Schedae Informaticae Vol. 25 (2016): 25–35
doi: 10.4467/20838476SI.16.002.6183

tfml 2017
theoretical foundations
of machine learning, Kraków

Misclassification-Driven Sample Relabeling for Supervised Kernel Principal Component Analysis

Krzysztof Adamiak, Krzysztof Ślot
Institute of Applied Computer Science, Łódź University Technology
Stefanowskiego 18/22, 90-924 Łódź, Poland
e-mail: *krzysztof.adam.adamiak@gmail.com, kslot@p.lodz.pl*

Abstract. Supervised kernel-Principal Component Analysis (S-kPCA) is a me thod for producing discriminative feature spaces that provide nonlinear decision regions, well-suited for handling real-world problems. The presented paper proposes a modification to the original S-kPCA concept, which is aimed at improving class-separation in resulting feature spaces. This is accomplished by identifying outliers (understood here as misclassified samples) and by an appropriate reformulation of the original S-kPCA problem. The proposed idea is to replace binary class labels that are used in the original method, by real-valued ones, derived using sample-relabeling scheme aimed at preventing potential data classification problems. The postulated concept has been tested on three standard pattern recognition datasets. It has been shown that classification performance in feature spaces derived using the introduced methodology improves by 4–16% with respect to the original S-kPCA method, depending on a dataset.

Keywords: pattern recognition, feature extraction, kernel methods, supervised kernel PCA.

1. Introduction

Common attributes of datasets corresponding to hard, real-world data classification problems are presence of outliers, complex nonlinear and multi-modal character of

Received: 11 December 2016 / Accepted: 30 December 2016

class decision boundaries, uneven representation of classes and class' modes as well as noise and erroneous sample labeling. These problems have to be addressed by pattern recognition procedures that aspire to be of practical use. In fact, all well-established pattern recognition methods that have been developed so far, such as Support Vector Machines (SVM [1]), neural networks (especially trained using deep-learning techniques [2]) or probabilistic classifiers [3], attempt to handle the aforementioned problems. Some of these methods operate on raw data, but typically they assume object representations in carefully selected feature spaces. Therefore, feature space derivation becomes an important element of pattern recognition procedure and an enormous amount of research has been done in this field. Appropriate feature spaces facilitate classification process, eliminate curse of dimensionality problem, thus reducing a risk of classifier overfitting, but also, enable new insights into intrinsic object properties, relations and dependencies that can be revealed by an adopted representation.

A variety of feature-space derivation strategies have been proposed so far. They emphasize various aspects of data representation that are of importance for a given application. Criteria used for derivation of new feature spaces range from maximization of data scatter (PCA and its nonlinear extension – kernel PCA, abbreviated henceforth using the term kPCA), through maximization of sample independence (Independent Component Analysis [4], and its kernelized extension [5]) to maximization of class discrimination (Linear Discriminant Analysis along with its kernelized version and supervised versions of PCA). Other concepts behind a search for reduced representations of samples involve for example preservation of original data structure (as e.g. in Multidimensional Scaling or Iso-mapping).

The presented paper is concerned with a modification of Supervised Kernel Principal Component Analysis (S-kPCA) [6], which is a supervised extension to the kPCA (labeled samples are considered in feature space derivation), proposed in [7]. Kernel PCA in turn generalizes the classical PCA in such a way that the discovered maximum scatter directions become nonlinear. Properties of kPCA address several basic requirements crucial for classification of real-world data, such as low sensitivity to outliers or nonlinear data mapping, which is crucial for solving linearly non-separable problems. Atop on that, S-kPCA provides features that maximize correlations between samples and their class labels, thus eliminating one of the main drawbacks of scatter-maximization based strategies.

Despite numerous advantages, S-kPCA is clearly not an ultimate solution to the problem of feature space derivation. For difficult datasets it fails to provide perfect data separation and one of the reasons behind its deteriorating performance is a lack of diversification of individual samples' role in building new data representation. This issue is explored in research reported in the presented paper. We postulate to diversify significance of different samples by their appropriate relabeling, so that samples that may potentially pose classification problems become more important. We verify this concept on three publicly available pattern recognition datasets and we show that significant improvement (4–16%, depending on dataset) over the original S-kPCA approach can be obtained.

A structure of the paper is the following. We begin with a short review of related concepts: kPCA and S-kPCA. Then we present in detail the proposed sample relabeling principles and the adopted feature space derivation procedure. Finally, we provide results of experimental evaluation of the concept, where the modified S-kPCA method is confronted with the original one and feature spaces derived for both approaches are indirectly compared using classification performance results.

2. Related work

A basis for the presented research is laid out by an impressive development of kernel methods for data classification and processing, which followed a success of the Support Vector Classification (SVM) [8–12]. Theory of kernel methods has been expanded also onto data preprocessing domain and several 'kernelized' versions of well-established concepts were formulated. These include concepts that are directly related to the presented research: kernel Principal Component Analysis and its supervised version S-kPCA.

Kernel Principal Component Analysis, proposed in [7], extends classical Principal Component Analysis concept in order to identify nonlinear data scatter directions. A concept of implicit problem-solving in high-dimensional, intermediate spaces, which can be accomplished using kernels, provides a means for making the relevant computations feasible. An objective of kPCA is to find directions of the maximum variability among samples $\mathbf{x}_i$ that are projected to some high dimensional space, using a transformation $\Phi(.)$ (i.e. $\mathbf{X}_i = \Phi(\mathbf{x}_i)$). In other words, an objective is to find eigenvectors $\mathbf{V} = [\mathbf{v}_0, \mathbf{v}_1, ...]$ of the projected data covariance matrix:

$$(\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})^T \mathbf{V} = \Lambda \mathbf{V} \tag{1}$$

where $\mathbf{M}$ is a matrix of mean-valued vectors $\mathbf{m}$, computed for the projections in high-dimensional space, and Λ is a diagonal matrix of eigenvalues. As eigenvectors lie in a subspace defined by projected samples:

$$\mathbf{v}^i = \sum_{j=0}^{n-1} \alpha_j^i (\mathbf{X}_j - \mathbf{m}) = (\mathbf{X} - \mathbf{m})\mathbf{a}^i,$$

premultiplying the equation (1) by the term $(\mathbf{X} - \mathbf{M})^T$ yields alternative formulation of the eigenproblem:

$$(\mathbf{X} - \mathbf{M})^T (\mathbf{X} - \mathbf{M}) \mathbf{A} = \Lambda \mathbf{A} \tag{2}$$

where $\mathbf{A} = [\mathbf{a}^0, \mathbf{a}^1, ...]$ comprises vectors of coefficients that become a solution to the modified eigenproblem. Observe, that only dot products are involved in computations of the eigenproblem (2), so they can be replaced by kernels. Introducing a Gramm matrix, with elements $G_{i,j} = \hat{K}(\mathbf{x}_i, \mathbf{x}_j)$, where $\hat{K}$ is some kernel function, centered in high-dimensional space, one can rewrite (2) in a compact form:

$$\mathbf{G}\mathbf{A} = \Lambda \mathbf{A} \tag{3}$$

A solution to (3), which can be found for reasonable amounts of samples, defines directions of the maximum variability in a high-dimensional space and can be used for projecting unknown samples:

$$(\Phi(\mathbf{z})-\mathbf{m})^T\mathbf{v}^i=(\Phi(\mathbf{z})-\mathbf{m})^T(\mathbf{X}-\mathbf{m})\mathbf{a}^i=\left[\hat{K}(\mathbf{z},\mathbf{x}_0),...\hat{K}(\mathbf{z},\mathbf{x}_{n-1})\right]\mathbf{a}^i \tag{4}$$

As it can be seen, projections onto each eigenvector $\mathbf{v}_i$ can be determined in the original, low-dimensional space, using kernel operations and the computed coefficient vectors $\mathbf{a}^i$.

The second concept relevant to the presented paper is a supervised version of kPCA. The proposed idea is to use Hilbert-Schmidt Independence Criterion (HSIC) [13] as an objective function that is to be maximized. HSIC measures a level of cross-covariance between samples and their labels:

$$\mathbf{C}_{x,y}=E(\mathbf{X}-\mathbf{m}_x)(\mathbf{Y}-\mathbf{m}_y)^T=E(\mathbf{XH})(\mathbf{YH})^T \tag{5}$$

where $\mathbf{X}$ is a matrix of input samples with a mean vector $\mathbf{m}_x$, $\mathbf{Y}$ is a matrix of labels, with their mean $\mathbf{m}_y$, and $\mathbf{H}$ is a centering matrix. HSIC uses a Hilbert-Schmidt norm, which, in essence, aggregates squared entries of the cross-covariance (5). It can be easily shown that this can be expressed as:

$$HSIC=k\cdot\mathbf{tr}(\mathbf{C}_{x,y}\mathbf{C}_{x,y}^T) \tag{6}$$

where **tr** denotes the trace of a matrix and k is a scaling factor. As the criterion (6) involves dot products, one can introduce kernels: on input samples - $\mathbf{K}=[k(\mathbf{x}_i,\mathbf{x}_j)]$ and on labels - $\mathbf{L}=[l(\mathbf{y}_i,\mathbf{y}_j)]$, and rewrite the criterion in the form:

$$HSIC=k\cdot\mathbf{tr}(\mathbf{KHLH}) \tag{7}$$

An objective of S-kPCA procedure is to find such a transformation matrix $\mathbf{U}$ of original samples $\mathbf{x}$, i.e.:

$$\mathbf{x'}=\mathbf{Ux}$$

which, after plugging $\mathbf{x'}$ into (5) provides maximization of the criterion (7). This can be seen as searching for such a combination of original samples that ensure data transformations (through kernel functions) that maximize correlation between samples and their labels.

3. Misclassification-driven sample relabeling

The original S-kPCA procedure is not addressing an issue of diversifying sample labels and is not exploring its impact on discriminative properties of a resulting feature space. By default, all class samples are treated evenly: for example, for a two-class problem (samples belong either to class $\mathcal{A}$ or class $\mathcal{B}$), each sample $\mathbf{x}$ is labeled with a two-element vector $\mathbf{y}_i$ with binary entries:

$$\forall_{s_i \in \mathcal{A}} \quad \mathbf{y}_i = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \forall_{s_j \in \mathcal{B}} \quad \mathbf{y}_j = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{8}$$

where $\mathcal{A}$ -class sample label is linked to its class by the first component of the label vector and $\mathcal{B}$-class samples, by the second component.

As the criterion (6) that underlies feature space derivation focuses on sample-label correlation, it is evident that varying the label affects the relevant computations and leads to a different solution. One can observe that by varying numerical label representations one can modify sample-class correlations in several possible ways. For the considered two-class problem, where labels are represented by two-element vectors, one can vary any of the two entries. Moreover, one can easily interpret such changes. An increase in a value of sample's 'own' label vector component (the first for $\mathcal{A}$ -class samples and the second for $\mathcal{B}$-class samples) makes a class-sample correlation stronger, whereas decreasing this value - weakens the correlation. This way, one can diversify a significance of samples in a process of constructing a novel feature space. In addition, one can observe that a sample can be forced to negatively correlate with the opposite class. This can be accomplished by substituting the 'zero-correlation' component of the label vector (the second one for $\mathcal{A}$ -class samples and the first one for $\mathcal{B}$-class samples) with a negative value. As a result, every sample would get individual labels that can be represented as:

$$\forall_{s_i \in \mathcal{A}} \quad \mathbf{y}_i = \begin{bmatrix} a_i^p \\ -a_i^n \end{bmatrix}, \quad \forall_{s_j \in \mathcal{B}} \quad \mathbf{y}_j = \begin{bmatrix} b_i^p \\ -b_i^n \end{bmatrix}, \tag{9}$$

where $a_{..}, b_{..}$ are positive real numbers and the superscripts p and n identify 'positive' and 'negative' correlation coefficients.

The presented concept of sample relabeling seems a viable way to modify a role of different samples in derivation of new feature spaces. In particular, one can apply this mechanism to increase significance of samples that are harder to be correctly classified over significance of samples that pose no serious classification problems.

To improve discriminative properties of feature spaces derived using S-kPCA concept, one needs to elaborate rules that enable reasonable sample relabeling, i.e. that enable determining for which samples label alterations should be made and what should be a magnitude of such an alteration. We propose to use training sample classification results statistics as the basis for both determination of samples to be affected and determination of amounts of label changes.

The proposed procedure has been schematically depicted in Figure 1. Original dataset is split into two parts: training/validation and test sets. Feature space derivation is performed only on samples from the former one and begins with its random split into temporary training and temporary test subsets. Next, the original S-kPCA procedure, which assumes binary class labels (8) is executed on the temporary training subset, followed by data classification performed in the derived space using Gaussian Mixture Model (GMM) classifier. All misclassified samples are then recorded and saved for a future use. The GMM classification method has been chosen, as it has a little in common with the adopted kernel-based feature space derivation methodology and can therefore be considered as an unrelated tool for feature space evaluation.

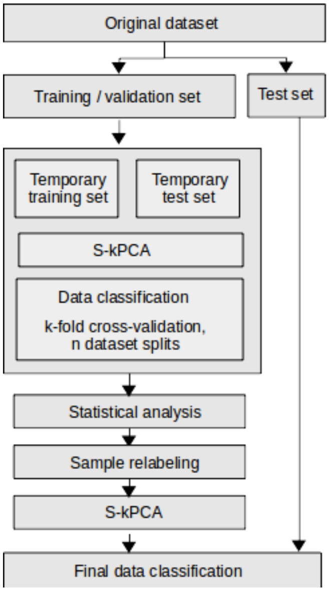

Figure 1. Block diagram of the proposed feature space derivation procedure.

For the current split of the training set, classification is performed using a k-fold cross validation scheme and once it is completed, the whole procedure is repeated n-times for another n random splits of the training set into new temporary training/test parts. After completion of this iterative procedure, misclassification percentage is determined for each sample. The computed coefficients are used as a basis for sample relabeling. Assuming that some $i-$th sample from a class $\mathcal{A}$ has been misclassified $m-$times, the corresponding correction coefficient is determined:

$$c_i = \alpha \frac{m}{n} \tag{10}$$

where α is a positive constant that controls a magnitude of label updates.

The coefficient (10) is then used to produce the 'positive' component of the sample label vector:

$$a_i^p = 1 + c_i \tag{11}$$

or the 'negative' component (updates for samples from a class $\mathcal{B}$ are analogous):

$$a_i^n = -c_i \tag{12}$$

Having the samples relabeled, another S-kPCA procedure is executed, producing the resulting feature space for classification of samples from the test set. Next, classification result is recorded and the whole procedure is repeated p-times for other random splits of the original dataset.

Table 1. Datasets used in experiments.

Database	Digits	Ionosphere	Pima	Glass	E-Coli	Parkinson	Heart
Classes	10	2	2	6	5	2	5
Attributes	64	34	8	10	8	23	14
Samples	1797	351	768	213	327	195	297

4. Experimental evaluation of the proposed concept

To verify the proposed concept, a series of experiments (using Python's sckit package [14]) on seven different pattern recognition datasets: Digits, Ionosphere, Pima Indian Diabetes, Glass, E-coli, Parkinson's Disease [15] and Cleveland Heart Disease [16] (all datasets available at UCI repository [17]) were performed. Dataset highlights are shown in Table 1. Three of these datasets correspond to binary classification problems, so that sample labels are represented by two-element vectors, as shown in (9). For the remaining datasets, the presented sample relabeling approach requires only a straightforward modification: label vectors are composed of multiple entries that get appropriately updated during the validation procedure.

Three different label alteration scenarios were considered during the experiments:

- Only positive components of sample's label vector (i.e. a_i^p or b_i^p depending on sample's class) were being updated according to (11)
- Only negative components of sample's label vector (i.e. a_i^n or b_i^n) were being updated according to (12)
- Both negative and positive components were modified

For each dataset, a procedure explained in the previous Section was executed. In each case we assumed a 5-fold cross validation ($k = 5$), the inner loop (i.e. estimations of misclassification rates for a given split of the original dataset) was executed 100 times (i.e. $n = 100$) and 100 classifications were made ($p = 100$) to asses an overall classification performance for each method. Throughout all experiments, a Gaussian kernel was used in S-kPCA procedure:

$$k(\mathbf{x}_i, \mathbf{x}_j) = exp\left(-\gamma \cdot \ ||\mathbf{x}_i - \mathbf{x}_j||^2\right) \tag{13}$$

where $||.||$ denotes a distance between samples and γ was chosen using the grid-search method [18].

An objective of the first phase of experiments was to compare the three adopted sample relabeling scenarios. Classification experiments (GMM method was used for data classification both in the validation and in the test step) were performed on the first three databases and the results, plotted as a function of the parameter $\alpha \in [0..2]$ (10) have been shown in Figure 2. The experiments have been summarized in Table 2. For comparison purposes, performance evaluation of GMM classification of raw data,

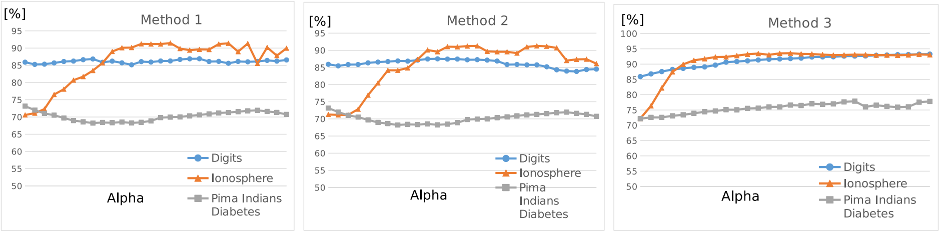

Figure 2. Average classification performance as a function of varying label update magnitudes for all considered scenarios and datasets.

Table 2. Classification performance using Gaussian Mixture Models (in percent) with 95% confidence intervals

Database	Digits	Ionosphere	Pima
Raw data	76,2 ± 3.6	72.4 ± 3.2	63.9 ± 4.7
S-kPCA	85.5 ± 1.8	72.1 ± 5.3	73.3 ± 2
Scenario 1	86.7 ± 1.9	91.4 ± 2.6	72.0 ± 2
Scenario 2	87.5 ± 2.1	91.2 ± 3	72.0 ± 2.4
Scenario 3	**93.2** ± 0.9	**93.0** ± 2	**76.0** ± 1.8

as well as GMM classification in a space derived using the original S-kPCA procedure (for identical splits into training and test parts) were provided. One can observe that classification in feature spaces derived using the proposed strategy outperforms the reference methods: classification made on raw data and classification made in a space derived using the original S-kPCA. Also, it can be seen that a combination of both types of updates, i.e. the last scenario used for label alterations, provides the most noticeable gain, so we consider this strategy to be the best one.

An improvement in classification performance is statistically significant for the two datasets: Digits and Ionosphere. In case of the last dataset (Pima Indian Diabetes), although average classification results are better than for the original S-kPCA, large variations of individual results do not allow making any definite statement on the proposed method's superiority. On the other hand one can observe that Pima dataset samples have relatively low initial dimensionality, so that dimensionality reduction in that case may not be necessary at all.

As sample relabeling involving alterations both to positive and negative label vector components has been found to provide the best results, this approach has been adopted in the following experiments. This time, classification performance of three different procedures, involving no dimensionality reduction (raw data classification), dimensionality reduction with original S-kPCA and dimensionality reduction with sample relabeling using the adopted third scenario, were done for all considered datasets. GMM was used as the classification strategy in validation phase, whereas test set samples were classified with either GMM, linear SVM and k-NN ($k = 5$ was used) classifiers.

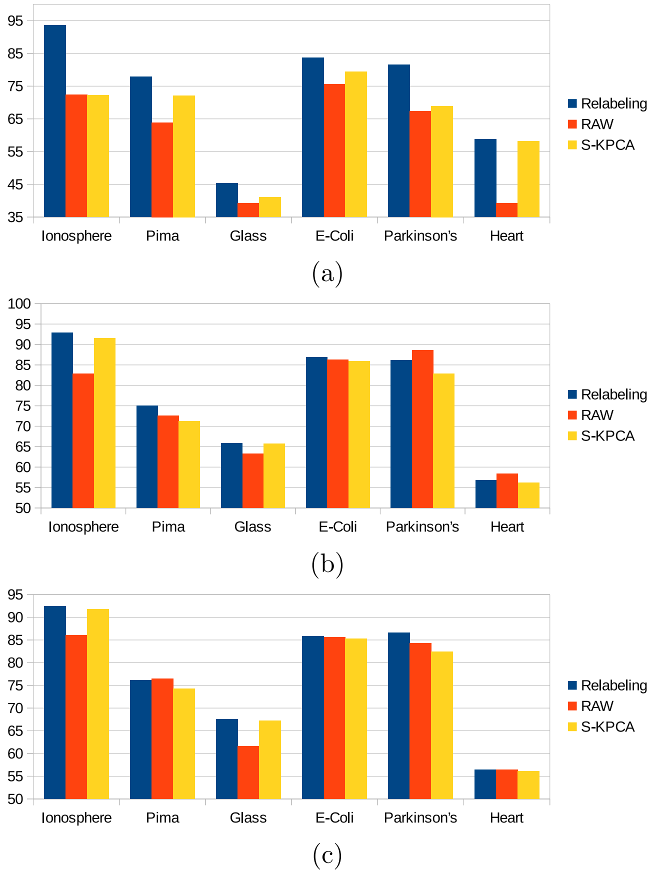

Figure 3. Maximum classification rates for different datasets and the three considered test-set classification methods: GMM (a), 5-NN (b) and linear SVM (c). 'Relabeling' denotes the proposed approach, 'RAW' denotes classification of original data and 'S-KPCA' denotes classification with the original method.

Performance comparison results are depicted in Figure 3, where the best performing α value for each dataset ($\alpha \in [0.3...1.1]$) was used for sample relabeling.

A few observations can be made from the presented results. The most important from the point of view of the proposed concept is that the proposed sample relabeling always results in feature spaces that have better class discrimination properties than the ones produced by the original S-kPCA. In each case, classification performance improves, however, in several cases this improvement is not statistically significant. Secondly, it seems that classification strategy adopted in validation phase (GMM) implies the best improvements if the same classification strategy is used in the test phase (differences between results obtained for the sample-relabeled and the original S-kPCA are the most salient). Finally, performance for different datasets depends on the adopted classification strategy.

5. Conclusion

A strategy for improving class separation properties for feature spaces derived using Supervised kernel Principal Component Analysis has been presented in the paper. It has been shown that appropriate modifications to sample labels, which diversify their significance in derivation of target space features, result in increased data classification performance and that this gain can be substantial for some datasets. Although the concept needs to be thoroughly verified using many other existing data sources, we believe that the observed tendency will hold, improving significance of the S-kPCA concept.

One needs to bear in mind, that PCA-based data recognition methods are inherently computationally complex, which limits their use in time-critical applications. This also applies to S-kPCA and to the proposed modification. On the other hand, the considered data preprocessing offers several crucial advantages – it reveals structures that exist among data and that can be relevant for class-discrimination, reduces a risk of classifier overfitting and reduces sensitivity to bad class examples.

6. References

[1] Burges C.J., *A tutorial on support vector machines for pattern recognition.* Data Mining and Knowledge Discovery, 1998, 2 (2), pp. 121–168.

[2] Bengio Y., *Learning deep architectures for ai.* Foundations and Trends in Machine Learning, 2009, 2 (1), pp. 1–127.

[3] Reynolds D., *Gaussian mixture models.* Encyclopedia of Biometrics, 2015, pp. 827–832.

[4] Comon P., *Independent component analysis: a new concept?* Signal Processing, 1994, 36 (3), pp. 287–314.

[5] Bach F.R., Jordan M.I., *Kernel independent component analysis.* Journal of Machine Learning Research, 2002, 3, pp. 1–48.

[6] Barshan E., Ghodsi A., Azimifar Z., Jahromi M.Z., *Supervised principal component analysis: visualization, classification and regression on subspaces and submanifolds.* Pattern Recognition, 2011, 44, pp. 1357–1371.

[7] Schölkopf B., Smola A., Müller K.R., *Nonlinear component analysis as a kernel eigenvalue problem.* Neural Computation, 1998, 10, pp. 1299–1319.

[8] Hofmann T., Schölkopf B., Smola A.J., *Kernel methods in machine learning.* The Annals of Statistics, 2008, 36 (3), pp. 1171–1220.

[9] Smola A.J., Schölkopf B., *Learning with Kernels.* MIT Press, 2002.

[10] Wang M., Sha F., Jordan M.I., *Unsupervised kernel dimension reduction.* Proc. of Conf. Advances in Neural Information Processing Systems, 2010, 23, pp. 2379–2387.

[11] Mika S., Rätsch G., Scholkoph W.J., Müller K.R., *Fisher discriminant analysis with kernels.* Proc. of IEEE Conf. Neural Networks for Signal Processing, 1999, pp. 41–48.

[12] Baudat G., Anouar F., *Feature vector selection and projection using kernels.* Neurocomputing, 2003, 55, pp. 21–38.

[13] Song L., Smola A., Gretton A., Bedo J., Borgwardt K., *Feature selection via dependence maximization.* Journal of Machine Learning Research, 2012, 13, pp. 1393–1434.

[14] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 2011, 12, pp. 2825–2830.

[15] Little M.A., McSharry P.E., Roberts S.J., Costello D.A., Moroz I.M., *Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection-6.* BioMedical Engineering OnLine, 2011, 6 (1), pp. 23.

[16] Hungarian Institute of Cardiology. Budapest: Andras Janosi M.D., University Hospital Zurich, Switzerland: William Steinbrunn M.D., University Hospital Basel, Switzerland: Matthias Pfisterer M.D., V.A. Medical Center Long Beach and Cleveland Clinic Foundation:Robert Detrano M.D. Ph.D., *Heart Disease Data Set.* [online].

[17] Moshe L., *UCI machine learning repository*, 2013.

[18] Chapelle O., Vapnik V., Bousquet O., Mukherjee S., *Choosing multiple parameters for support vector machines.* Machine Learning, 2002, 46 (1), pp. 131–159.