About Analysis and Robust Classification of Searchlight fMRI-Data Using Machine Learning Classifiers

M. Lange, M. Kästner and T. Villmann

Abstract—In the present paper we investigate the analysis of functional magnetic resonance image (fMRI) data based on voxel response analysis. All voxels in local spatial area (volume) of a considered voxel form its so-called searchlight. The searchlight for a presented task is taken as a complex pattern. Task dependent discriminant analysis of voxel is then performed by assessment of the discrimination behavior of the respective searchlight pattern for a given task. Classification analysis of these patterns is usually done using linear support vector machines (linSVMs) as a machine learning approach or another statistical classifier like linear discriminant classifier. The test classification accuracy determining the task sensitivity is interpreted as the discrimination ability of the related voxel. However, frequently, the number of voxels contributing to a searchlight is much larger than the number of available pattern samples in classification learning, i.e. the dimensionality of patterns is higher than the number of samples. Therefore, the respective underlying mathematical classification problem has not an unique solution such that a certain solution obtained by the machine learning classifier contains arbitrary (random) components. For this situation, the generalization ability of the classifier may drop down. We propose in this paper another data processing approach to reduce this problem. In particular, we reformulate the classification problem within the searchlight. Doing so, we avoid the dimensionality problem: We obtain a mathematically well-defined classification problem, such that generalization ability of a trained classifier is kept high. Hence, a better stability of the task discrimination is obtained. Additionally, we propose the utilization of generalized learning vector quantizers as an alternative machine learning classifier system compared to SVMs, to improve further the stability of the classifier model due to decreased model complexity.

I. INTRODUCTION

The analysis of functional magnetic resonance images (fMRI) is still a challenging topic. In particular, investigations about task dependent activities became into the focus of research during the last years [31], [21], [33], [34], [54]. Multivariate voxel analysis based on the underlying fMRI time series (BOLD-responses\textsuperscript{1}) is a popular mathematical tool for investigation. A single value is calculated from the BOLD-signal taken as voxel response. Depending on the given tasks, the behavior of each voxel is analyzed. The respective discrimination ability of a voxel is assessed according to its task discrimination behavior for a given problem, i.e. the test accuracy is considered. The results are then mapped onto the locations of the voxels in the brain forming all together a so-called accuracy map related to the considered tasks [31], [54], [40]. Accuracy maps are considered as essential tools for local brain discrimination ability detection.

Yet, the response of a voxel, typically is influenced by noise, such that an assessment of the discrimination ability needs further information for reliable results. For this purpose, the so-called searchlight method was suggested [31], [21]: For a voxel to be considered, a surrounding spatial volume of voxels inside a certain diameter is taken as a local field denoted as searchlight assigned to that voxel. The voxel itself is denoted as central voxel. The searchlight should reflect the classification behavior of the central voxel [31]. One underlying assumption is that BOLD-responses in a small volume should show correlated response behavior\textsuperscript{2}. More precisely, the searchlight forms a high-dimensional pattern, which is used for classification analysis [36], [59]. Hence, a searchlight classification analysis, taking into account these correlation information, decreases the influence of the noise compared to a single voxel classification analysis.

Yet, frequently the number of samples of an experiment is less than the number of voxels within the searchlight area (volume) [5], [15], [23], [24]. As we will explain later in this article, this causes mathematical problems in searchlight classification analysis (SCA). In particular, the classification behavior of a statistical or machine learning classifier is affected when it is applied to such problems: the generalization ability is decreased and, hence, the SCA results are difficult to interpret.

One standard machine learning classifier frequently applied in SCA is the (linear) support vector machine (SVM) [47]. The application of SVMs is far from being trivial although many software tools are available. In particular, taking only the obtained classification accuracy is not the full truth for a correct assessment. An important additional information is the number of so-called support vectors compared to the number of training samples, which determines the generalization ability of the trained classifier.

We emphasize at this point the following: ‘Correlation’ means that the voxels contributing to a searchlight contain related information such that the searchlight forms an informative pattern. Hence, correlations could have positive and negative signs as well as may take place on different orders.

\textsuperscript{1}BOLD-response - blood oxygenation level-dependent-response, see [6], [32].

\textsuperscript{2}We emphasize at this point the following: 'Correlation' means that the voxels contributing to a searchlight contain related information such that the searchlight forms an informative pattern. Hence, correlations could have positive and negative signs as well as may take place on different orders.
model complexity and influences the generalization ability of the trained SVM classifier model.

In this contribution we consider a typical experimental setting in fMRI analysis of only a few data sample. We investigate for this scenario the common data processing and SCA. We show that this procedure can be misleading and, hence, may provide delusive results. In consequence, we propose a different data processing that lessen these problems. Instead of investigation of searchlight patterns, we propose to consider response patterns of the voxels inside the searchlight to the given tasks. This processing includes a different utilization of the data matrix provided by the fMRI searchlights as well as we suggest an alternative machine learning classifier.

The paper is structured as follow: First, we shortly review standard data processing in the context of the searchlight fMRI data analysis. In particular, we will focus on the underlying mathematical assumptions and consider them in the few-sample-scenario. Thereafter, we reformulate the classification task for this problem to reduce the difficulties.

II. Typical fMRI Data Processing in the Context of the Searchlight Method

A. fMRI Data Dimension Reduction

General fMRI analysis is done for large brain volumes or the whole brain. Thereby, the considered volume is discretized into voxels. Modern scanners have a resolution of up to $1mm^3$ per voxel. This results in a huge number $N$ of voxel data for an experiment. Collecting all voxel responses in a large $N$-dimensional data vector, as frequently done in general multivariate pattern analysis (MVPA, [59]), causes numerical difficulties. Therefore, feature reduction methods are frequently applied to reduce the complexity including principal component analysis (PCA), independent component analysis or information based selection schemes to name just a few [9], [14], [35], [38], [39], [52], [58]. The (spherical) searchlight approach in fMRI analysis takes a local view to reduce the number of considered voxels [31]. In particular, it selects a spherical spatial area of voxels surrounding the central voxel and, the searchlight diameter $\delta_s$ is typically in a range of a few voxels [54]. The method assumes that the fMRI responses of the voxels are this spatial region form an informative pattern, such that their contained information contributes to noise reduction, when task dependent classification analysis is applied for the considered central voxel.

Let $N_s$ be the number of voxels belonging to a searchlight area including the central voxel with spatial coordinates $r = (r_x, r_y, r_z)$. The standard procedure to extract the response of a voxel for a given experiment with several experimental conditions (tasks) is well-known and can be found in [5], [9], [15], [36]. Here, we denote by $v_k$ the response of the $k$th voxel in the searchlight. The central voxel has the index $k = 1$. The overall response of a searchlight for a given experiment sample $l$ is fused in the vector $v_r(l) \in \mathbb{R}^{N_s}$ for the central voxel located at position $r$. We denote this vector as searchlight response for a given experiment. Let $K$ be the number of experiments (trials) for a given task and $C$ be the number of experimental conditions. Then, all response data of a searchlight are collected in the data matrix $D_r \in \mathbb{R}^{M \times N_s}$ as

$$D_r = \begin{pmatrix}
v_r(1) \\
v_r(l) \\
v_r(M)
\end{pmatrix}$$

with $M = C \cdot K$, see Fig. 1a). We denote the matrix $D_r$ as searchlight matrix. We remark at this point that for the rank of the matrix $D_r$ always the relation

$$\text{rank}(D_r) \leq \min(M, N_s)$$

is valid. Without loss of generality, we suppose that the data assigned to a certain experimental condition $c$ with $1 \leq c \leq C$ are stored in the sub-matrix $D_{rc} \in \mathbb{R}^{K \times N_s}$, whereby we dropped here the index $r$ for better readability. Thus, $D_r$ is
structured as

\[ D_r = \begin{pmatrix} D_1 \\ \vdots \\ D_N \end{pmatrix} \]

and contains all response information for each experimental condition and trial.

B. Mathematical Investigation of the Searchlight Data Matrix in Context of Classification Tasks for Information Mapping

Information mapping based on the searchlight method is a popular method to visualize the discrimination ability of a voxel for given tasks [30], [31]. For this purpose, each voxel, located at position \( r \), is investigated as a central voxel with searchlight matrix \( D_r \). Its discrimination ability is determined as the estimated test classification accuracy of a classifier trained on \( D_r \), reflecting the overall generalization ability of the voxels contained in the searchlight. The row vectors \( \mathbf{v}_r(l) \) are taken as samples for classifier training.

Several studies have shown that linear classifiers deliver sufficient precise results in SCA [15], [20], [40], [36]. Besides standard linear discriminant classifier (LDC) [43], machine learning classifiers like k-nearest-neighbors, decision trees or support vector machines are frequently used [3], [12]. SVMs implicitly map the data into a feature kernel mapping space for better class separation using kernel properties [1], [37], [53]. Among them linear support vector machines (linSVMs) play a key role in SCA [15], [30], [35], [41].

1) Comments on Classifiers: Although many software tools for MVPA are available, the application of classifier methods in the context of the SCA may be crucial and can lead to misleading results and interpretations if they are applied without further, detailed considerations: In many studies, only two experimental conditions (two-class-problem) are compared such that \( C = 2 \) [35], [36], [41], [54], [40]. For this case, the discriminant function of LDC has \( 2 \times N_s \) parameters to be determined. In this situation, the searchlight matrix \( D_r \) should have a rank \( \text{rank}(D_r) \geq 2N_s \). Otherwise, the system has too many degrees of freedom and the generalization ability of the LDC becomes worse. Similarly, any MVPA, as frequently applied in searchlight analysis [4], [15], [59], would be affected by this low rank problem.

For SVMs, the binary case is most convenient, because SVMs were designed originally for this scenario [47]. The separating hyperplane \( \mathbf{h} \), implicitly calculated by the SVM, is a \( N_s \)-dimensional vector with only \( (N_s - 1) \) free parameters due to the normalization condition \( || \mathbf{h} || = 1 \) valid for SVMs. For cases \( C > 2 \) heuristics have to be applied, for example successive one-versus-all considerations [11], [51]. This machine learning procedure is a convex optimization in an appropriate dual space (kernel mapping space) with inherent regularization according to the classification separation margin. The optimization determines so-called support vectors, which define the decision boundaries between the classes and, hence, are not class-typical representatives. Remember at this point that support vectors are always data points [11]. Depending on the used kernel, the underlying implicit mapping is linear or nonlinear realizing a linear or nonlinear classifier. The complexity of the SVM classifier is determined by the number of support vectors \( N_{SV} \leq N_T \) and the dimensionality of response vectors \( \mathbf{v}_r(l) \). Here, \( N_T \) is the number of training samples, which usually is chosen to be \( N_T \leq M \). If the relation \( N_{SV} \leq N_T \) is valid, i.e. the ratio \( \gamma_{SV} = N_{SV}/N_T \) is approximately 100%, the number of support vectors equals approximately the number of training samples. Then, we have to distinguish two cases: if the number of training samples \( N_T \) is less than the data dimensions, which is for \( D_r \) the number of \( N_s \) of voxels in the searchlight, the SVM model memorizes the training data. Otherwise, this conclusion cannot be drawn, because this effect could be also dedicated to the fact that the data are not separable. Hence, in the first case, the generalization ability is reduced, which becomes manifest in a large difference between training and test accuracy [56]. Yet, the number of samples \( N_T \) should be kept as high as possible, such that the classifier is able to extract the knowledge about class separabilities from the statistical properties of the training data.

An alternative to SVMs are prototype based learning vector quantizers (LVQ, [29]). In contrast to support vectors in SVMs, here the prototypes \( \mathbf{w} \in \mathbb{R}^n \), trained during the learning process for each class, are class typical but not necessarily data points [26]. In case of the above searchlight model we have \( n = N_s \). If the number of prototypes \( N_P \), which has to be chosen in advance, is significantly smaller than the number \( N_T \) of training samples, i.e. \( \gamma_{LVQ} = N_P/N_T \ll 1 \) holds, a good generalization ability can be expected [10]. Otherwise, the generalization ability may become worse.

Variants of the basic LVQ schemes, the so-called generalized LVQs (GLVQ, [44]), also can be interpreted as margin optimizers [17]. Yet, here the hypothesis margin is maximized [10], [45], which ensures the robustness of the model with respect to small changes. Variants of LVQ consider adaptive metrics instead of the Euclidean metric for optimal classification performance including quadratic forms

\[ d(x, w) = (x - w)^T \Lambda (x - w) \]

with positive semi-definite but adaptive classification correlation matrices \( \Lambda \) [19], [49]. Generally, LVQ models outperform k-nearest-neighbor approaches. The number of prototypes per class defines the model complexity in LVQ and has to be specified by the user. Hence, the model complexity has to be fixed in advance. Therefore, we can guaranty that the number \( N_P \) of prototypes is always \( N_P < N_T \) to ensure generalization ability. Again, the number of training samples should be kept as high as possible to achieve a good learning
statistics. If less than three prototypes per class are chosen, a linear classifier (with respect to the used metric) is obtained. Kernel variants, adequate to nonlinear SVMs, are recently proposed showing comparable high performance [42], [46], [25], [26], [57].

2) Comments on Mathematical Structure of the Searchlight Data Matrix: Unfortunately, for many investigations in fMRI analysis only a few trials are available. Frequently, the number $K$ of trials is less than 10 down to only $K = 5$ [8], [9], [24], [23], [54]. Yet, the searchlight diameter commonly is chosen to be $3 \leq \delta_s \leq 11$ and should depend on the spatial resolution of the fMRI scanner. The respective numbers of voxels within a searchlight ranges from 19 to 691 [54]. Thus, $\text{rank}(D_s) = M = 10$ according to (2), and, particularly $\text{rank}(D_s) \ll N_s$ is valid (low rank case). While the respective classification problem is undetermined in this situation [13], the responses are completely separable in training equivalent to a training accuracy of 100%. In the low rank case, LDC has $(N_s - \text{rank}(D_s))$ free parameters, which can be chosen arbitrarily without any performance loss on the training set. However, this arbitrariness leads to obscure classification performance for different test data. Of course, this indeterminacy influences each statistical or machine learning classifier to be applied.

One of the key ideas of searchlight is that it is a complex informative pattern of voxel responses. We consider the searchlight matrix $D_s$ ($N_s$) in dependence of the number $N_s$ of voxels inside the searchlight. Obviously, $N_s$ increases with increasing searchlight diameter $\delta_s$. Suppose, we start with a small number $N_s^0 \gtrsim M$. Additionally, we assume a high classification accuracy for $D_s$ ($N_s^0$) for a given classifier. Now, voxels are added successively increasing the diameter. If the added voxels are independent or only weakly correlated to those forming $D_s$ ($N_s^0$), the classification accuracy is lost. This can be dedicated to the irrelevant but contributing data stream, which prevents a classification related knowledge extraction of the classifier during the training. If strong correlations exist with only small added noise, the learning would be only slightly affected and the classification accuracy is kept high after a stabilization phase.

In the following experiment we demonstrate the effect: Suppose a data matrix $M \in \mathbb{R}^{n \times d}$ of $n$ d-dimensional row vectors $m_i$ equally assigned to two classes playing the role of ground truth (informative features), to simulate a small searchlight with high accuracy. The data matrix is taken as training set for a machine learning classifier, which achieves an accuracy $\alpha_0$ for this problem. Now, the matrix $M$ is successively increased adding noisy irrelevant features (dimensions) such that $m_i(j)$ contains $j$ additional components compared to $m_i$. Two cases are of interest: a) the added component is a copy of one of the original components with a small additive noise; b) the new components are weakly correlated or completely independent from the original.

Obviously, in the case a) the new components are strongly correlated to the original and, hence, this is the underlying assumption of the strongly correlated searchlight model. In this case, the classification accuracy of a classifier should be decreased, if only a few components are added. However, adding more and more correlated but slightly noisy dimensions should stabilize the accuracy on the level of the system without noise. Otherwise, in case b) the accuracy should decrease continuously to the value 0.5 (by chance), because complete independence without any information regarding the original problem destroys the original configuration. Simulations for both settings are visualized in Fig. 2 using a linear SVM (LIBSVM, [7]) as the classifier model for $n = 10$ and $d = 5$. As we can see, the simulations emphasize the above discussions.

The addressed problems can be diminished applying additional dimensionality reduction or feature selection schemes before classification learning to limit the number of voxels contributing to the data matrix accordingly. Another solution could be to apply metric adaptive classifiers, which automatically weight out voxels not contributing to class discrimination of a searchlight. These methods are known as relevance or classification-correlation learning (matrix learning) [19], [49]. Principal component analysis also can help to reduce these problems [22], but is independent from the classification task.

In conclusion of these investigations we can state that the choice of a suitable searchlight diameter remains crucial. However, the problem frequently is neglected [5], [8], [23], [24], [40], [54].
C. A New View on the Searchlight Data Matrix for Adequate Classification Analysis

Now, we take another view onto the data matrix $D_r$ from (3). We assume for each task (class) the same number $K$ of trials and consider for each voxel the response vector $x^c_i \in \mathbb{R}^K$ a given task, i.e. for each voxel $i$ and a class $c \in \{1, \ldots, C\}$ the vector $x^c_i$ collects the response information. With other words, the vector $x^c_i$ forms the overall response of a voxel $i$ for task $c$ (response patterns of the voxels). Hence, the matrix $X_c = \mathbb{R}^{C \times N_s \times K}$ of the response vectors of all voxels is $X_c = (D_c)^\top$ and we obtain the new data matrix, denoted as voxel response matrix,

$$X_r = \begin{pmatrix} X_1 \\ \vdots \\ X_r \\ \vdots \\ X_C \end{pmatrix} = \begin{pmatrix} (D_1)^\top \\ \vdots \\ (D_r)^\top \\ \vdots \\ (D_C)^\top \end{pmatrix}$$

for the central voxel at location $r$, see Fig. 1b), which obviously has a rank $\text{rank}(X_r) \leq K$. A label $y_i \in \{1, \ldots, C\}$ is assigned to each row vector $x_i^c \in \mathbb{R}^K$ of $X_r$ indicating its class origin according to $x_i^c$.

For the voxel response matrix $X_r$ the number of training samples $N_T$ is determined by the product $C \cdot N_s$ usually with $C \cdot N_s \gg K$ is valid in SCA. Hence, the above mentioned difficulties regarding the high-dimensional vectors $\mathbf{v}_c(l)$ in the original searchlight data matrix $D_r$ for machine learning classifiers are drastically reduced. With high probability, we can expect that $\text{rank}(X_r) = K$ holds. Therefore, the training behavior for a machine learning classifier becomes more serious in that approach. In consequence, a better generalization ability should be achievable, i.e. the difference between training and test classification accuracy should be decreased in comparison to the original data handling in SCA.

III. Numerical Investigation on an Exemplary Real Data Set

In this chapter we give numerical results for an exemplary but real data set. The fMRI data (gradient EPI) were collected on a Siemens 3 T system (Trio) with a standard head coil. The scans contained 36 axial slices covering the whole brain (TR=2000ms, TE=24ms, slice thickness 4mm with 1mm gap, in plane resolution 3x3mm²). A standard hemodynamic response function model was fitted to the data to estimate the statistical parameters (scaling parameters, beta values) resulting in one β-value for every voxel per experimental condition as the response value [54]. We split the functional data into five parts such that we obtain $K = 5$ trials per task. The number of tasks was $C = 2$. We used the standard spherical searchlight approach with a diameter $\delta_s = 5$ such that each searchlight contains $N_s = 57$ voxels. Using this setting, we have for the rank of a searchlight data matrix $D_r$ the inequality $\text{rank}(D_r) \leq 10 \ll N_s$ according to (2). All numerical experiments were done for three different fMRI data sets.

For comparison with the simulations depicted in Fig. 2, we considered the averaged behavior of a reduced data matrix $D_r$ of a searchlight depending on the number of additional voxels. For this purpose, first we determined for each voxel the searchlight and selected from this randomly 4 voxels. Additionally, the central voxel belongs to reduced matrix, such that $D_r \in \mathbb{R}^{10 \times 5}$ with $C \cdot K = 10$. Hence, the reduced data matrices are accordingly to those start matrices $M$ of the simulations. For each reduced data matrix $D_r$ we calculated the classification accuracy by 5-fold cross validation using linSVM (to be comparable with the above simulations). From all central voxels with a test accuracy greater than 70% we randomly selected 5000 central voxels for the investigations to be done: For each of the selected voxels, we started with the reduced data matrix $D_r$ of the respective searchlight, which plays the role of originals as in the simulations above. Then we added more and more voxel of the considered searchlight to the reduced data matrix $D_r$, again in accordance to the simulations above. Finally, the obtained accuracy curves are averaged over all investigated (selected) 5000 central voxels. The resulting curves for the three probands are depicted in Fig. 3. As we observe, the resulting averaged accuracy curves obtained are similar to that of case b) in the simulations above depicted in Fig. 2. Thus, we can conclude that the correlations in the full searchlight are at least weak, which should cause a large number of support vectors in SVMs in comparison to the number of training samples ($\gamma_{SVM}$ close to 1).

Generally, we should expect from this first investigation a loss in the generalization ability for a trained machine learning classifier in a cross validation setting for overall searchlights. Therefore, we applied a five-fold cross valida-
tion procedure to verify this property applying two classifiers (linSVM and GLVQ). For linSVMs we observed in average (over all three fMRI data sets) $N_{SV} = 7.1$ support vectors compared to exactly $N_T = 8$ training vectors resulting in the ratio $\gamma_{SVM} = N_{SV}/N_T = 0.8875$, while $N_T \ll N_s$ is valid and $N_s$ is the data dimension for the classifier. Hence, a loss of the generalization ability is expected. This loss was estimated calculating the differences between training and test accuracies and obtained in average as $\Delta_{linSVM} = 0.5093$. For GLVQ with only one prototype per class chosen, the averaged loss is $\Delta_{GLVQ} = 0.5581$. The corresponding value $\gamma_{GLVQ}$ is obtained as the ratio $N_P/N_T = 0.25$, also indicating the reduced generalization ability.

The respective averaged test accuracy variance is $\bar{\sigma}_{linSVM} = 0.0947$. Application of GLVQ yields $\bar{\sigma}_{GLVQ} = 0.075$. The reduced model complexity of GLVQ compared to the SVM model gives a reduced variance value indicating a slightly improved stability the GLVQ model.

If we switch to the new voxel response matrix $X_r$ for the same data set instead, the data dimension becomes $K = 5 \ll N_s$. We get the following results, again for a five-fold cross validation: Now, we have in this model $N_T = C \cdot N_s/5 \gg K$ vectors for training during the cross validation procedure, such that the learning of the classifiers becomes more serious. The averaged ratio $\gamma_{SVM}$ for the linSVM is obtained as $N_{SV}/N_T = 0.5440$, which should lead to a better generalization ability. In fact, we achieved a difference value $\Delta_{linSVM} = 0.0612$, which is a magnitude smaller than before. For GLVQ with one prototype per class, i.e. $\gamma_{GLVQ} = N_P/N_T = 0.0351$, we got $\Delta_{GLVQ} = 0.098$, which also is a reduced loss value compared to that obtained for the original data model above.

The respective averaged test accuracy variances are calculated as $\bar{\sigma}_{linSVM} = 0.0092$ and $\bar{\sigma}_{GLVQ} = 0.0080$, i.e. both classifiers show the same behavior decreasing the variance by a magnitude indicating a better stability of the test accuracy.

We conclude that for the new data model an improved generalization ability is obtained. This can be addressed to the better statistical properties of the data matrix $X_r$ in the context of classification learning, which lead to lower ratios $\gamma_{SVM}$ and $\gamma_{GLVQ}$ for SVM and GLVQ, respectively indicating improved generalization abilities. At the same time, a sufficient high number of training samples for good learning statistics is achieved utilizing the new data model.

IV. CONCLUSION

In this article we highlighted some mathematical and numerical difficulties for machine learning classifiers, when applied to determine the discrimination ability of BOLDS in fMRI data analysis and information mapping using the frequently applied searchlight method. In particular, we consider the mathematical properties of the searchlight data matrix and their consequence for the learning and application of machine learning classifiers. One key point is the low rank problem of those matrices compared to the number of voxels in a searchlight and the very small number of training samples. A large discrepancy causes unstable behavior of the classifier, i.e. the generalization ability is drastically reduced. This behavior is independent from the used classifier. However, GLVQ with predetermined low model complexity gives slightly better generalization, which become more significant for greater searchlight diameters. If the new data model is used, the low rank problem for the classifiers is reduced and the generalization ability is improved.

Further, we investigated the classification accuracy of a searchlight depending on the searchlight diameter with respect weak and strong correlations. We compared simulated results with those of real world data. In conclusion of these experiments, the appropriate choice of the diameter crucially influences the classification accuracy. In particular, if the components of the searchlight matrix are only weakly correlated, the classification accuracy drops down.

Yet, the demonstrated difficulties for analysis of searchlight data are not restricted to crisp classification problems. Obviously, other approaches like fuzzy classification and regression schemes are affected in a similar way, because they base on the same algebraic methods in MVPA [2], [16], [15]. Further, similar problems of huge-dimensional data with only a small number of samples occur in in other fields of data mining like gene expression data analysis [18], [55]. Thus, respective conclusions should can be of interest also there [28].

Based on our considerations, we strongly recommend a faithful analysis of the searchlight data and an adequate feature selection. Further, we want to draw the attention to alternatives to SVMs in machine learning classification, namely the family of generalized LVQs. In addition to the above highlighted property of pre-defined model complexity, these models are well-suited for non-standard (non-Euclidean) and adaptive metrics including adaptive quadratic forms [48], [49], [50] as well as kernel distances known from SVMs [18], [27], [26], [57].

ACKNOWLEDGMENTS

M.L. is supported by a grant of the University of Applied Sciences Mittweida, Germany. M. K. is supported by the European Social Fond (ESF), Saxony. The authors acknowledge helpful discussions with J. Stelzer and G. Lohmann both of the Max-Planck-Institute for Human Cognitive and Brain Sciences, Leipzig, Germany as well as providing the real world data. Additionally, we thank J. Stelzer for carrying out the simulations for the artificial data matrix in sec. II-B2.

REFERENCES


