P-SVM Variable Selection For Discovering Dependencies Between Genetic And Brain Imaging Data

Johannes Mohr, Imke Puls, Jana Wrase, Sepp Hochreiter, Andreas Heinz and Klaus Obermayer

Abstract— **The joint analysis of genetic and brain imaging data is the key to understand the genetic underpinnings of brain dysfunctions in several psychiatric diseases known to have a strong genetic component. The goal is to identify associations between genetic and functional or morphometric brain measurements. We here suggest a machine learning method to solve this task, which is based on the recently proposed Potential Support Vector Machine (P-SVM) for variable selection, a subsequent k-NN classification and an estimation of the effect of 'correlations by chance'. We apply it to the detection of associations between candidate single nucleotide polymorphisms (SNPs) and volumetric MRI measurements in alcohol dependent patients and healthy controls.**

I. Introduction

An important topic in current psychiatric research is the study of the genetic influence on brain (dys)functions in patients suffering from diseases like alcohol dependence and schizophrenia. Often, the limbic system is affected, a group of brain structures, including the hippocampus and the amygdala, that are associated with arousal, motivation, emotion and recent memory. A genetic variability of serotonergic and dopaminergic neurotransmitter systems, which are targeted by drugs, might influence the treatment outcome. It is important to better understand these genetic modulations of higher brain functions, in order to get a deeper insight into the the pathophysiology of the brain and to identify genetic variables which might be used for screening. A promising strategy is the combined analysis of genetic data with magnetic resonance imaging (MRI) measurements of the brain.

Here, we consider genetic data consisting of single nucleotide polymorphisms, or SNPs (see Fig. 1). These are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. Each individual has many single nucleotide polymorphisms that together create a unique DNA pattern for that person. MRI measurements can be used to get morphological information, like form or size of brain structures of interest. The combination of genetic, clinical and MRI data results in datasets with a high number of variables and small sample size, which often cannot be dealt with by standard statistical methods. Therefore, we here propose an analysis using machine learning, which has three goals:

Johannes Mohr is with the Bernstein Center for Computational Neuroscience Berlin (BCCNB) and the Department of Psychiatry and Psychotherapy, Charité University Medicine Campus Mitte, Berlin, Germany (email: johann@cs.tu-berlin.de).

Imke Puls, Jana Wrase and Andreas Heinz are with the Department of Psychiatry and Psychotherapy, Charité University Medicine Campus Mitte, Berlin, Germany

Sepp Hochreiter is with the Institute for Bioinformatics, Johannes Kepler University, Linz, Austria

Klaus Obermayer is with the Department of Electrical Engineering and Computer Science, Berlin University of Technology, Berlin, Germany

A. Detection of Dependencies

The first aim is to check whether the data provides evidence for a statistical dependency between a set of d input variables $\mathbf{x}$ and a class label y. Under the assumption that a dependency exists, a powerful enough predictor should be able to learn it. The assessment of statistical dependency should be based on the generalization error of the prediction. This means that if the predictor was trained on a training dataset assumed to be sampled from a certain distribution, one wants to estimate the expected error it will make on a yet unseen data point from the same distribution. For small sample size, this generalization error can be estimated from the average test error using a leave-one out cross-validation method. The lower the generalization error, the stronger is the detected dependency. If the generalization error is very high, no evidence for a dependency could be found with the current predictor; either it is not present in the dataset, or the learning machine was not powerful enough. Therefore, this method cannot affirm independencies, only dependencies. However, any established dependency must still be assessed for "correlations by chance" (see next paragraph).

B. Assessing the Effect of Correlations by Chance

Dependencies between input and class variables might arise due to noise-induced "correlations by chance". Therefore, when judging the scientific relevance of a dependency found in the dataset, one has to assess how likely it is that a "correlation by chance" could have led to an equally low generalization error. The result of this assessment (a probability value) will strongly depend on the size of the dataset, the specific noise level, the distributions of input and class variables and the power of the predictor. The higher the probability that a generalization error as low as the one found on the dataset could have been produced by "correlations by chance", the lower is the evidence the data provides for suspecting a dependency. In case of a high probability it is nevertheless possible that with a larger sample the dependency could be affirmed, but given the present sample, not enough evidence is found.

C. Finding the Relevant Variables

If there was a dependency detected, and if the effect of correlations by chance is estimated to be low, the next thing we are interested in is to know which of the input variables

are most relevant for predicting the class label, in order to increase the interpretability of the results. Therefore, the task is to select a subset of the input variables which seems to be important for good prediction performance of the learning machine.

In this work, we suggest a new method to solve these tasks, which is based on the recently proposed P-SVM ([1], [2], [3]). The P-SVM selects a compact subset of the input variables, which is then used in the subsequent k-NN-classification of the target label. A Jackknife estimate of the correct classification rate per class gives a robust estimate of generalization performance, thus judging the evidence for a dependency between input variables and class label. To provide a measure of confidence, the probability that an equally good performance results from "correlations by chance" is estimated. While this method can be used in general, we apply it to the joint analysis of candidate SNPs, clinical variables (e.g. sex and age) and volumetric MRI brain measurements in alcoholic patients and healthy controls.

The paper is organized as follows: Section II reviews the medical background of the problem we focus on. This is followed by a description of the used methods (section III). First, the techniques used in the acquisition of the genetic (section III-A) and the MRI (section III-B) data are summarized. Then the conduction of the volumetric measurements (III-C) and the pre-processing of the SNP variables (III-D) are described. Section III-E deals with the methods used for variable selection and ranking. Section III-F reviews the P-SVM algorithm, section III-G the k-NN-classifier. The assessment of "correlations by chance" is described in section III-H. In section IV the dataset and the experimental setup are described, the results are given and their clinical interpretation is discussed. The paper closes with conclusions and outlook (V).

II. Medical Background

Studies over the last years have shown that the dopaminergic reward system is involved in the development of alcohol craving and reduced control of alcohol intake, factors that are known to be associated with an increased risk for alcohol dependence (for review see [4]). Case-control association studies and genome wide linkage analyses have identified associations between alcoholism and common functional polymorphisms in several candidate genes of the dopaminergic system, including dopamine receptor D2, dopamine transporter, and catechol-O-methyltransferase (COMT), one of the main enzymes involved in dopamine degradation (for review see [5]).

A single nucleotide polymorphism ("SNP", see Fig. 1) of COMT in exon 4 (allele $g \rightarrow a$) causes an amino acid exchange from valine to methionine ($Val^{158}Met$) and thereby reduces the enzyme activity to almost $\frac{1}{4}$. In several studies, the Met allele was associated with an increased risk to develop alcoholism ([6], [7]). Whereas subjects with the Met allele perform significantly better in working memory tasks, they also seem to be more sensitive to exogenous stress and anxiety-generating situations, and they show higher reactivity to unpleasant stimuli in the limbic system ([8]). However, other studies did not find a significant association between the $Val^{158}Met$ polymorphism and the risk to develop alcohol dependence ([9], [10], [11]).

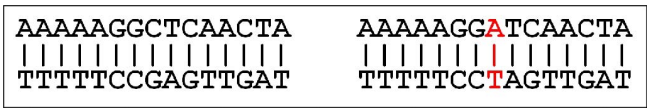

Fig. 1. The main chromosomal component is desoxyribonucleic acid (DNA), which is the carrier of genetic information. The molecular structure of DNA are two intertwined chains held together by hydrogen bounds. The main elements of DNA are the purine bases guanine (G) and adenine (A) and the pyrimidine bases cytosine (C) and thymine (T), where G pairs exclusively with C, and A with T. A SNP (single nucleotide polymorphism) represents an exchange of one single nucleotide within the genome sequence that is transmitted to all offspring. By definition a SNP occurs in at least 1% of the population. SNPs can be found at approximately every 300-1000 base pairs. These sequence variations are the main basis of interindividual differences in humans. On the left the original double-stranded DNA is shown, on the right the mutated DNA with the introduction of a SNP, in this case a substitution from C to A. The corresponding base on the second chain is altered accordingly.

Reasons for the contradictory results might be manifold and include multifactorial pathogenesis of complex disorders, locus heterogeneity, inhomogeneous patient samples, and differing ethnic backgrounds. Recent studies have started to dissect complex and heterogeneous disorders by using endophenotypes which generate more homogeneous diagnostic subgroups. The number of potentially influencing factors is reduced and genetic contributions can be more easily identified. Among a variety of promising alcohol-related endophenotypes, brain imaging studies seem to be a very robust tool ([12], [13]).

The observation of an association between a disease and a disease gene does not simply point to the causative genetic locus within the gene. Genotyping of several markers or SNPs and reconstructing their association, i.e. the construction of haplotypes, increases the chances to detect an association and to receive more reliable and stable results. SNPs are selected based on haplotype information in public databases, localization within the gene and locus heterogeneity. Recently, several haplotype analyses have been published on COMT polymorphisms, which show that genetic variations beside the $Val^{158}Met$ polymorphism are of functional relevance ([14]). There was also a significant association between a certain COMT haplotype and nicotine dependence ([15]).

By combining Magnetic Resonance Imaging (MRI) with human genome analysis, often datasets are created which are characterized by containing many variables and having a rather small sample size. Potential relations between volumetric data and a certain SNP might be rather complex and involve multiple morphometric variables. Furthermore, it is quite likely that on a small and noisy dataset random correlations arise between variables which are in fact independent.

Here, we consider a dataset consisting of detoxified alcohol-dependent patients and healthy control subjects which contains clinical data, eight COMT SNPs, and MRI-

based brain structure volume measurements from 75 subjects. The question of interest is whether the dataset provides evidence for any association between COMT haplotypes and hippocampal, amygdala and mammilary body atrophy.

We apply the novel variable selection and ranking method based on the Potential Support Vector Machine (P-SVM) to this task. In this case, the set of input variables $\mathbf{X}$ are the clinical and volumetric variables and the class label y corresponds a specific allele configuration of a SNP.

III. Methods

A. Genetic Methods

Genomic DNA was extracted from 75 individuals (34 alcohol dependent patients, 41 healthy controls). SNPs were selected based on HapMap data and other publications, localization within the gene and locus heterogeneity available from public databases. Oligonucleotide primers were designed for each SNP (primer information available on request), amplification was performed by polymerase chain reaction (PCR) according to general procedures. PCR products were incubated with suitable restriction enzymes (information from NEB-cutter). Restriction products were run on agarose gel electrophoresis to separate potential restriction products.

B. MRI Acquisition

The MRI scans were acquired using a 1.5T clinical whole-body MRI (Magnetom VISION; Siemens, Erlangen, Germany) equipped with a standard quadrature head coil. The automatic Siemens MAP shim was used for shimming. A morphological 3D T1-weighted MPRAGE (magnetization prepared rapid gradient echo) image dataset (1x1x1 mm voxel size, FOV 256 mm, 162 slices, TR=11.4 ms, TE=4.4 ms, $\alpha = 12°$) covering the whole head was acquired for anatomical study.

C. Morphometric Measurements

The morphometric measurements were based on the anatomical MRI scans. Several regions of interest (amygdala, hippocampal, and mammillary body) were segmented and their total volumes were measured. Volumes, calculated in cubic centimeters, for each individual structure were derived by multiplying the number of voxels assigned to that structure on each slice by the slice thickness and summing across all slices in which the structure appeared ([16]). To rule out gross volumetric effects contributing to the differences in local anatomic measures, total head circumference was measured in each subject and used as a subject-wise correction for total amygdala, hippocampal, and mammillary body volumes.

Moreover, a symmetry index (SI) was determined for each of the three brain regions ([17]):

$$SI = 1/100 \cdot \frac{LV - RV}{0.5(LV + RV)}$$

where RV and LV denote the volumes of the same anatomic region in the right and left hemisphere. Positive values indicate that the anatomic region of concern is larger in the left hemisphere. This symmetry index is a unitless quantity.

D. Pre-Processing of the Nominal SNP Variables

In order to make the nominal SNP variables compatible to a machine learning algorithm requiring vectorial data, they are transformed into three new binary variables, each indicating a specific allele configuration. This is necessary, because there is no a priori order to the three allele configurations which would justify the transformation into a single, discrete ordered variable.

The described data analysis problem can be cast in the form of a classification problem. The aim in the analysis was to predict a SNP configuration from brain imaging and clinical data. For a given SNP, each of the created binary SNP variables can serve as class label.

E. Variable Selection and Ranking Method

In terms of machine learning, the final goal of our analysis is to determine whether a dataset provides sufficient evidence that there is a dependency between the set of input variables and the class label. Moreover, we would like to determine the subset of the input variables which is responsible for the dependency. In this section we describe a method to solve this problem for small, noisy datasets with nominal and continuous variables.

For a given training dataset, the P-SVM variable selection, which will be described in section III-F, provides a compact, set of variables, which are suspected to be "informative" about the class label. Using this set of selected variables, a k-NN classifier ([18]) performs the actual prediction of the class label for a given test data point. In order to render the selection method robust against outlier samples, the P-SVM variable selection is run within a 5-fold cross validation[1](CV). This results in a ranking of the Variables according to how often they were chosen.

While any predictor (even the P-SVM itself) could be used as a classifier, we here chose a weighted k-NN classifier, which is described in more detail in section III-G. The hyper-parameters of this classifier are (1) the number of variables which are used from the ranking and (2) the number of neighbors of the k-NN algorithm. The optimal values are determined via leave-one-out cross validation (CV) on the training set. In the following we refer to the classifier using the optimal hyper-parameters as the "optimal classifier".

To get an estimate of the generalization performance of the predictor, a second (outer) leave-one-out CV loop is used. This provides a Jackknife estimate ([18]) of the correct classification rate per class. This performance measure is defined as the percentage of correctly classified data points in a class averaged over both classes, and is usually chosen in

[1]The q-fold cross validation method makes use of all available data, by splitting a dataset containing m examples into q disjoint subsets (folds) of the same size $\equiv m/q$. The algorithm is trained q times, each time using a different fold as hold-out test set and the remaining $q-1$ subsets as training set. If q is equal to the number of data points, this method is called leave-one-out cross validation. Because of the repeated training, cross validation is a computational intensive method

classification problems where the classes are of unequal size. Classifying all data points as belonging to the larger class would trivially result in a high total correct classification rate, whereas the correct classification rate per class yields a value of exactly 50%. The final variable ranking is calculated by combining the results from the single cross-validation folds. First, the variables are ranked according to how often they appear on the first h positions of the variable ranking of the optimal classifier (we used $h = 4$ in the experiments). Then, all other variables which were used in an optimal classifier are ranked according to their number of appearances.

The Jackknife estimate of generalization performance tells us how much the predictor (consisting of variable selector and classifier) is expected to be able to "learn" possible dependencies between the input variables and the class variable. A value of more than 50% means that this was done successfully. However, dependencies between input and class variables might arise due to noise-induced "correlations by chance", an effect which is stronger the smaller the sample and the noisier the data. It can never be totally ruled out that a dependency found by the predictor is in fact a result of such "correlations by chance". This fact is a basic principle of statistics, and is independent of the specific statistical method. Therefore, in section III-H we suggest a numerical procedure to estimate the probability that, given the input variables are in fact independent of the class label, "correlations by chance" lead to a correct classification rate per class which is at least equal to the one achieved on the true dataset. The lower this probability, the more evidence for a detected dependency is found in the dataset. The pseudocode of the whole variable selection and ranking algorithm is shown in Algorithm 1.

Algorithm 1 Variable Selection and Ranking Algorithm

BEGIN PROCEDURE
  for all leave-one-out CV folds i **do**
    training set $Train(i)$
    test point $Test(i)$
    P-SVM variable selection (5-fold CV) on $Train(i)$
    $\Rightarrow$ variable ranking (contains $R(i)$ variables)
    for all leave-one-out CV folds t **do**
      training set $Train(i,t)$
      test point $Test(i,t)$
      for $r = 1$ to $R(i)$ **do**
        for $k = 1$ to K_{max} **do**
          training of k-NN classifier on $Train(i,t)$
            using the r leading variables from the variable ranking
          test with k-NN classifier on the $Test(i,t)$
          remember test error $e(t,r,k)$
        end for
      end for
    end for
    $[r_{opt}(i), k_{opt}(i)] = min(e(t,r,k))$.
    $Feat(i) = \{$the r_{opt} first variables from the variable ranking$\}$
    training of the k_{opt}-NN classifier on $Train(i)$ using variables $Feat(i)$
    test with k_{opt}-NN-classifier on the test data point $Test(i)$
    remember test error $E(i)$
  end for
  $Variables$=Rank($Feat$)
END PROCEDURE

F. The P-SVM Algorithm

The P-SVM is a recently introduced ([1], [2], [3]) large-margin method for classification, regression and variable selection. It uses a novel objective function, which minimizes a scale-invariant capacity measure, and novel constraints, which enforce a low empirical error. In contrast to standard support vector machine approaches ([19], [20], [21], [22]), the P-SVM can also handle negative definite and non-square kernel matrices. A standard support vector machine expresses the classification or regression function via a subset of the data points, the so-called "support vectors". The basic idea behind P-SVM variable selection is to interpret the data matrix as a kernel matrix and to exchange the role of variables and data points. The weight vector w is expanded into a sparse set of "support variables", thus extracting a small number of "informative" variables from the set of all variables. The P-SVM as a filter method for variable selection was proposed in [3], where it was also bench-marked against other variable selection methods on the NIPS 2003 feature selection challenge and was shown to be one of the best methods for selecting a compact set of variables. Once a set of "support variables" was determined, it can be used as input to an arbitrary predictor.

In the following, we will briefly outline the mathematical formulation of P-SVM variable selection (for further details, see [1]).

We consider a two class classification task, where the m (d-dimensional) input vectors and class labels are summarized in the matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ and the vector $\mathbf{y}$. The learning task is to select a classifier g with minimal risk, $\mathrm{R}(g) = \min$, from the set of classifiers

$$g(\mathbf{x}) = sign\left(\mathbf{w} \cdot \mathbf{x} + b\right), \tag{1}$$

which are parameterized by the weight vector $\mathbf{w}$ and the offset b. Standardization (mean subtraction and dividing by the standard deviation) of the data leads to $\mathbf{X}^T\mathbf{1} = 0$. The primal P-SVM optimization problem for variable selection can then be formulated as

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{X}^T\mathbf{w}\|^2 \tag{2}$$

subject to

$$\begin{array}{l} \mathbf{X}^T\left(\mathbf{X}^T\mathbf{w} - \mathbf{y}\right) + \epsilon\mathbf{1} \geq \mathbf{0} \\ \mathbf{X}^T\left(\mathbf{X}^T\mathbf{w} - \mathbf{y}\right) - \epsilon\mathbf{1} \leq \mathbf{0}. \end{array} \tag{3}$$

The corresponding dual problem can be derived as

$$\begin{aligned} \min_{\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-} \quad & \frac{1}{2}\left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right)^T \mathbf{X}^T\mathbf{X}\left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) \\ & -\mathbf{y}^T\mathbf{X}\left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) + \epsilon\mathbf{1}^T\left(\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-\right) \\ \text{subject to} \quad & \mathbf{0} \leq \boldsymbol{\alpha}^+, \mathbf{0} \leq \boldsymbol{\alpha}^-. \end{aligned} \tag{4}$$

where ϵ is a parameter to determine the number of variables (a larger ϵ will result in fewer variables). The vectors $\boldsymbol{\alpha}^+$ and $\boldsymbol{\alpha}^-$ are the Lagrange multipliers for the constraints. The non-zero components α_j mark the support variables. The α_j^+ correspond to variables relevant for the positive class, while variables with non-zero α_j^- are indicative for the negative class. Eqs. (4) can be solved using a new sequential minimal optimization (SMO) technique [2]. Using $\boldsymbol{\alpha} = \boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-$, the weight vector $\mathbf{w}$ and the offset b are given by

$$\mathbf{w} = \boldsymbol{\alpha} \text{ and } b = \frac{1}{m}\sum_{i=1}^{m} y_i. \tag{5}$$

The resulting classifier is then given by

$$\begin{aligned} g(\mathbf{x}) &= sign(\mathbf{w}\cdot\mathbf{x} + b) \\ &= sign(\sum_{j=1}^{d}\alpha_j(\mathbf{x}\cdot\mathbf{e}_j) + b). \end{aligned}$$

G. The k-NN Classifier

We use a weighted k-NN classifier as predictor. It evaluates the Euclidean distance of a test data point $\mathbf{x}$ to all points of the training dataset. It then determines the k nearest neighbors, and orders them according to their distance from x. Then it assigns a weighting factor to each neighbor, which depends linearly on the Euclidean distance. A distance of zero gets assigned a value of five, while the most remote neighbor gets assigned a value of one. If the choice of the k-th nearest datapoint is not uniquely possible because multiple data points possess exactly the same distance, all candidates are included. Finally, the datapoint $\mathbf{x}$ gets assigned the sign of the weighted sum of the class labels of the k nearest neighbors as a class label.

H. Assessing the Effect of "Correlations by Chance"

It is assumed that the dataset is sampled independently and identically distributed (i.i.d.) from an underlying distribution $P(y, \mathbf{x})$. In case there is a dependency between the input variables and the class label, this distribution will not factorize, and it is possible to predict the class label from the conditional probability distribution $P(y|\mathbf{x})$. In this section, we describe a numerical method which allows to assess the probability that a sample drawn from the factorizing distribution $P(y, \mathbf{x}) = P(y)\cdot P(\mathbf{x})$, where input variables and class label are statistically independent, achieves a correct classification rate per class equal or higher that the one achieved on the actual dataset.

For this, many new datasets are randomly sampled from the factorizing distribution. This is done by randomly permuting the class labels of the examples from the original dataset, while keeping the input vectors. The random permutations destroy any existing correlation between $\mathbf{x}$ and y, but leave the distributions $P(\mathbf{x})$ and $P(y)$ unchanged. However, "correlations by chance" between $\mathbf{x}$ and y might appear. For each of these randomly sampled new datasets, the whole variable selection procedure as described in section III-E is run. The resulting correct classification rates per class are compared to the one achieved on the true data. The relative percentage of equal or higher values achieved under random label permutation is an estimate of the desired probability.

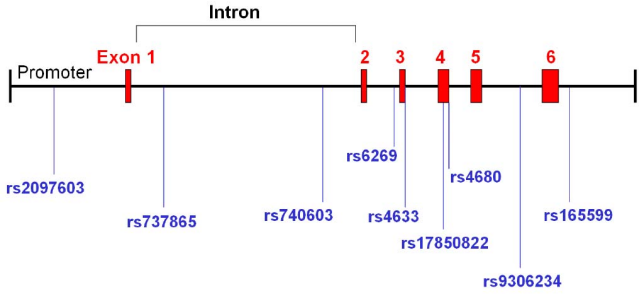

Fig. 2. Scheme of the catechol-O-methyltransferase (COMT) gene. Red boxes depict exons (i.e. the coding sequence) which is translated into amino acids. SNPs (marked in blue) are scattered throughout the gene, i.e. the promoter region (expression regulating sequence), exons and introns (sequence between exons, some might have regulatory functions, others are functionally inactive).

IV. Experiments

A. Dataset

The dataset this analysis is based on contained 75 subjects (22 female, 53 male, age: 25-61 years, 41 healthy controls and 34 alcohol dependent patients). For the analysis we used the following variables:

1) Clinical Variables: Clinical variables used were the Age (in years) the sex (male/female) and the status (patient/control). The first is an positive integer number, the latter two are binary variables.

2) Genetic Variables: The genetic variables (see section III-A) consist of 9 different COMT-SNPs, each of which can be considered as nominal variable with three levels, each of which corresponds to a certain configuration of the two alleles: rs2097603, rs737865, rs740603, rs6269, rs4633, rs17850822, rs4680, rs9306234 and rs165599. For details see Figure 2.

3) Morphometric Variables: The morphometric variables (see section III-C) are summarized in Table I

B. Experimental Set-Up

The first experiment used the status as class label, and the SNPs, age and sex as input variables. The second experiment again used the status as class label, and the volumetric data, age and sex as input variables. The remaining experiments were an explorative search: Each of the three allele configurations of the eight SNPs was taken as target value, while

Variable	Description
r_hppcps	volume of right hippocampus
l_hppcps	volume of left hippocampus
t_hppcps	total volume of hippocampus
r_amyg	volume of right amygdala
l_amyg	volume of left amygdala
t_amyg	total volume of amygdala
r_mambdy	volume of right mammillary body
l_mambdy	volume of left mammillary body
t_mambdy	total volume of mammillary body
si_hippo	SI for hippocampus
si_amygd	SI for amygdala
si_mam	SI for mammillary body

TABLE I

Name and description of the morphometric variables used in the experiments

the input variables always consisted of the 12 morphometric variables, sex and age.

In each case, the feature selection method described in section III-E was applied. Subjects with missing values in either $\mathbf{x}$ or y were left out. On the remaining examples the Jackknife estimate of the correct classification rate per class was calculated. In case it exceeded a value of 50 percent, the method for assessing the effect of "correlations by chance", which is described in section III-H, was applied using 100 runs on datasets generated under random permutations of the labels. In case the resulting probability estimate was lower than 10%, additional 2000 runs were conducted to improve the numerical accuracy of the estimate. This two-step strategy was used, because the computational demands of the estimation procedure were rather high. The strategy is very conservative, since in case the initial runs are not representative for the whole distribution, some true dependencies might not be detected, whereas a detected dependency is always verified by the following more accurate numerical estimation. Since the label permutations are conducted independently of each other, the task can be split in several calculations which are run in parallel.

C. Results

In the first two experiments, no evidence for a dependency neither between SNPs and status, nor between volumetric data and status was found. In the other experiments, 18 allele configurations had a generalization performance of more than 50%. However, only for five allele configurations of four of the SNPs the probability that "correlations by chance" achieve equally good prediction was smaller than 10%. These results are listed in Table II and in each case the selected variables are ranked with decreasing importance from top to bottom. In the experiments, the initial estimates from 100 runs were always close to the final results, so they seem suitable to judge whether it is worthwhile to conduct further runs in order to increase the numerical accuracy.

How should the results be interpreted? A value over 50% for the correct classification rate per class (CCRPC) indicates at least a small dependency which was detected on the dataset. Note that this fact is based on the generalization properties of the learned predictor on unseen data, estimated via leave-one-out CV. The size of the dependency is indicated by the value of the CCRPC, which amounts to the estimated probability per class that an unseen data point will be correctly classified. The value needed in order to assess how likely such a prediction performance could be produced by "correlations by chance" is shown in the last column of Table II. This value indicates how much confidence one can place in the dependency given the dataset.

Name	m_+/m_-	CCR	CCRPC	Ranking	P
rs2097603_2	37 / 36	66%	66%	r_hppcps l_hppcps si_hippo l_amyg si_mam	1.95%
rs740603_2	42 / 31	62%	63%	l_amyg r_hppcps r_amyg	7.33%
rs740603_3	15 / 58	79%	87%	l_mambdy r_mambdy si_hippo	1.52%
rs4633_3	15 / 59	80%	87%	l_hppcps l_mambdy r_hppcps r_mambdy	1.00%
rs4680_2	39 / 33	61%	62%	l_hppcps l_mambdy r_amyg r_mambdy si_amygd r_hppcps	9.95%

TABLE II

Results from the experiments. The name of the SNP is followed by an underscore and the allele configuration. m_+ and m_- denote the number of examples in the positive and negative class, respectively. "CCR" denotes the total correct classification rate, independent of the class, whereas "CCRPC" denotes the correct classification rate per class, which is the performance measure one is interested in. "Ranking" lists the variable names in the order in which they were ranked. "P" gives the probability that "correlations by chance" achieved a CCRPC equal to or better than the one found on the dataset, based on 2100 random permutations of the labels. Only the results with $P \leq 10\%$ are listed.

D. Discussion

This study provides an interesting and alternative view on the involvement of COMT in the limbic system. Associations between several SNPs of COMT and volume of hippocampus, amygdala and mammillary body have been found. The limbic system is the center of emotion, motivation and emotional association with memory, features affected in patients with alcoholism. COMT is one of the main dopamine degrading enzymes, that partly controls dopamine brain level. Dopamine is one of the main actors within the limbic system. It was previously shown that one SNP of COMT, the $Val^{158}Met$ is associated with reactivity to unpleasant stimuli in the limbic system ([8]), however, other genetic variations in the COMT gene were not investigated.

Our data indicate that the genetic constitution of the COMT gene affects the volume of certain limbic structures. This might be not surprising since dopamine and norepinephrine activity can affect the underlying morphology of brain tissue, e.g. via stimulation of cyclic AMP (cAMP), which gates synaptic actions of brain derived neurotrophic

factor (BDNF) ([23]), which is involved in neuronal plasticity. Carriers with the less active Met allele have higher extracellular dopamine and norepinephrine levels. This might activate the BDNF receptor Trk B via cAMP-dependent phosphorylation and its translocation to spines in mature hippocampal neurons ([23]). These effects may modulate the neurotoxic effects of alcohol on brain tissue and contribute to the processing of affective cues in the hippocampus and potentially also in other limbic brain areas.

Our data implicate that the genetic constitution of the COMT gene not only as an impact on functional activity of the limbic system but also affects volume of certain limbic structures. This might not be surprising, since dopamine activity could affect underlying morphology of brain tissue. Carriers with the less active Met allele have high dopamine brain level, thereby receiving permanent dopamine input. This might activate plasticity of neurons, synapses are spouting and further interneuronal connections might be established. This would also fit to the functional highly active areas in the limbic system in subjects with one or two Met alleles.

No association was found between diagnosis of alcoholism and any genetic variations within the COMT gene. This finding is not surprising, since COMT may interact with the effects of excessive alcohol intake once it was established, rather than contribute to the risk of excessive alcohol intake. Also, since alcoholism is a complex disorder with several genetic loci affected, very large patient samples are needed to detect genetic associations with behaviourally heterogeneous disease categories. Smaller sample sizes might be sufficient to detect gene effects on closer related intermediate phenotypes.

Altogether, this pilot study illustrates the use of the proposed machine learning method in elucidating the interaction between several genetic polymorphisms and a complex clinical dataset. However, further studies willbe needed to gain more information: Replications have to be performed with a new dataset to control for results of this study and other statistical methods have to be used to receive information concerning the causality of the observed relations.

V. Conclusions and Outlook

In this paper, we proposed a method for variable selection and ranking based on the P-SVM and a k-NN classifier and applied it to the task of determining statistical dependencies between a group of variables and a class label. The Jackknife estimate of the correct classification rate per class was used to detect dependencies, while the estimated probability that "correlations by chance" achieve equally good performance is used to provide a measure of confidence. The application of two layers of leave-one-out cross-validation loops allows to extract robust estimates of existing dependencies. For large sample sizes m, however, the computational demands would be too high, therefore one would use q-fold cross validation with $q < m$ instead. While in the current paper we employed a k-NN classifier, in principle any predictor could be used.

The method was applied to a clinically relevant question on a real-world dataset containing SNPs and volumetric MRI measurements. However, the proposed method is not restricted to this specific application. Determining dependencies between a set of variables and a class label on datasets of small sample size is a task often encountered in clinical resarch. Future applications will include the joint analysis of genetic and functional brain data (fMRI).

Acknowledgment

We thank Peter Sander for his help with the numerical simulations. This work was funded by the Bernstein Center for Computational Neuroscience Berlin (BMBF grant 01GQ0411).

References

[1] S. Hochreiter and K. Obermayer, "Support vector machines for dyadic data," *Neural Computation*, 2006, in press.

[2] ——, "Classification, regression, and feature selection on matrix data," Technische Universität Berlin, Fakultät für Elektrotechnik und Informatik, Tech. Rep. 2004/2, 2004.

[3] ——, "Nonlinear feature selection with the potential support vector machine," in *Feature extraction, Foundations and Applications*, I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Eds. Springer, 2005.

[4] A. Heinz, M. Schafer, J. Higley, J. Krystal, and D. Goldman, "Neurobiological correlates of the disposition and maintenance of alcoholism," *Pharmacopsychiatry*, vol. 36, no. suppl3, pp. 561–70, 2003.

[5] A. Heinz, D. Goldman, J. Gallinat, G. S. G, and I. Puls, "Pharmacogenetic insights to monoaminergic dysfunction in alcohol dependence," *Psychopharmacology*, vol. 174, pp. 561–570, 2004.

[6] T.Wang, P. Franke, H. Neidt, S. Cichon, M. Knapp, D. L. D, W. Maier, P. Propping, and M. M. Nothen, "Association study of the low-activity allele of catechol-o-methyltransferase and alcoholism using a family-based approach," *Mol Psychiatry*, vol. 6, pp. 109–111, 2001.

[7] J. Tiihonen, T. Hallikainen, H. Lachman, T. Saito, J. Volavka, J. Kauhanen, J. T. Salonen, O. P. Ryynanen, M. Koulu, M. K. Karvonen, T. Pohjalainen, E. Syvalahti, and J. Hietala, "Association between the functional variant of the catechol-o-methyltransferase (comt) gene and type 1 alcoholism," *Mol Psychiatry*, vol. 4, pp. 286–289, 1999.

[8] M. N. Smolka, G. Schumann, J. Wrase, S. M. Grusser, H. Flor, K. Mann, D. F. Braus, D. Goldman, C. Buchel, and A. Heinz, "Catechol-o-methyltransferase val158met genotype affects processing of emotional stimuli in the amygdala and prefrontal cortex," *J Neurosci*, vol. 25, pp. 836–842, 2005.

[9] Y. S. Kweon, H. K. Lee, C. T. Lee, and C. U. Pae, "Association study of catechol-o-methyltransferase gene polymorphism in korean male alcoholics," *Psychitr Genet*, vol. 15, pp. 151–154, 2005.

[10] Y. Liu, K. Yoshimura, T. Hanaoka, S. Ohnami, S. O. amd T. Kohno, T. Yoshida, H. S. H, T. Sobue, and S. Tsugane, "Association of habitual smoking and drinking with single nucleotide polymorphism (snp) in 40 candidate genes: data from random population-based japanese samples," *J Hum Genet*, vol. 50, pp. 62–68, 2005.

[11] T. Hallikainen, H. Lachman, T. Saito, J. Volavka, J. Kauhanen, J. T. Salonen, O. Ryynanen, M. Koulu, M. Karvonen, T. Pohjalainen, E. Syvalahti, J. Hietala, and J. Tiihonen, "Lack of association between the functional variant of the catechol-o-methyltransferase (comt) gene and early-onset alcoholism associated with severe antisocial behavior," *Am J Med Genet*, vol. 96, pp. 348–352, 2000.

[12] G. Oroszi and D. Goldman, "Alcoholism: genes and mechanisms," *Pharmacogenomics*, vol. 5, no. 1037-1048, 2004.

[13] M. A. Enoch, M. Schuckit, B. A. Johnson, and D.Goldman, "Genetics of alcoholism using intermediate phenotypes," *Alcohol Clin Exp Res*, vol. 27, pp. 169–176, 2003.

[14] N. J. Bray, P. R. Buckland, N. M. Williams, H. J. Williams, N. Norton, M. J. Owen, and M. C. O'Donovan, "A haplotype implicated in schizophrenia susceptibility is associated with reduced comt expression in human brain," *Am J Hum Genetic*, vol. 73, pp. 152–161, 2003.

[15] J. Beuten, T. J. Payne, J. Z. Ma, and M. D. Li, "Significant association of catechol-o-methyltransferase (comt) aplotypes with nicotine dependence in male and female smokers of two ethnic populations," *Neuropsychopharmacolgy*, 2006.

[16] D. Kennedy, P. A. Filipek, and V. S. Caviness, "Anatomic segmentation and volumetric calculations in nuclear magnetic resonance imaging," *IEEE Transactions on Medical Imaging*, vol. 8, pp. 1–7, 1989.

[17] A. M. Galaburda, G. D. Rosen, and G. F. Sherman, "Individual variability in cortical organization - its relationship to brain laterality and implications to function," *Neuropsychologia*, vol. 28, pp. 529–546, 1990.

[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley & Sons, 2001.

[19] B. E. Boser, I. M. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, D. Haussler, Ed. Pittsburgh, PA: ACM Press, 1992, pp. 144–152.

[20] C. Cortes and V. N. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[21] B. Schölkopf and A. J. Smola, *Learning with kernels – Support Vector Machines, Reglarization, Optimization, and Beyond*. MIT Press, Cambridge, 2002.

[22] V. N. Vapnik, *Statistical Learning Theory*, ser. Adaptive and learning systems for signal processing, communications, and control. New York: Wiley, 1998.

[23] Y. Ji, P. T. Pang, L. Feng, and B. Lu, "Cyclic amp controls bdnf-induced trkb phosphorylation and dendritic spine formation in mature hippocampal neurons," *Nature Neuroscience*, vol. 8, pp. 164–172, 2005.