac{\partial \mathcal{L}}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k \quad k = 1, \ldots, N \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \rightarrow e_k - v^T(\phi(x_k) - \hat{\mu}_\phi) = 0, \quad k = 1, \ldots, N. \end{cases}$$

By elimination of variables e and w, one obtains

$$\frac{1}{\gamma} \alpha_k - \sum_{l=1}^{N} \alpha_l (\phi(x_l) - \hat{\mu}_\phi)^T (\phi(x_k) - \hat{\mu}_\phi) = 0 \quad k = 1, \ldots, N.$$

Defining $\lambda = \frac{1}{\gamma}$, one obtains the following dual problem

$$\Omega_c \alpha = \lambda \alpha$$

where Ω_c denotes the centered kernel matrix with ijth entry: $\Omega_{c,i,j} = K(x_i, x_j)$ - $\frac{1}{N} \sum_{r=1}^{N} K(x_i, x_r)$ - $\frac{1}{N} \sum_{r=1}^{N} K(x_j, x_r) + \frac{1}{N^2} \sum_{r=1}^{N} \sum_{s=1}^{N} K(x_r, x_s)$.

Data-Driven Bandwidth Selection for KPCA

Model selection is a prominent issue in all learning tasks, especially in KPCA. Since KPCA is an unsupervised technique, formulating a data-driven bandwidth selection criterion is not trivial. Until now, no such data-driven criterion was available to tune the bandwidth (h) and number of components (k) for KPCA. Typically these parameters are selected by trial and error. Analogue to least squares cross validation [31, 32] in kernel density estimation, we propose a new data driven selection criterion for KPCA. Let

$$z_n(x) = \Sigma_{i=1}^{N} \alpha_i^{(n)} K(x_i, x)$$

where $K(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2h^2})$ (RBF kernel with bandwidth h) and set the target equal to 0 and denote by $z_n(x)$ the score variable of sample x on n^{th} eigenvector $\alpha^{(n)}$. Here, the score variables are expressed in terms of kernel expressions in which every training point contributes. These expansions are typically dense (nonsparse). In Equation 2, the KPCA uses L_2 lose function. Here we have chosen the L_1 loss function to induce sparsness in KPCA. By extending the formulation in Equation 2 to L_1 loss function, the following problem can be formulated for kernel PCA.

$$\max_{v,e} J_P(v, e) = \gamma \frac{1}{2} \sum_{k=1}^{N} L_1(e_k) - \frac{1}{2} v^T v$$

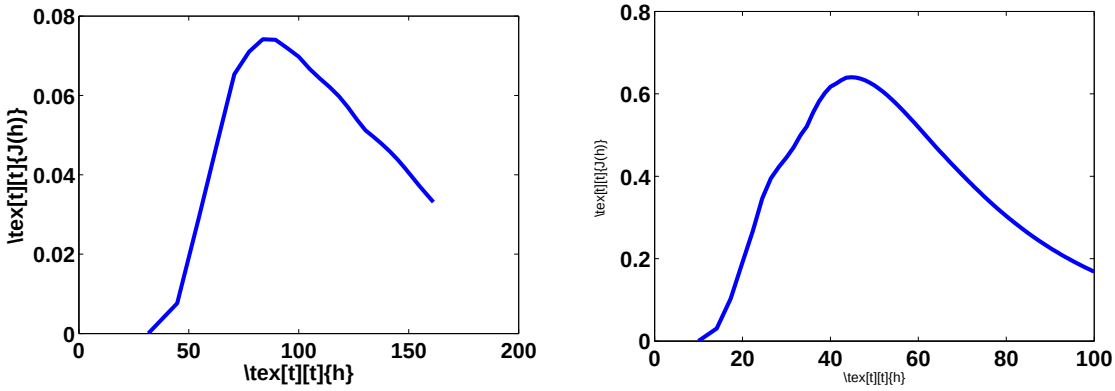
such that

$$e_k = v^T(\phi(x_k) - \mu_\phi), k = 1, ..., N.$$

We propose the following tuning criterion for the bandwidth h which maximizes the L_1 loss function of KPCA:

$$J(h) = \underset{h \in \mathbb{R}_0^+}{\text{argmax}} \, \mathrm{E} \int |z_n(x)| dx, \tag{3}$$

where E denotes the expectation operator. Maximizing Equation (3) would lead to overfitting since we used all



(a) (b)

Figure 1: Bandwidth selection of KPCA for a fixed number of components. Retaining (a) 5 components for cervical cancer data set (b) 15 components for colon cancer data set.

the training data in the criterion. Instead, we work with Leave-One-Out cross validation (LOOCV) estimation of $z_n(x)$ to obtain the optimum bandwidth h of KPCA, which gives projected variables with maximal variance. A finite approximation to Equation (3) is given by

$$J(h) = \underset{h \in \mathbb{R}_0^+}{\text{argmax}} \frac{1}{N} \sum_{j=1}^{N} \int |z_n^{(-j)}(x)| dx \tag{4}$$

where N is the number of samples and $z_n^{(-j)}$ denotes the score variable with the jth observation is left out. In case the leave-one-out approach is computationally expensive, one could replace it with a leave v group out strategy (v-fold cross-validation). Integration can be performed by means of any numerical technique. In our case, we have used trapezoidal rule. The final model with optimum bandwidth is constructed as follows:

$$\Omega_{c,\hat{h}_{max}}\alpha = \lambda\alpha,$$

where $\hat{h}_{max} = \max_{h\in\mathbb{R}_0^+} \frac{1}{N}\sum_{j=1}^{N}\int |z_n^{(-j)}(x)|dx$. Figure 1 shows the bandwidth selection for cervical and colon cancer data sets for fixed number of components. To also retain the optimum number of components of KPCA, we modify Equation (4) as follows:

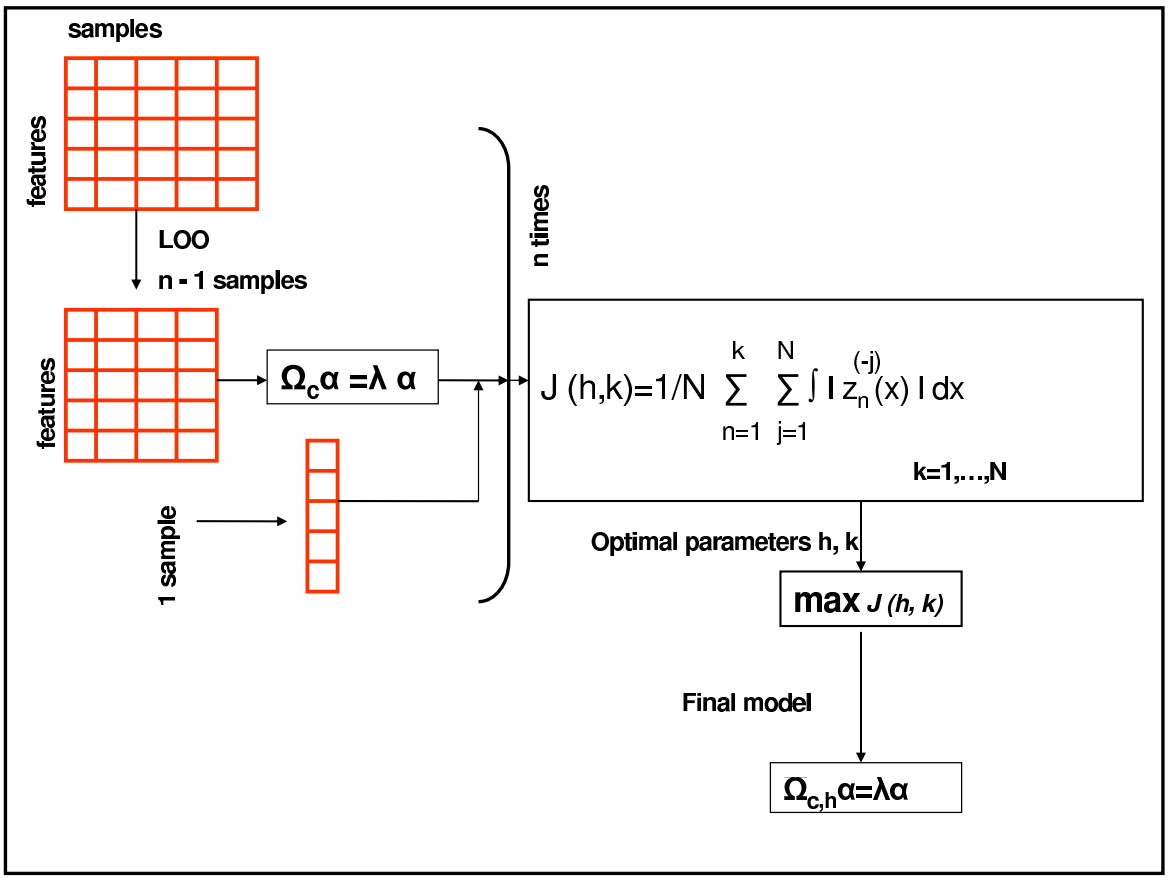
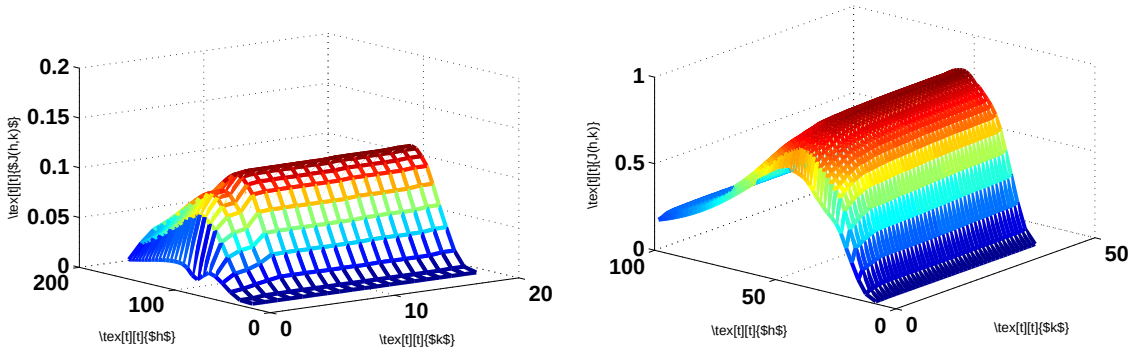


Figure 2: Data-Driven Bandwidth Selection for KPCA

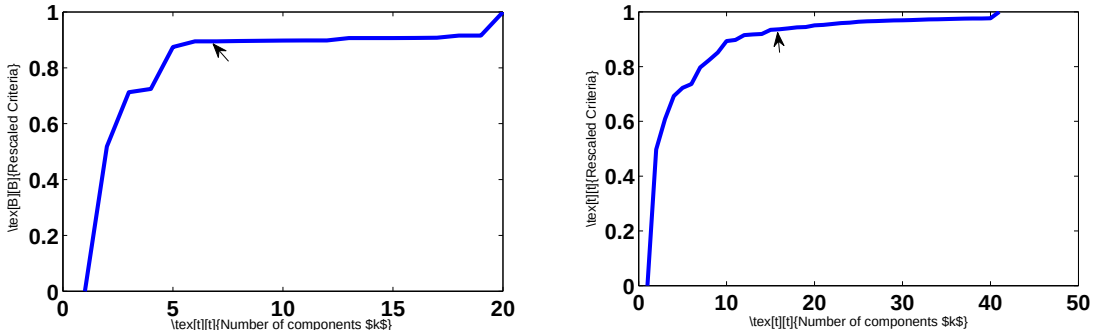
$$J(h,k) = \underset{h\in\mathbb{R}_0^+,k\in\mathbb{N}_0}{\operatorname{argmax}} \frac{1}{N}\sum_{n=1}^{k}\sum_{j=1}^{N}\int |z_n^{(-j)}(x)|dx \tag{5}$$

where $k = 1,\ldots,N$. Figure 2 illustrate the proposed model. Figure 3 shows the surface plot of Equation (5) for



(a) (b)

Figure 3: Model selection for KPCA-optimal bandwidth and number of components.(a) Cervical cancer (b) Colon cancer.



(a) (b)

Figure 4: Slice plot for the Model selection for KPCA for the optimal bandwidth.(a) Cervical cancer (b) Colon cancer.

various values of h and k. Thus, the proposed data-driven model can obtain the optimal bandwidth for KPCA, while retaining minimum number of eigenvectors which capture the majority of the variance of the data. Figure 4 shows a slice of the surface plots. The values of the proposed criterion were re-scaled to be maximum 1. The parameters that maximize Equation (5) are $h = 70.71$ and $k = 5$ for cervical cancer data and $h = 43.59$ and $k = 15$ for colon cancer data.

Classification Models

The constrained optimization problem for an LS-SVM [16, 33] for classification has the following form:

$$\min_{w,b,e}(\frac{1}{2}w^T w + \gamma\frac{1}{2}\Sigma_{k=1}^{N} e_k^2)$$

subject to:

$$y_k[w^T\phi(x_k)+b] = 1-e_k, \qquad k=1,\ldots,N$$

where $\phi(.)$: $\mathbb{R}^d \to \mathbb{R}^{d_h}$ is a nonlinear function which maps the d-dimensional input vector x from the input space to the d_h-dimensional feature space, possibly infinite. In the dual space the solution is given by

$$\begin{bmatrix} 0 & y^T \\ y & \Omega + \frac{I}{\gamma} \end{bmatrix} \begin{bmatrix} b \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 1_v \end{bmatrix}$$

with $y = [y_1,\ldots,y_N]^T$, $1_N = [1,\ldots,1]^T$, $e = [e_1,\ldots,e_N]^T$, $\beta = [\beta_1,\ldots,\beta_N]^T$ and $\Omega_{i,j} = y_i y_j K(x_i,x_j)$ where $K(x_i,x_j)$ is the kernel function. The classifier in the dual space takes the form

$$y(x) = \text{sign}[\sum_{k=1}^{N} \beta_k y_k K(x,x_k) + b] \tag{6}$$

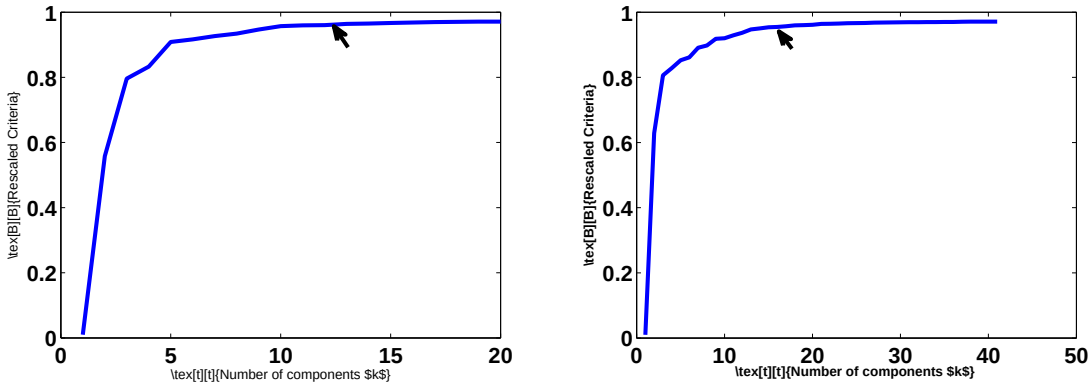
where β_k are Lagrange multipliers.

Results

First we considered nine data sets described in Table 1. We have chosen the RBF kernel $K(x_i,x_j) = \exp(-\frac{||x_i-x_j||^2}{2h^2})$ for KPCA. In this section all the steps are implemented using Matlab R2012b and LS-SVMlab v1.8 toolbox [34]. Next, we compared the performance of the proposed method with classical PCA and an existing tuning algorithm for RBF-KPCA developed by Pochet *et al.* [17]. Later, with the intention to comprehensively compare PCA+LS-SVM and KPCA+LS-SVM with other classification methods, we applied four widely used classifiers to the microarray data, being LS-SVM on whole data sets, t-test + LS-SVM, PAM and Lasso. To fairly compare kernel functions of the LS-SVM classifier; linear, RBF and polynomial kernel functions are used (in Table 2 referred to as linear/poly/RBF). The average test accuracies and execution time for all these methods when applied to the 9 case studies are shown in Table 2 and Table 4 respectively. Statistical significance test results (two-sided signed rank test) are given in Table 3 which compares the performance of KPCA with other classifiers. For all these methods, training on 2/3rd of the samples and testing on 1/3rd of the samples was repeated 30 times.

Comparison between the proposed criterion and PCA

For each data set, the proposed methodology is applied. This methodology consists of two steps. First, Equation (5) is maximized in order to obtain an optimal bandwidth h and corresponding number of components k. Second, the reduced data set is used to perform a classification task with LS-SVM. We retained 5 and 15 components respectively for cervical and colon cancer data sets. For PCA, the optimal number of components were selected by slightly modifying the Equation 5, i.e., which performed only for the components k. Figure 5 shows the plots of the



(a) (b)

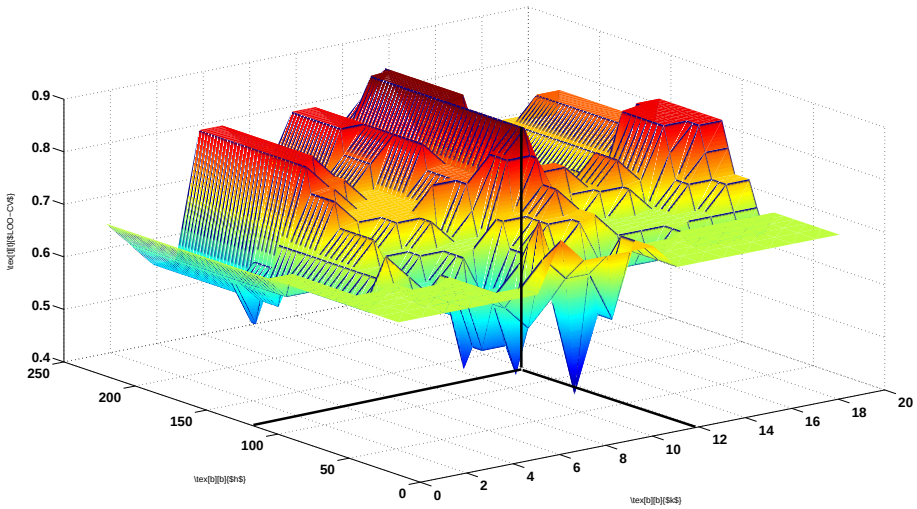
Figure 5: Plot for the selection of optimal number of components for PCA.(a) Cervical cancer (b) Colon cancer.

optimal components selection of PCA. Thus we retained 13 components and 15 components for cervical and colon cancer respectively for PCA. Similarly, we obtained number of components of PCA and the number of components with corresponding bandwidth for KPCA for the remaining data sets.

The score variables (projection of samples onto the direction of selected principal components) are used to develop an LS-SVM classification model. The averaged test AUC values over the 30 random repetitions were reported.

Comparison between the proposed criterion and an existing optimization algorithm for RBF-KPCA

We selected two experiments from Pochet *et al.* [17] (last two data sets in Table 1), being high-grade glioma and breast cancer II data sets. We repeated the same experiments as reported in Pochet *et al.* [17] and compared with the proposed strategy. The results are shown in Table 5. The three dimensional surface plot of LOOCV performance of the method proposed by [17] for the high-grade glioma data set is shown in Figure 6, with the optimal $h = 114.018$ and $k = 12$. The optimum parameters are $h = 94.868$ and $k = 10$ obtained by the proposed strategy (see Equation (5)) for the same data set. When looking at test AUC in Table 5, both case studies applying the proposed strategy, perform better than the method proposed by Pochet *et al.* [17] with less variability. In addition, the tuning method Pochet *et al.* [17] appears to be quite time consuming, whereas the proposed model enjoys a distinctive advantage with its low time complexity to carry out the same process.



(a)

Figure 6: LOO-CV performance of optimization algorithm [17] on high-grade glioma data set

Comparison between the proposed criterion and other classifiers

When looking specifically at all these methods in term of test AUC, we note that LS-SVM performance was slightly low on PCA. On breast cancer I, cervical cancer and HIV encephalitis data sets LS-SVM with linear kernel performs significantly better in terms of test AUC. The t-test + LS-SVM classifier shows the best test AUC for Leukemia and EDMD data sets. LS-SVM with linear kernel and t-test + LS-SVM classifiers have approximately the same test AUC on ovarian cancer and head & neck squamous cell carcinoma data sets. The proposed strategy with LS-SVM (RBF) classifiers offer better test AUC for colon cancer, breast cancer I, cervical cancer and head & neck squamous cell carcinoma data sets. Only on pancreatic data set, Lasso outperformed all other case studies. The test AUC of PAM was significantly worse on all data sets except DMD data set.

Discussions

The obtained test AUC of different classifiers on nine data sets, do not direct to a common conclusion that one method outperforms the other. Instead, it shows that each of these methods have its own advantage in classification tasks. When considering classification problems without dimensionality reduction, the regularized LS-SVM classifier shows a good performance on 50 percentage of data sets. Up till now, most microarray data sets are smaller in the sense of number of features and samples, but it is expected that these data sets might become larger or perhaps represent more complex classification problems in the future. In this situation, dimensionality reduction processes

(feature selection and feature transformation) become quite prominent.

The selected features of feature selection method such as t-test, PAM and Lasso widely vary for each random iteration. Further, the classification performance of these methods on each iteration depends on the number of features selected. Table 6 shows the range, i.e. minimum and maximum number of features selected on 30 iterations. PAM and Lasso outperformed only in two case studies. However PAM is a user friendly toolbox for gene selection and classification tasks, its performance depends heavily on the selected features. In addition, it is interesting that the Lasso selected only very small subsets of the actual data sets. But, in the Lasso, the amount of shrinkage varies, depending on the value of the tuning parameter, which is often determined by cross validation [35]. The number of genes selected as the outcome-predictive genes, generally decrease as the value of the tuning parameter increases. The optimal value of the tuning parameter, that maximizes the prediction accuracy is determined; however, the set of genes identified using the optimal value contains the non-outcome-predictive genes (ie, false positive genes) in many cases [9].

The test AUC on all nine case studies shows that KPCA performs better than classical PCA. But the parameters of KPCA need to be optimized. But here we have used LOOCV approach for parameters selection (bandwidth and number of components) of KPCA. In the optimization algorithm proposed by Pochet *et al.* [17], the combination of KPCA with RBF kernel and selection of principal components followed by FDA tends to result in overfitting. The proposed parameter selection criterion of KPCA with RBF kernel, often results in test set performances (see Table 4) that is better than using KPCA with a linear kernel, which reported in Pochet et al. Thus it means that LOOCV in the proposed parameter selection criterion does not encounter an overfitting for KPCA with RBF kernel function. In addition, the optimization algorithm proposed by Pochet *et al.* is completely coupled with the subsequent classifier and thus it appears to be very time-consuming.

In combination with classification methods, microarray data analysis can be useful to guide clinical management in cancer studies. In this study, several mathematical and statistical techniques were evaluated and compared in order to optimize the performance of clinical predictions based on microarray data. Considering the possibility of increasing size and complexity of microarray data sets in future, dimensionality reduction and nonlinear techniques have its own significance. In many cases, in a specific application context the best feature set is still important (e.g. drug discovery). While considering the stability and performance (both accuracy and execution time) of classifiers, the proposed methodology has its own importance to predict classes, of future samples of known disease cases.

Conclusion

The objective in, class prediction with microarray data is an accurate classification of cancerous samples which allows directed and more successful therapies. In this paper, we proposed a new data-driven bandwidth selection criterion for KPCA (which is a well defined preprocessing technique). In particular, we optimize the bandwidth and the number of components to maximize, the projected variance of KPCA. In addition, we compared several data preprocessing techniques prior to classification. In all the case studies, most of these data preprocessing steps performed well on classification with approximately similar performance. We observed that in feature selection methods selected features widely vary on each iteration. Hence it is difficult, even impossible to design a stable class predictor for future samples with these methods. Experiments on nine data sets show that the proposed strategy provides a stable preprocessing algorithm for classification of high dimensional data with good performance on test data.

The advantages of the proposed KPCA+LS-SVM classifier were presented in four aspects. First, we propose a data-driven bandwidth selection criterion for KPCA by tuning the optimum bandwidth and the number of principal components. Second, we illustrate that the performance of the proposed strategy is significantly better than an existing optimization algorithm for KPCA. Third, its classification performance is not sensitive to any number of selected genes, so the proposed method is more stable than others proposed in literature. Fourth, it reduces the dimensionality of the data while keeping as much information as possible of the original data. This leads to computationally less expensive and more stable results for massive microarray classification.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MT performed bandwidth selection, subsequent classification and drafted the paper. KDB participated in the design and implementation of framework. KDB and BDM helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

BDM is full professor at the Katholieke Universiteit Leuven, Belgium. Kris De Brabanter is supported by an FWO fellowship grant. Research supported by Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, START 1, OT 09/052 Biomarker, several PhD/postdoc and fellow grants; Flemish Government: IOF:

IOF/HB/10/039 Logic Insulin, IOF: HB/12/022 Endometriosis; FWO: PhD/postdoc grants, projects: G.0871.12N (Neural circuits) research community MLDM; G.0733.09 (3UTR); G.0824.09 (EGFR); IWT: PhD Grants; TBM-Logic Insulin, TBM Haplotyping; FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate); Hercules Stichting: Hercules III PacBio RS; iMinds 2013; IMEC: phd grant; EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED; COST: Action BM1104: Mass Spectrometry Imaging, Action BM1006: NGS Data analysis network. The scientific responsibility is assumed by its authors.

References

1. Roth V, Lange T: **Bayesian Class Discovery in Microarray Data**. *IEEE Transactions on Biomedical Engineering* 2004, **51**.
2. Qiu P, Plevritis SK: **Simultaneous Class Discovery and Classification of Microarray Data Using Spectral Analysis**. *Journal of Computational Biology* 2009, **16**:935–944.
3. Somorjai RL, Dolenko B, Baumgartner R: **Class prediction and discovery using gene microarray and proteomics mass spectroscopy data:curses,caveats, cautions**. *Bioinformatics* 2003, **19**:1484–1491.
4. Conde L, Mateos A, Herrero J, J D: **Improved Class Prediction in DNA Microarray Gene Expression Data by Unsupervised Reduction of the Dimensionality followed by Supervised Learning with a Perceptron**. *Journal of VLSI Signal Processing* 2003, **35**:245–253.
5. Tibshirani RJ, Hastie TJ, Narasimhan B, G C: **Diagnosis of multiple cancer types by shrunken centroids of gene expression**. *PNAS* 2002, **99**:6567–6572.
6. Chu F, Wang L: **Application of Support Vector Machine to Cancer Classification with Microarray Data**. *International Journal of Neural systems, World Scientific* 2005, **5**:475–484.
7. Chun LH, Wen CL: **Detecting differentially expressed genes in heterogeneous disease using half Student's t-test**. *Int.I.Epidemiol* 2010, **10**:1–8.
8. Tibshirani R: **Regression shrinkage and selection via the lasso**. *J. Roy. Statist. Soc. B* 1996, **58**:267–288.
9. Kaneko S, Hirakawa A, Hamada C: **Gene Selection using a High-Dimensional Regression Model with Microarrays in Cancer Prognostic Studies**. *Cancer Inform* 2012, **11**:29–39.
10. Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R: *Advances in Knowledge Discovery and Data Mining*. AAAI/ MIT Press 1997.
11. Pechenizkiy M, Tsymbal A, Puuronen S: **PCA-based Feature Transformation for Classification:Issues in Medical Diagnostics**. In *Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems* 2004.
12. Ng A, Jordan M, Weiss Y: **On spectral clustering: Analysis and an algorithm**. In *Advances in Neural Information Processing Systems 14, Proceedings of the 2001* 2001:849–856.
13. Liu Z, Chen D, Bensmail H: **Gene Expression Data Classification With Kernel Principal Component Analysis**. *J Biomed biotechnol* 2005, **2**:155–159.
14. Scholkopf B, Smola AJ, Muller KR: **Nonlinear component analysis as a kernel eigenvalue problem**. *Neural Computation.* 1998b, **10**:1299–1319.
15. Suykens JAK, Van Gestel T, De Moor B: **A support vector machine formulation to PCA analysis and its kernel version**. *IEEE Transactions on Neural Networks* 2003, **14**:447–450.
16. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J: *Least Squares Support Vector Machines*. Singapore: World Scientific 2002.
17. Pochet N, De Smet F, Suykens JAK, De Moor B: **Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction**. *Bioinformatics* 2004, **20**:3185–3195.
18. **Bioinformatics Research Group** [http://www.upo.es/eps/bigs/datasets.html].
19. Alon U, Barkai N, A Notterman D, Gish K, Ybarra S, Mack D, J Levine A: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays**. *PNAS* 1999, **96**:6745–6750.
20. Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA, Booser D, Theriault RL, Buzdar AU, Dempsey PJ, Rouzier R, Sneige N, Ross JS, Vidaurre T, Gómez HL, Hortobagyi GN, Pusztai L: **Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer**. *J Clin Oncol* 2006, **24**:4236–4244.
21. **FDA-NCI Clinical Proteomics Program Databank** [http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp].
22. Hingorani SR, Petricoin EF, Maitra A, Rajapakse V, King C, Jacobetz MA, Ross S, Conrads TP, Veenstra TD, Hitt BA, Kawaguchi Y, Johann D, Liotta LA, Crawford ME H Cand Putt, Jacks T, Wright CV, Hruban RH, Lowy AM, Tuveson DA: **Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse**. *Cancer Cell* 2003, **4(6)**:437–50.

23. Wong YF, Selvanayagam ZE, Wei N, Porter J: **Expression genomics of cervical cancer: molecular classification and prediction of radiotherapy response by DNA microarray**. *Clin Cancer Res* 2003, **9(15)**:5486–92.

24. Stirewalt DL, Meshinchi S, Kopecky KJ, Fan W: **Identification of genes with abnormal expression changes in acute myeloid leukemia**. *Genes Chromosomes Cancer* 2008, **47(1)**:8–20.

25. Kuriakose MA, Chen WT, He ZM, Sikora AG: **Selection and validation of differentially expressed genes in head and neck cancer**. *Cell Mol Life Sci* 2004, **61(11)**:1372–83.

26. Pescatori M, Broccolini A, Minetti C, Bertini E: **Gene expression profiling in the early phases of DMD: a constant molecular signature characterizes DMD muscle from early postnatal life throughout disease progression**. *FASEB J* 2007, **21(4)**:1210–26.

27. Masliah E, Roberts ES, Langford D, Everall I: **Patterns of gene dysregulation in the frontal cortex of patients with HIV encephalitis**. *J Neuroimmunol* 2004, **157(1-2)**:163–75.

28. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd U Cand Pohl, Hartmann C, McLaughlin ME, Batchelor TT, Black PM, von Deimling A, Pomeroy SL, Golub TR, Louis DN: **Gene expression-based classification of malignant gliomas correlates better with survival than histological classification**. *Cancer Res.* 2003, **63**:1602–1607.

29. van't Veer LJ, Dai H, Van De Vijver MJ, HeY D, Hart AAM, Mao M, Peterse HL, Van Der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernard R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**:530–536.

30. Mercer J: **Functions of positive and negative type and their connection with the theory of integral equations**. *Philosophical Transactions of the Royal Society A* 1909, **209**:415–446.

31. Bowman AW: **An Alternative Method of Cross-Validation for the Smoothing of Density Estimates**. *Biometrika* 1984, **71**:353–360.

32. Rudemo M: **Empirical choice of histograms and kernel density estimators**. *Scand. J. Statist.* 1982, **9**:65–78.

33. Suykens JAK, Vandewalle J: **Least Squares Support Vector Machine classifiers**. *Neural Processing Letters* 1999, **9**:293–300.

34. De Brabanter K, Karsmakers P, Ojeda F, Alzate C, De Brabanter J, Pelckmans K, De Moor B, Vandewalle J, Suykens JAK: **LS-SVMlab Toolbox User's Guide version 1.8** 2010.

35. Verweij PJ, Houwelingen HC: **Cross-validation in survival analysis**. *Stat Med* 1993, **12**:2305–14.

Figures

Figure 1 - Bandwidth selection for KPCA: cervical and colon cancer

Bandwidth selection of KPCA for a fixed number of components (A): 5 components for cervical cancer (B) 15 components for colon cancer.

Figure 2 - Data-Driven Bandwidth Selection for KPCA

Leave-one-out cross validation (LOOCV) for KPCA.

Figure 3 - Model selection for KPCA

(A) cervical cancer (B) colon cancer

Figure 4 - Slice plot for the model selection for KPCA

(A) cervical cancer (B) colon cancer

Figure 5 - LOOCV performance of the optimization algorithm [17] on high-grade glioma data set
Figure 6 - Plot for the selection of optimal number of components for PCA

(A) cervical cancer (B) colon cancer.

Tables

Table 1: Summary of the 11 binary disease data sets

Table 1: Summary of the 11 binary disease data sets.

Data set	#Samples		#Genes
	Class 1	Class2	
1: Colon	22	40	2000
2: Breast cancer I	34	99	5970
3: Pancreatic	50	50	15154
4: Cervical	8	24	10692
5: Leukemia	26	38	22283
6: Ovarian	91	162	15154
7: Head & neck squamous cell carcinoma	22	22	12625
8: Duchenne muscular dystrophy	23	14	22283
9: HIV encephalitis	16	12	12625
10: High grade glioma	29	21	12625
11: Breast cancer II	19	78	24188

Table 2: Comparison of classifiers: Mean AUC(std) of 30 iterations
Table 3: Statistical significance test results which compares KPCA with other classifiers: whole data, PCA, t-test, PAM and Lasso.
Table 4: Summary of averaged execution time of classifiers over 30 iterations in seconds
Table 5: Comparison of performance of proposed criterion with the method proposed by Pochet et al. [17]
Table 6: Summary of the range (minimum to maximum) of features selected over 30 iterations

Table 2: Comparison of classifiers: Mean AUC(std) of 30 iterations

Data set	Kernel function for Classification		preprocessing +	LS-SVM classifier		PAM	Lasso
		whole data	PCA	KPCA	t-test($p<0.05$)		
	RBF	0.769(0.127)	0.793(0.081)	0.822(0.088)	0.835(0.078)		
I	lin	0.822(0.068)	0.837(0.088)	**0.864**(0.078)	**0.857**(0.078)	0.787(0.097)	**0.837**(0.116)
	poly	0.818(0.071)	0.732(0.072)	0.825(0.125)	0.845(0.017)		
	RBF	0.637(0.146)	0.749(0.093)	**0.780**(0.076)	0.779(0.082)		
II	lin	**0.803**(0.059)	0.772(0.094)	**0.790**(0.075)	0.751(0.071)	0.659(0.084)	0.766(0.074)
	poly	0.701(086)	0.752(0.063)	0.753(0.072)	0.784(0.059)		
	RBF	0.832(0.143)	0.762(0.066)	0.879(0.058)	0.921(0.027)		
III	lin	**0.915**(0.043)	0.785(0.063)	0.878(0.066)	**0.941**(0.036)	0.707(0.067)	**0.9359**(0.0374)
	poly	0.775(0.080)	0.685(0.105)	0.8380(0.068)	0.858(0.042)		
	RBF	0.615(0.197)	0.853(0.112)	0.867(0.098)	0.808(0.225)		
IV	lin	**0.953**(0.070)	0.917(0.083)	**0.929**(0.077)	**0.987**(0.028)	0.759(0.152)	0.707(0.194)
	poly	0.762(0.118)	0.811(0.140)	0.840(0.131)	0.779(0.123)		
	RBF	0.807(0.238)	0.790(0.140)	0.976(0.035)	0.998(0.005)		
V	lin	**0.997**(0.005)	0.528(0.134)	0.982(0.022)	**0.998**(0.006)	0.923(0.062)	0.934(0.084)
	poly	0.942(0.051)	0.804(0.121)	0.975(0.028)	**0.965**(0.049)		
	RBF	**0.998**(0.001)	0.982(0.002)	0.984(0.012)	**0.998**(0.004)		
VI	lin	0.990(0.005)	0.973(0.002)	0.978(0.013)	0.993(0.013)	0.960(0.016)	0.951(0.045)
	poly	**0.998**(0.006)	0.985(0.016)	0.973(0.018)	0.995(0.011)		
	RBF	0.946(0.098)	0.941(0.057)	0.932(0.071)	0.967(0.048)		
VII	lin	**0.983**(0.025)	0.947(0.047)	**0.954**(0.051)	**0.987**(0.022)	0.931(0.058)	0.952(0.030)
	poly	0.785(0.143)	0.903(0.078)	0.915(0.080)	0.920(0.025)		
	RBF	0.823(0.159)	0.923(0.096)	0.858(0.113)	0.950(0.150)		
VIII	lin	0.840(0.164)	0.969(0.044)	0.800(0.019)	**0.999**(0.005)	**0.982**(0.050)	0.890(0.081)
	poly	0.781(0.186)	0.870(0.117)	0.785(0.121)	**0.998**(0.007)		
	RBF	0.638(0.210)	0.823(0.159)	0.852(0.180)	0.815(0.200)		
IX	lin	**0.931**(0.126)	0.840(0.164)	0.846(0.143)	**0.930**(0.139)	0.703(0.175)	0.705(0.174)
	poly	0.841(0.176)	0.781(0.186)	0.798(0.193)	0.768(0.193)		

p-value: False Discovery Rate (FDR) corrected.

Table 3: Statistical significance test which compares KPCA with other classifiers: whole data, PCA, t-test, PAM and Lasso. P-values of two-sided signed test are given.

kernel function	Dataset	I	II	III	IV	V	VI	VII	VIII	IX
	whole data	0.572	0.201	0.185	3.00E-04	0.115	8.16E-03	0.041	0.004	0.003
	PCA	1.00E-03	4.20E-05	7.25E-05	3.00E-04	5.65E-09	1.00E-07	0.005	1.00E-04	2.80E-03
RBF	t-test	0.856	0.711	0.999	0.458	1.30E-04	8.70E-05	0.064	0.0001	0.678
	PAM	0.362	1.22E-05	5.00E-05	0.016	0.029	8.52E-03	0.69	4.82E-05	3.00E-04
	Lasso	0.016	0.585	2.82E-05	0.029	0.23	3.00E-04	0.987	7.80E-06	0.004
	whole data	0.919	0.061	0.997	0.919	0.989	0.664	0.791	1.87E-04	0.839
	PCA	1.53E-05	2.16E-04	2.60E-10	1.54E-09	2.56E-09	5.43E-06	1.23E-07	1.86E-09	3.40E-08
lin	t-test	0.988	0.043	0.144	0.664	0.031	0.023	0.995	0.109	0.989
	PAM	0.008	7.53E-05	8.43E-03	3.00E-04	7.53E-05	3.53E-05	3.00E-04	0.876	0.005
	Lasso	0.099	0.099	9.84E-04	4.23E-07	1.00E-04	9.86E-03	5.00E-04	0.963	4.23E-07
	whole data	0.956	0.002	0.901	8.31E-12	9.00E-04	1.18E-08	1.54E-08	0.327	0.424
	PCA	2.60E-03	7.60E-05	8.70E-09	9.00E-06	1.54E-08	8.91E-04	6.55E-09	5.43E-06	1.00E-11
poly	t-test	0.557	0.585	0.005	0.856	0.031	0.043	0.985	1.00E-04	0.3612
	PAM	0.024	1.00E-04	0.003	0.008	0.006	0.016	3.00E-04	0.004	0.013
	Lasso	0.002	0.998	9.22E-06	3.51E-09	0.100	0.016	1.26E-08	2.16E-04	0.087

Table 4: Summary of averaged execution time of classifiers over 30 iterations in seconds.

Dataset	whole data	PCA	KPCA	t-test ($p < 0.05$)	PAM	Lasso
1: Colon	17	10	18	13	8	72
2: Breast	56	38	54	42	12	258
3: Pancreatic	17	12	26	19	20	453
4: Cervical	43	28	29	33	43	106
5: Leukemia	225	185	184	195	28	680
6: Ovarian	51	25	39	44	19	865
7: Head & neck squamous cell carcinoma	59	39	45	47	30	238
8: Duchenne muscular dystrophy	146	115	113	110	80	20100
9: HIV encephalitis	45	27	27	28	88	118

Table 5: Comparison of performance of proposed criterion with the method proposed by Pochet *et al.* [17]: Averaged test AUC(std) over 30 iterations and execution time in minutes

Data set	proposed strategy		Pochet *et al.* [17]	
	Test AUC	time	Test AUC	time
high-grade glioma data	0.746(0.071)	2	0.704(0.104)	38
breast cancer II	0.6747(0.1057)	4	0.603(0.157)	459

Table 6: Summary of the range (minimum to maximum) of features selected over 30 iterations.

Dataset	t-test ($p < 0.05$)	PAM	Lasso
1: Colon	197-323	15-373	8-36
2: Breast	993-1124	13-4718	7-87
3: Pancreatic	2713-4855	3-1514	12-112
4: Cervical	5858-6756	2-10692	5-67
5: Leukemia	1089-2654	137-11453	2-69
6: Ovarian	7341-7841	34-278	62-132
7: Head and neck squamous cell carcinoma	307-831	1-12625	3-35
8: Duchenne muscular dystrophy	973-2031	129-22283	8-24
9: HIV encephalitis	941-1422	1-12625	1-20

p-value: False Discovery Rate (FDR) corrected.