Anal. Chem. **2010,** *82,* 7000–7007

Visualization and Recovery of the (Bio)chemical Interesting Variables in Data Analysis with Support Vector Machine Classification

Patrick W. T. Krooshof,[†] Bülent Üstün,[‡] Geert J. Postma,[†] and Lutgarde M. C. Buydens*,[†]

Radboud University Nijmegen, Institute for Molecules and Materials, Analytical Chemistry, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands, and Analytical Sciences Chemicals, Quality Unit API/Biotech, MSD Oss, P.O. Box 20, 5340 BH Oss, The Netherlands

Support vector machines (SVMs) have become a popular technique in the chemometrics and bioinformatics field, and other fields, for the classification of complex data sets. Especially because SVMs are able to model nonlinear relationships, the usage of this technique has increased substantially. This modeling is obtained by mapping the data in a higher-dimensional feature space. The disadvantage of such a transformation is, however, that information about the contribution of the original variables in the classification is lost. In this paper we introduce an innovative method which can retrieve the information about the variables of complex data sets. We apply the proposed method to several benchmark data sets and a metabolomics data set to illustrate that we can determine the contribution of the original variables in SVM classifications. The corresponding visualization of the contribution of the variables can assist in a better understanding of the underlying chemical or biological process.

In the past decade support vector machines (SVMs) have become a popular technique in pattern recognition and regression estimation. Applications of SVMs are among the fields of bioinformatics,[1,2] medicine,[3−6] drug discovery,[7−9] text categorizing,[10] gene expression analysis,[11−13] face recognition,[14] spam categorizing,[15,16] financial forecasting,[17,18] and many others. Especially because of the possibility to model complex nonlinear relationships the application of SVMs has grown substantially.[19−22] By transforming the original input space into a high dimensional feature space, the nonlinear relationships can be presented in a linear form. This transformation is performed by using a specific kernel function.[6,23,24] Several kernel functions are proposed in the literature for this purpose and include variance-covariance based linear and polynomial kernels, the Euclidean distance based radial basis function (RBF) and the Pearson VII Universal Kernel (PUK) functions.[23−25] The transformation by a kernel function has also been introduced in other algorithms, such as Kernel Principal Component Analysis,[26] Kernel Partial Least Squares,[27,28] and Kernel Fisher Discrimination.[29]

However, the disadvantage of using such a kernel function is that the correlation between the obtained SVM model and the original input space is lost. Therefore it is not possible to determine which variables (e.g., spectral ranges) contribute to the final SVM results and a direct interpretation of the SVM model is not straightforward.[30,31] This seriously hampers the ultimate (bio)chemical interpretation of the resulting classification model. In this manuscript we propose a novel method to overcome this problem and reveal the importance of the original variables.

* To whom correspondence should be addressed. Phone: +31 24 3653180. Fax: +31 24 3652653. E-mail: l.buydens@science.ru.nl.

† Radboud University Nijmegen.

‡ Analytical Sciences Chemicals.

(1) Yang, Z. R. *Briefings Bioinf.* **2004**, *5* (4), 328–338.
(2) Ramo, P.; Sacher, R.; Snijder, B.; Begemann, B.; Pelkmans, L. *Bioinformatics.* **2009**, *25*, 3028–3030.
(3) Akay, M. F. *Expert Syst. Appl* **2009**, *36* (2), 3240–3247.
(4) Magnin, B.; et al. *Neuroradiology.* **2009**, *51* (2), 73–83.
(5) Luts, J.; Heerschap, A.; Suykens, J. A. K.; Van Huffel, S. *Artif. Intell. Med.* **2007**, *40* (2), 87–102.
(6) Conforti, D.; Guido, R. *Comput. Oper. Res.* **2010**, *37*, 1389–1394.
(7) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.* **2001**, *26*, 5–14.
(8) Warmuth, M. K.; et al. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.
(9) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048–2056.
(10) Leopold, E.; Kindermann, J. *Mach. Learn.* **2002**, *46*, 423–444.
(11) Furey, T. S.; et al. *Bioinformatics.* **2000**, *16* (10), 906–914.
(12) Clarke, R.; et al. *Nat. Rev. Cancer.* **2008**, *8*, 37–49.
(13) Noble, W. S. *Nat. Biotechnol.* **2006**, *24* (12), 1565–1567.
(14) Guo, G.; Li, S. Z.; Chan, K. L. *Image Visualization Comput.* **2001**, *19*, 631–638.
(15) Drucker, H.; Wu, D.; Vapnik, V. N. *IEEE Trans. Neural Networks* **1999**, *10* (5), 1048–1054.
(16) Guzella, T. S.; Caminhas, W. M. *Expert Syst. Appl.* **2009**, *36* (7), 10206–10222.
(17) Tay, F. E. H.; Cao, L. *Omega.* **2001**, *29*, 309–317.
(18) Kim, K. J. *Neurocomputing.* **2003**, *55*, 307–319.
(19) Vapnik, V. *Estimation of Dependence Based on Empirical Data*; Springer Verlag: New York, 1982.
(20) Vapnik, V. *the Nature of Statistical Learning Theory*; Springer Verlag: New York, 1995.
(21) Cortes, C.; Vapnik, V. *Mach. Learn.* **1995**, *20*, 273–297.
(22) Vapnik, V. *Statistical Learning Theory*; John Wiley and Sons: New York, 1998.
(23) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press.: Cambridge, 2000.
(24) Schölkopf, B.; Smola, A. J. *Learning with Kernels*; MIT Press.: Cambridge, 2002.
(25) Üstün, B.; Melssen, W. J.; Buydens, L. M. C. *Chemom. Intell. Lab. Syst.* **2006**, *81*, 29–40.
(26) Schölkopf, B.; Smola, A. J.; Müller, K. R. *Neural Comput.* **1998**, *10*, 1299–1319.
(27) Walczak, B.; Massart, D. L. *Anal. Chim. Acta* **1996**, *331*, 177–185.
(28) Rosipal, R.; Trejo, L. J. *J. Mach. Learn. Res.* **2001**, *2*, 97–123.
(29) Mika, S.; Rätsch, G.; Wetson, J.; Schölkopf, B.; Müller, K. R. *Proc. NNSP'99;* **1999**, 41–48.

10.1021/ac101338y
Published on Web 07/20/2010

Another approach to handle this limitation is automatic relevance determination,[32,33] which was developed as a feature selection procedure in SVM models. Even though this method selects the variables that are important for the model, the relation between the variables and, for example, the class separation is not visualized. Generally, researchers report the high performance obtained by using a SVM model for classification and regression estimation, but do not comment on the relationship between the input variables and the modeled output data. SVM is often used as a black box approach.

In this paper we present an innovative approach to open this black box for the SVM classifier and give insight in the transformation by the kernel function to make the SVM model more transparent. This approach is based on the nonlinear biplot principles described by Gower and Harding in 1988[34] and is used to visualize and determine the importance and influence of the input variables to the final SVM classifier. The resulting information can then be used to reduce the number of input variables to improve the performance or to reduce the complexity of the model. Furthermore, the visualization of the contribution of the original variables can assist in a better understanding of the underlying chemical or biological process. We will present the proposed approach, and illustrate and validate the methodology by applying it for classification problems: two benchmark data sets and a relevant metabolomics data set obtained from magnetic resonance spectroscopic images to diagnose human brain tumors. The effectiveness of the method is verified by a comparison of the classification performance obtained by using the entire set of input variables and a selection of variables, determined by our approach.

EXPERIMENTAL SECTION

Theory and Computational Strategy. As the theory of SVMs is described extensively in the literature[13,21−24] and the use of a kernel transformation results in the loss of information about the input variables, we will focus in the next section briefly on the concepts of the kernel function. Subsequently, we will explain the basic steps of the nonlinear biplot technique, as described by Gower and Harding,[34] to discuss the proposed method to determine the contribution of the input variables in SVM classifications. The full theory and algorithm of the proposed method can be found in the Supporting Information (SI).

Kernel Transformations. In the various kernel-based methods a specific mapping function is used to project the original input data in a higher-dimensional feature space.[24] The data in this new feature space can subsequently be used as a new input for pattern recognition. The advantage of such an approach is that the method can deal with complex nonlinear problems. The typical (two-dimensional) example to illustrate this approach is shown in the inset of Figure 1. This data set (**X**) contains two classes that we would like to separate.

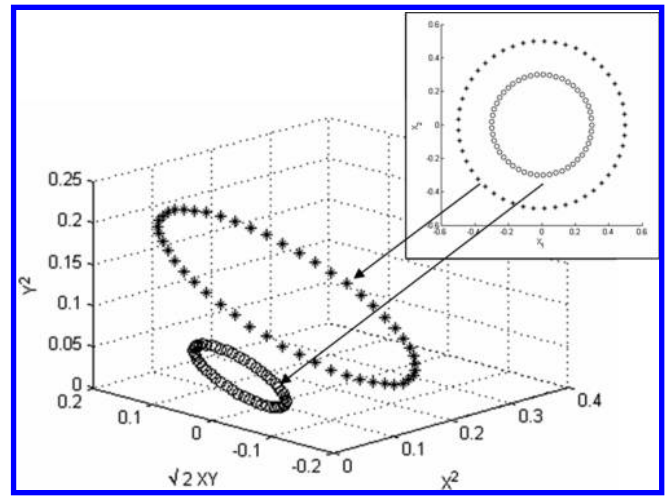

Figure 1. The synthetic benchmark data set. Mapping of a two-dimensional data set which contains two nonlinearly separable classes (outer and inner circles) into a three-dimensional feature space. The transformation is performed by applying the nonlinear mapping function $\phi(\mathbf{x}_{.1}, \mathbf{x}_{.2}) = (\mathbf{x}_{.1}^2, \sqrt{2}\mathbf{x}_{.1}\mathbf{x}_{.2}, \mathbf{x}_{.2}^2)$, which makes the two classes linearly separable.

The inner circle represents one class and the outer circle represents the other class. From the figure it is obvious that we are not able to separate the classes by a linear model. However, if we project this data in a higher dimensional space by using a mapping function, we are able to obtain the three-dimensional feature space which is represented in Figure 1. In this feature space the two classes can be linearly separated by a plane between the circles. In this case the transformation is performed by the mapping function ϕ. It has been shown that, since the (nonlinear) mapping function is in general unknown beforehand and is difficult to determine, the feature space can be constructed implicitly by invoking a generic kernel function (see, e.g., refs 23−25, 35). Such a kernel function is a function (K) which operates on two vectors, such that

$$K(\mathbf{x}_{i.}, \mathbf{x}_{j.}) = \langle \varphi(\mathbf{x}_{i.}), \varphi(\mathbf{x}_{j.}) \rangle \quad (1)$$

where $\mathbf{x}_{i.}$ and $\mathbf{x}_{j.}$ are two objects in the data set and ϕ represents the actual nonlinear mapping function. The use of a kernel function makes it unnecessary to know the actual underlying feature map in order to be able to construct a linear model in the feature space. The application of such a kernel function will result in a square symmetric matrix: the kernel matrix K. This matrix is a weighted dissimilarity matrix, of which each position represents a dissimilarity (distance) measure between two objects. The specific kernel function and the optimal parameter settings of the function are determined in combination with the applied classification algorithm (e.g., SVM) by optimizing the classification performance. However, because the kernel function transforms the original input space into a feature space with a higher dimension, information about the original variables is not preserved.

Nonlinear Biplots. To explain the concepts of nonlinear biplots, we will first describe the classical biplots technique which is used extensively in (chemometric) data analysis.

(30) Üstün, B.; Melssen, W. J.; Buydens, L. M. C. *Anal. Chim. Acta* **2007**, *595*, 299–309.
(31) Devos, O.; Ruckebusch, C.; Durand, A.; Duponchel, L.; Huvenne, J. P. *Chemom. Intell. Lab. Syst.* **2009**, *96*, 27–33.
(32) Van Gestel, T.; Suykens, J. A. K.; De Moor, B.; Vandewalle, J. *Proc. Eur. Symp. Artif. Neural Networks.* **2001**, 13–18.
(33) MacKay, D. J. C. In *Neural Networks and Machine Learning*, NATO Asi Series. Series F, Computer and Systems Sciences 168; Bishop, C. M., Ed.; Springer: Berlin, 1998; pp 133−165.
(34) Gower, J. C.; Harding, S. A. *Biometrika* **1988**, *75*, 445–455.
(35) Gunn, S. R. *Support Vector Machines for Classification and Regression. Technical Report*; Image Speech and Intelligent Systems Research Group, University of Southampton: Southampton, 1997.

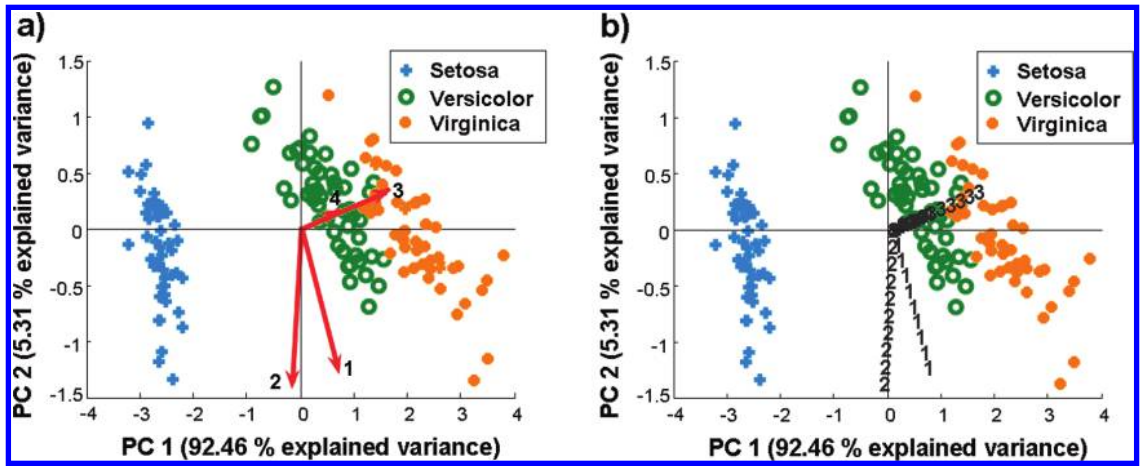

Figure 2. Biplots of the *Iris benchmark* data set. The scores (representing the samples) are visualized as symbols, whereas the loadings are visualized by the vectors. The loadings are obtained by (a) projection of the first two columns in the loading matrix and (b) projection of the scores of 10 pseudosamples (for each variable).

A biplot is a visualization in which the samples and the variables of a data set are represented together. This technique has been introduced by Gabriel in 1971[36] and is based on singular value decomposition (SVD) or principal component analysis (PCA)[37] of the column mean-centered data matrix **X**, containing n samples and m variables. By SVD **X** is decomposed into scores and loadings according to

$$\mathbf{X}_{(n\times m)} = \mathbf{U}_{(n\times r)}\Lambda_{(r\times r)}\mathbf{V}^T_{(m\times r)} = \mathbf{S}_{(n\times r)}\mathbf{L}^T_{(m\times r)} \quad (2)$$

The positions of the samples (rows of **X**) in a two-dimensional biplot are subsequently given by the elements in the columns of **UΛ**, often called the scores **S**. The variables (columns of **X**) are usually represented as vectors pointing from the origin to the coordinates that are given by the elements of the columns of **V**, called the loadings **L**. To construct a biplot for the first two singular vectors (or principal components, PCs), the elements in the first two columns of **S** and in the first two columns of **L** are used. This is illustrated for the *Iris* data set (see the Iris benchmark Data Set section) in Figure 2a. The coordinates of a new sample in this biplot can be obtained by premultiplying the sample as a row vector with **V**:

$$\mathbf{x}_{(1\times m)}\mathbf{V}_{(m\times 2)} = \mathbf{s}_{(1\times 2)} \quad (3)$$

The key idea leading to the nonlinear biplot is the interpretation of the loadings in the classical biplot (the representation of the variables) as projections of special so-called pseudosamples that carry all their weight in one variable. This means that these pseudosamples have a value of 0 for all variables, except for one variable. Projecting this pseudosample in the two-dimensional PCA plot yields coordinates that are equal to the loadings of the variable whose weight it carries. For example, [1, 0, 0, ..., 0] is a $(1 \times m)$ pseudosample with a value 1 for variable 1 and a value 0 for all other variables. Then

$$[1, 0, 0, ..., 0]_{(1\times m)}\mathbf{V}_{(m\times 2)} = \mathbf{s}_{(1\times 2)} = \mathbf{v}_{(1\times 2)} \quad (4)$$

where **v** is the first row of **V** and contains exactly the loading of variable 1, defining its position in the biplot. If the value of 1 is replaced by p different values z, a total of p pseudosamples are obtained. The projection of these pseudosamples in the PCA plot will result in a trajectory along the direction of the variable vector, as illustrated in Figure 2b.

Gower et al.[34] extend this idea to the visualization of variable information in principle coordinate analysis. In principal coordinate analysis (or classical metrics scaling) an SVD is performed not on the original rectangular data matrix **X**, but on the symmetric matrix of squared Euclidean distances. This approach allows the visualization of the relative distances of the objects. It is often remarked that the information of the original variables (such as in a biplot) is lost. Gower et al., however, showed that this can be overcome by using the concept of trajectories of pseudosamples, as explained earlier. To make this approach feasible, one must be able to calculate the squared Euclidean distances of the pseudosamples with all other samples (e.g., data **X**). By projecting the rows of the resulting distance matrix in the principal component space, the trajectories that represent the variables are obtained. Gower et al. showed that this concept can be extended from Euclidean distances to many nonlinear distance metrics, as long as the same distance metric can be used to calculate the distances of the pseudosamples to the original data samples. In the classical linear biplot, only one pseudosample per variable is sufficient to represent the variables. For a nonlinear distance metric, the trajectory will be curved and multiple pseudosamples per variable are required.

Nonlinear Biplots in SVM Classification. We propose to apply the above approach to the kernel-based SVM method, since the kernel is a (nonlinear) distance metric between objects. This procedure comprises the following steps (shortened):

0. Optimize the kernel function and parameter settings for the SVM classification used (resulting in the kernel function $K(\mathbf{x}_i,\mathbf{x}_j)$).

1. From the data **X** calculate the kernel matrix **K**, by using the optimized kernel function. This matrix is subsequently centered, resulting in matrix $\mathbf{K}^c$ of size $n \times n$.

2. Apply SVD on K^c and construct different score plots (for the various combinations of principal components) for the n samples in K^c. Inspect these score plots to find the direction(s)

(36) Gabriel, K. R. *Biometrika.* **1971**, *58*, 453–467.
(37) Massart, D. L. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier Science Publishers: Amsterdam, 1997.

(principal components, PCs) in which maximum class separation is obtained. Note that this is not necessarily along the first PCs.

3. Construct a matrix $\mathbf{P}_j$ for the j^{th} variable which contains p pseudosamples. The range of these pseudosample values z_j should vary between the minimum and maximum value of the original variable.

4. Apply the kernel function $K(p_{i.},x_{j.})$ to the pseudosample data $\mathbf{P}_j$ to obtain the kernel matrix of the pseudosamples $\mathbf{C}$ (i.e., calculate the kernel distances of the pseudo samples to the original samples).

5. Project the rows of the centered pseudosample kernel matrix $\mathbf{C}^c$ in the score plot that was found in step 2.

6. Repeat steps 3−5 for each variable j.

The resulting plot contains a trajectory of pseudosamples for each original variable. These trajectories yield information about the relative contribution of the variables to the SVM classification. The curvature of a trajectory indicates a nonlinear kernel transformation.

SVM Classification and Validation Procedure. Each data set was analyzed by SVM using the RBF kernel. The RBF kernel was chosen because of its simplicity for optimization. Because the SVM application that we have used is a binary classifier, each different class in a data set that contains more than two classes was analyzed by a one-against-all approach.[38] The kernel parameters parameter σ and the SVM parameter C (C is a cost parameter[23]) were optimized by leave-10%-out cross-validation.

All calculations are performed in the software program Matlab (The MathWorks Inc.) version 6.5 release 13. The nonlinear biplot approach was implemented in Matlab by using the commercially available PLS toolbox from eigenvector Research Inc.

Data Sets. To illustrate the applicability of the proposed method we have used several benchmark data sets (synthetic and real) and a metabolomics data set based on Magnetic Resonance Spectroscopic Imaging (MRSI) data.

Synthetic Benchmark Data Set. A synthetic data set was constructed based on the data presented in Figure 1, which is generally used to illustrate the power of SVM classification. The data set contains two variables to represent an inner and outer circle, both consisting of 50 objects. The two circles can not be separated in a linear way and therefore a kernel-based classification method is required to separate the data.

To construct the *Synthetic* data set we added five variables containing random noise (normally distributed) to the data. The variance of these five variables was set to be about 10% of the variance of the two variables which represent the circles. Because noise contains no information, the added variables do not contribute to the classification performance. This data set was constructed to confirm that the first two variables (representing the circles) are only (equally) important in the SVM classification.

Iris Data Set. A widely used benchmark data set to exemplify discriminant and cluster analysis is the data published by Fisher in 1936.[39] This *Iris* data set contains fifty specimens of each of the three species *Iris Setosa*, *Iris Versicolor*, and *Iris Virginica*, resulting in a total of 150 samples. Four properties of the species are measured to determine differences between the three classes. These properties (representing four variables) are *Sepal Length*, *Sepal Width*, *Petal Length*, and *Petal Width*, all measured in millimeters. The *Setosa* class can easily be separated from the other two classes using only one variable (either *Petal Length* or *Petal Width*). The two other classes are partly overlapping, as shown in Figure 2.

We will use the *Iris* data set to demonstrate the applicability of our proposed method to determine the most discriminative variable for the three species.

Metabolomics Data Set. To illustrate the proposed method for variable selection on a more complex data set, we have used a metabolomics data set which consists of magnetic resonance spectroscopic (MRS) spectra obtained from MRSI data.[40] This data set was constructed during a European project called *Interpret*, which was funded by the European Commission to develop new methodologies to automatically classify tumors in the human brain (see http://azizu.uab.es/INTERPRET). Data from a total of 24 patients and 4 volunteers were acquired by MRS at different positions in the brain, according to an acquisition protocol defined by the *Interpret* Consortium. The study was approved by the ethical committee and followed the rules of the World Health Organization. After reaching consensus about the histopathology, three tumor types were identified according to the World Health Organization classification system. These three classes contained glial tumors with different grades: *Grade II* (10 cases), *Grade III* (4 cases), and *Grade IV* (7 cases). A fourth class consists of spectra acquired from patients with *Meningioma* (3 cases). Additionally, a class consisting of *Healthy* tissues was created from patient (4 cases) and volunteer (4 cases) data. For each predefined class a selection of spectra from the different patients was made. Only spectra acquired at regions which clearly consisted of tissue belonging to the particular class were selected. The data for the *Healthy* class was selected from the volunteers or from the contralateral brain region of the patients.[41] The resulting data set contains 569 spectra, consisting of five different classes. Each spectrum contains 229 data points, covering the chemical shifts between 4.0 and 0.5 ppm. Details about the acquisition parameters and preprocessing of the data are described in Simonetti et al[42] and are beyond the scope of this paper.

RESULTS

Synthetic Benchmark Data Set. Classification of the *Synthetic* data set by using SVMs resulted in a leave-10%-out cross-validated accuracy of 100%. This accuracy (see also SI Table 1) illustrates that SVMs are able to separate the two circles. As noise contains no information, the particular noise-variables should not have contributed to the class separation. To verify the contribution of each variable in the final classification first we have to find and visualize the optimal (linear) separation between the classes in the resulting kernel matrix. Because the *Synthetic* data set contains one hundred objects, the feature space in which the original data is mapped (i.e., the space of the kernel matrix) consists as a consequence of one hundred dimensions. Therefore PCA is used to reduce the dimensionality of the feature space. Analysis of the score plots of combinations of only the first five

(38) Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: Singapore, 1999.
(39) Fisher, R. A. *Annu. Eugen.* **1936**, *7*, 179–188.
(40) Barker, P. B.; Lin, D. D. M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2006**, *49*, 99–128.
(41) Simonetti, A. W.; et al. *Anal. Chem.* **2003**, *75*, 5352–5361.
(42) Simonetti, A. W.; et al. *NMR Biomed.* **2005**, *18*, 34–43.

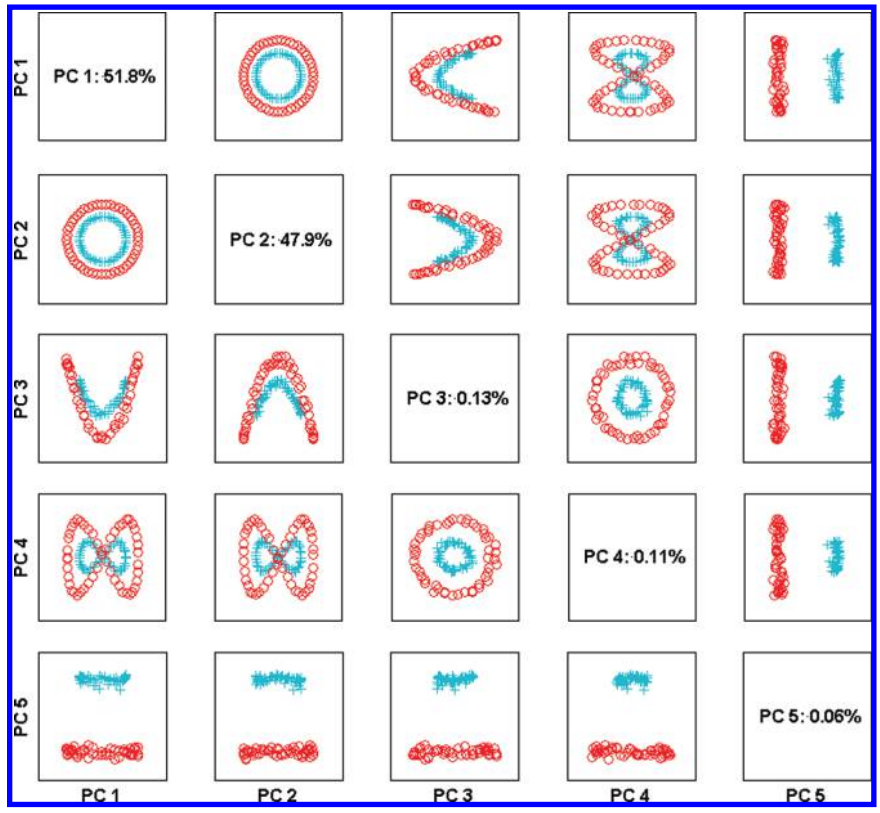

Figure 3. Principle component score plots of the synthetic benchmark data set. Pairs-plot of the first five principal components (PCs) are shown. The separation between the inner circle (blue +-symbol) and the outer circle (red o-symbol) is obtained on PC 5.

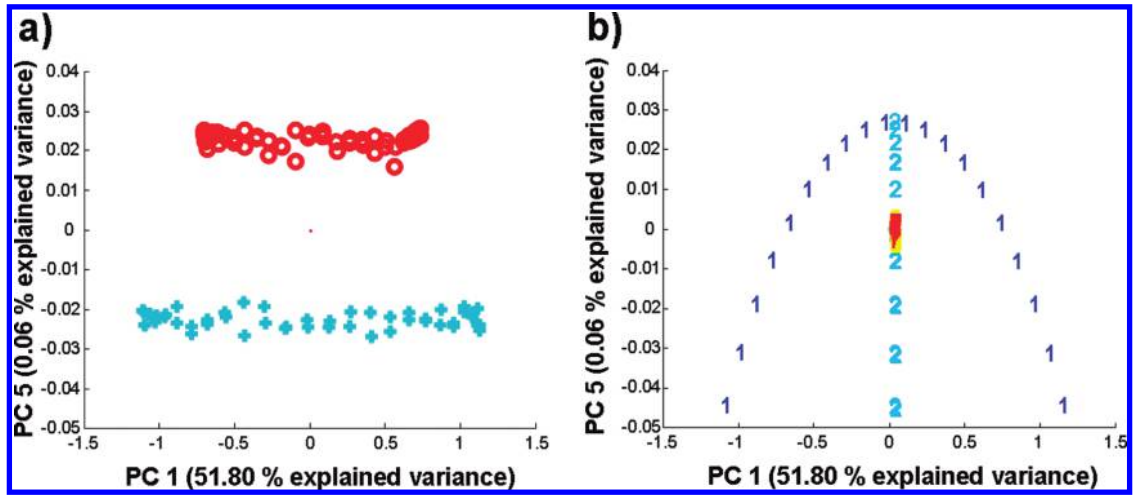

Figure 4. Relevant score plot and pseudo sample trajectories for the *Synthetic* benchmark data set. (a) Projection of the objects of the *Synthetic benchmark* data set (after the kernel transformation) in the feature space spanned by PC 1 and PC 5. (b) Pseudosample trajectories projected in the same feature space. The trajectories of the two variables representing the two circles are indicated by the numbers "1" and "2". The five noise-variables (variables "3" to "7") are distributed around the origin of the plot and are therefore not clearly visible.

PCs show that "PC 5" can be used to visualize the class separation of the two classes. The pairs-plot for the first five PCs of the *Synthetic* data set is given in Figure 3. Apparently, the variance which accounts for the separation between the classes is captured by PC 5. The contribution of each original variable can be determined by projecting the kernel matrix of the corresponding pseudosamples in the feature space spanned by PC 5 and any of the other PCs and by subsequently analyzing the obtained trajectories.

The pseudosample matrices were constructed by varying the intensities of the individual variables over twenty objects (uniformly distributed). The application of the kernel function resulted in seven kernel matrices (one for each variable) of size (20×100). These matrices were then projected in the feature space obtained by PCA on the original kernel matrix $\mathbf{K}$. The trajectories of the seven variables obtained by the proposed method are visualized for the space spanned by "PC 1" and "PC 5" in Figure 4b. As shown, the trajectories of the five noise-variables (variables 3−7) are along the direction of the class separation (PC 5), but are relatively small compared to the two trajectories of the variables representing the inner and outer circle (variables 1 and 2). This indicates that the noise-variables have a small contribution to the

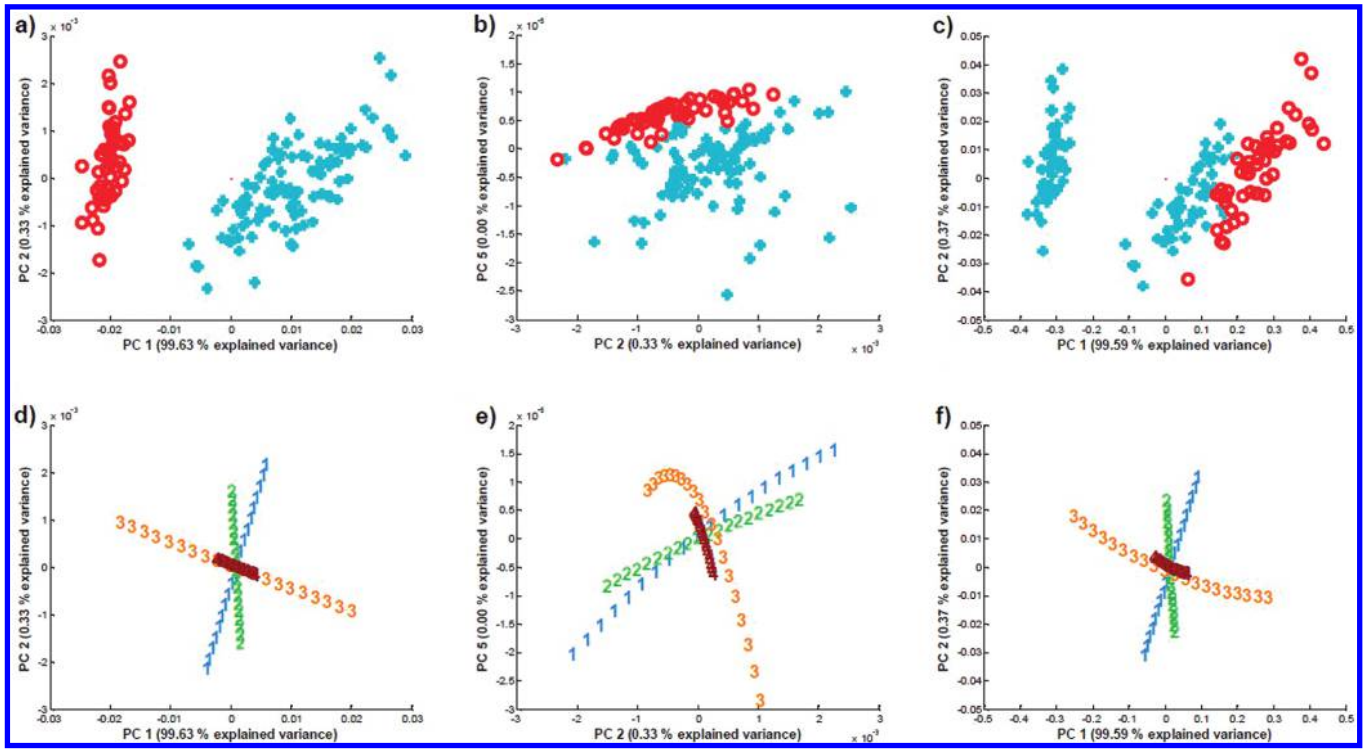

Figure 5. Relevant score plots and pseudo sample trajectories for the *Iris* benchmark data set. Projection of the objects of the *Iris* data set after kernel transformation and PCA, to illustrate the optimal separation of the (a) *Setosa*, (b) *Versicolor*, and the (c) *Virginica* class (red o-symbols). The pseudosample trajectories are projected in the same feature space for the (d) *Setosa*, (e) *Versicolor*, and the (f) *Virginica* class. The trajectory of each variable is indicated by a different number: "1" for *Sepal Length*, "2" for *Sepal Width*, "3" for *Petal Length*, and "4" for *Petal Width*.

class separation, which is in accordance with our hypothesis. As the trajectories of variable 1 and 2 have a similar length on PC 5, both variables are important for the class separation. A (linear) classification based on only one variable is therefore not possible. A visual inspection of the *Synthetic* data set (see the Theory and Computational Strategy section) already confirmed this conclusion.

Inspection of the trajectories in the score plots of combinations of other PCs showed that the variables representing the two circles can have a relatively small loading in the particular feature space (results not shown). For example, in the space spanned by "PC 2" and "PC 4" variable 4 (representing noise) has the largest loading compared to the other variables. However, no class separation was found in this particular score-plot (see Figure 3) and therefore this feature space is not informative.

Iris Benchmark Data Set. The application of the three possible one-against-all classifications resulted in cross-validated accuracies of at least 96.7% (see also SI Table 1). After PCA was applied to the three kernel matrices (for each one-against-all classifier), we searched for the PCs resulting in the optimal separation between the classes in the pairs-plots. These optimal separations are visualized in Figure 5a−c and as shown the *Setosa* class is completely separated from the other classes by the variance captured by PC 1. The *Versicolor* and *Virginica* class requires two principal components to capture the variance which accounts for the class separation, that is, PC 2 and PC 5 for *Versicolor* and PC 1 and PC 2 for the *Virginica* class.

To determine the contribution of the original variables, pseudosample trajectories were constructed and projected in the space spanned by the corresponding PCs. These projections are visualized in Figure 5d−f and as illustrated the variables *Petal Length* (variable 3) and *Petal Width* (variable 4) are most discriminative for the classifications, confirming our hypothesis. However, the length of the trajectory of *Petal Width* is much shorter compared to *Petal Length* (for unscaled data) and is therefore less important for the classifications. To confirm these observations we have applied SVM to the *Iris* data set by including only *Petal Length*. The resulting cross-validated accuracies are given in SI Table 2. Although the accuracies for the *Versicolor* and *Virginica* class are somewhat lower compared to the results where all the variables are used (95.3% versus 96.7%), the accuracies are still comparable. If one of the other variables was chosen for the classification, that is, *Sepal Width*, the accuracies are much lower (<84%, results not shown), indicating that *Petal Length* is the most important variable in the classification.

Metabolomics MRSI Brain Tumor Data Set. The *MRSI* data set was used to study the proposed method for data sets with many variables (229). For illustration purposes we only consider the classification of the *Healthy* class against all the tumor classes. The application of SVM resulted in an accuracy of 98.7% as shown in Table 1 of the SI. By the application of PCA on the kernel matrix, the variance between the *Healthy* class and the other classes is captured by PC 1 and PC 3. The projection of the samples in the space spanned by PC 1 and PC 3 is presented in Figure 6a. If the trajectory of each pseudosample (representing the individual variables) is projected in the same space, many trajectories are located around the origin as shown in Figure 6b. Only two groups of trajectories show a clear elongated pattern.

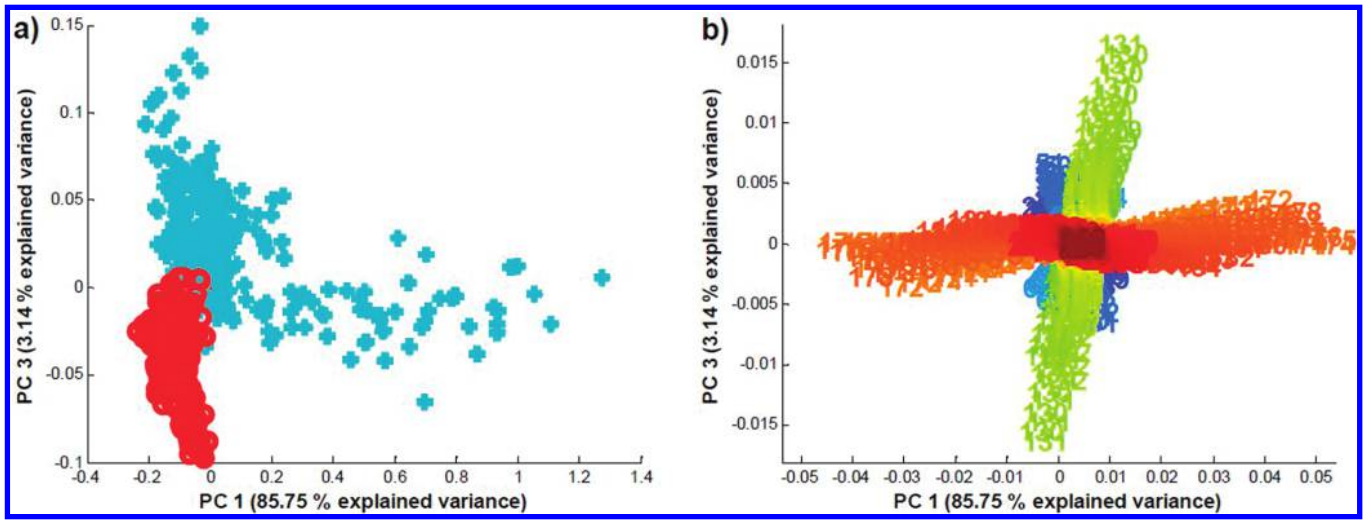

Figure 6. Relevant score plot and pseudo sample trajectories for the *metabolomics* data set. (a) Space spanned by PC 1 and PC 3, obtained by PCA applied on the kernel matrix of the *MRSI* data set. The *Healthy* class is indicated by red o-symbols and the tumor classes by blue +-symbols. (b) Pseudosample trajectories, projected in the same feature space as in (a). Each trajectory is indicated by a different number and represents a different variable. The trajectories are also color-coded using the variable numbers, resulting in similar colors for trajectories of variables representing specific regions in the NMR spectrum.

One group of trajectories consists of pseudosamples representing the variable numbers 170−184 (1.40−1.19 ppm; the trajectories in the direction of PC1, colored orange in Figure 6b), and the other group the variables 129−132 (2.03−1.98 ppm; the trajectories more or less in the direction of PC3, colored green in Figure 6b). Because the group of variables within 2.03−1.98 ppm is along the direction of the class separation we postulate that these variables have a large contribution to the SVM classification. This observation corresponds to results published in the literature, in which the researchers stated that this specific region corresponds to *N*-acetyl-aspartate, which is a neuronal marker for viable neurons, and that the concentration is reduced or absent in most brain tumors.[43]

Note that the other group of trajectories, representing the variables 170−184 (1.40−1.19 ppm), corresponds to lactate and lipids regions of the spectra. These trajectories are located in the direction along PC 1 and direction corresponds to increasing tumor grade. This observation also corresponds to the results of Howe et al.[43]

With only the variables selected within the region of 2.03−1.98 ppm (four variables), the SVM classification results in a leave-10%-out cross-validated accuracy of 95.4% (see SI Table 2). This value is in agreement with the accuracy obtained when all the variables are included in the SVM model (98.7%), indicating that these variables have a large contribution to the classification. If a set of four variables was selected randomly, the cross-validated accuracy was <76%. Classification of the other classes results in similar conclusions. This indicates that the proposed method can be used to determine the variables which contribute to the optimal class separation in kernel-based methods.

DISCUSSION

The application of the method to the different classification problems shows that the proposed procedure can be used to visualize and determine the relative contribution of the original variables in SVM classifications. From the examples with classes that are not linearly separable, the curvatures of the trajectories illustrate the effect and importance of the nonlinear kernel transformation. This can be concluded especially for the *Synthetic* and *Iris* benchmark data sets, in which several trajectories are curved. Such a determination of the relative contribution of the original variables and the effect of the kernel transformation can also be applied to kernel PLS in regression problems. Another possible extension is the application the method to multiclass kernel classifiers. Even though promising results for the kernel PLS case are already obtained, these approaches are still subject to further research.

The metabolites identified and visualized using the proposed method on the metabolomics data set were also identified applying the ARD method on this data set.[44] The advantage of our method, however, is the visualization and identification of (non-) linearity of the contribution. Moreover, our method is computationally simple.

The results of the data sets illustrate that the optimal class separations are captured by any or a combination of the first five principal components. However, a combination of several principal components has still to be (visually) analyzed. Another approach is to apply pattern recognition techniques[45,46] to the feature space in order to find the direction which described the class separation. This approach is also still under investigation.

The proposed method could possibly fail if no class separation can be found after inspection of the score plots with different combinations of principal components. This is, however, not expected, as the SVM algorithm searches for a kernel function that provides such a linear class separation.

(43) Howe, F. A.; Opstad, K. S. *NMR Biomed.* **2003**, *16* (3), 123–131.

(44) Postma, G. J.; Luts, J.; Idema, A. J.; Julià-Sapé, M.; Moreno-Torrese, Á.; Gajewicz, W.; Suykens, J. A. K.; Heerschap, A.; Van Huffel, S.; Buydens, L. M. C. *Comp. Biol. Med.* (submitted).

(45) Jain, A. K.; Murty, M. N.; Flynn, P. J. *ACM Comput. Surv.* **1999**, *31*, 264–323.

(46) Webb, A. *Statistical Pattern Recognition*; Wiley: Malvern, 2002.

CONCLUSIONS

In this paper we have introduced a new method to successfully visualize and identify the important variables in the classification by a kernel-based method. By constructing a set of pseudosamples we are able to determine the effect of the kernel function to the individual variables. We have applied the proposed method to several synthetic and real data sets and have shown that our method is able to find and visualize the most discriminative variables. To confirm this conclusion, we have selected only the most discriminative variables and reapplied the proposed method. The cross-validated classification accuracies of these reduced data sets are similar to the accuracies obtained by using the data sets with the full set of variables. This illustrates the validity of our method to determine the most important variables for classification purposes by kernel-based techniques. The interpretation of SVM models with our proposed visualization method has the potential to extract relevant chemical and biological knowledge from complex data, such as omics data. This will greatly enhance the applicability of the powerful SVM classifiers.

SUPPORTING INFORMATION AVAILABLE

List of symbols, discussion of theory and computational strategy, two additional figures, and two additional tables. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review May 26, 2010. Accepted July 7, 2010.

AC101338Y