Knowl Inf Syst (2007) 11(2): 243–258
DOI 10.1007/s10115-006-0043-5

Knowledge and Information Systems

REGULAR PAPER

Glenn Fung · Jonathan Stoeckel

SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information

Received: 30 November 2005 / Revised: 12 January 2006 / Accepted: 20 February 2006 /
Published online: 21 September 2006

Abstract Alzheimer's disease is the most frequent type of dementia for elderly patients. Due to aging populations, the occurrence of this disease will increase in the next years. Early diagnosis is crucial to be able to develop more powerful treatments. Brain perfusion changes can be a marker for Alzheimer's disease. In this article, we study the use of SPECT perfusion imaging for the diagnosis of Alzheimer's disease differentiating between images from healthy subjects and images from Alzheimer's disease patients. Our classification approach is based on a linear programming formulation similar to the 1-norm support vector machines. In contrast with other linear hyperplane-based methods that perform simultaneous feature selection and classification, our proposed formulation incorporates proximity information about the features and generates a classifier that does not just select the most relevant voxels but the most relevant "areas" for classification resulting in more robust classifiers that are better suitable for interpretation. This approach is compared with the classical Fisher linear discriminant (FLD) classifier as well as with statistical parametric mapping (SPM). We tested our method on data from four European institutions. Our method achieved sensitivity of 84.4% at 90.9% specificity, this is considerable better the human experts. Our method also outperformed the FLD and SPM techniques. We conclude that our approach has the potential to be a useful help for clinicians.

Keywords Alzheimer's disease · Support vector machines · Medical imaging · Mathematical programming

G. Fung · J. Stoeckel
Siemens Medical Solutions USA, Computer Aided Diagnosis, Malvern, PA 19355, USA

Present address:
G. Fung(✉)
Computer Aided Diagnosis & therapy group, 51 valley Stream Parkway, Nalvern , PA, USA
E-mail: glenn.fung@siemens.com

1 Introduction

Alzheimer's disease (AD) is the most frequent type of dementia for elderly patients. Due to aging populations, its occurrence will still increase. Even though no definitive cure has been found for this disease, reliable diagnosis is useful for excluding other dementias, choosing the right treatment and for the development of new treatments.

AD is diagnosed using the criteria from the National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) [1]. These criteria include dementia established by examination and objective testing; deficits in two or more cognitive areas; progressive worsening of memory and other cognitive functions; no disturbance in consciousness; and onset between ages 40 and 90. Absence of systemic disorders or other brain diseases, which could account for the deficits in memory and cognition, should also be established. In practice the main tool for evaluating patients are neuro-psychologic tests, that test abilities like memory and language. The Mini Mental State Examination (MMSE) is the most widely used of these tests [2].

Brain images can also provide some helpful indication of AD. Magnetic resonance imaging (MRI) is used to study possible anatomical changes of the brain [3]. The shrinkage of the hippocampus, a region of the brain showing some of the first signs of Alzheimer's disease, occurs very early in the disease process, long before the illness spreads to the cerebral cortex and results in cognitive and memory impairment and its volume change is an important sign for the detection of Alzheimer's disease in MRI images [3]. Images showing the local perfusion (amount of blood flow) of the brain can be used for the diagnosis of AD because the perfusion pattern is affected by the disease. In this article, we will look into the use of cerebral perfusion imaging acquired by single photon emitting computer tomography (SPECT) using technetium-99m hexamethylpropylene amine oxime (HMPAO) as the tracer. SPECT imaging is a largely accepted clinical modality for AD diagnosis. Even though the perfusion pattern and its evolution is not the same for all patients some hypo-perfusion patterns seem to be typical for the disease. There are three main regions mentioned in literature attained by hypo-perfusion [4]:

Temporo-parietal region: Many studies have shown this region to be typical for Alzheimer's disease; however, most studies were carried out with patients included having advanced Alzheimer's disease who are no longer characterized by a specific cognitive impairment but by general cognitive decline. So this region was not found for early Alzheimer's disease. Although bilateral temporo-parietal abnormalities, with or without other regional defects, are known as the predominant pattern for Alzheimer disease [5–8], they appear to be neither sensitive nor specific for early Alzheimer's disease [9, 10].

Posterior cingulate gyri and precunei: Kogure et al. [11] show these regions to be affected in a study on mild cognitive impaired subjects which turn out to have Alzheimer's disease after a 2 years follow-up. Other studies confirm these findings on early Alzheimer as well [12–14]. These perfusion deficits are probably more specific and more frequent in early Alzheimer's disease than temporo-parietal deficits.

Medial temporal lobe: Hypo-perfusion in these regions was only noticed during follow-up of the patients [11]. These results are surprising as previous pathological and anatomical studies have suggested that these regions are the first affected by the disease [15, 16]. Some research suggests that hypo-perfusion in these regions, like in the hippocampus, is not found in mild Alzheimer's disease due to the difficulties of imaging these deep brain structures with SPECT [17]. The discussion of the hypo-perfused regions above and of the differences that exist even between healthy subjects shows the difficulties that exist for physicians when analyzing Alzheimer's disease perfusion images. Thus it might be useful to have tools that could assist physicians in this difficult task. In this article, we will present a method that does not need any explicit knowledge about the perfusion pattern of Alzheimer's disease patients.

Some approaches for a computer aided diagnosis (CAD) system for the analysis of SPECT images for AD can be found in literature. The first family is based on the analysis of regions of interest. The mean values for these regions are analyzed using some discriminant functions (see e.g. [18, 19]).

The second approach is statistical parametric mapping (SPM) and its numerous variants. Statistical parametric mapping is widely used in the neuro-sciences. Its framework was first developed for the analysis of SPECT and PET studies, but is now mainly used for the analysis of functional MRI data. It was not developed specifically to study a single image, but for comparing groups of images. One can use it for diagnostics by comparing the image under study to a group of normal images.

Statistical parametric mapping consists of doing a voxel-wise statistical test, in our case a t-test, comparing the values of the image under study to the mean values of the group of normal images. Subsequently, the significant voxels are inferred by using random field theory (see e.g. [20] for a full description of SPM). A largely used freely available implementation called SPM99 [21] has been developed and is used in this article as comparison to our approach.

In this article, we will propose another approach using as less a-priori information about the pathology as possible, by obtaining it implicitly from image databases. Another important aspect is that our approach is global. that all the information in the image can be used at once in contrast to more local approaches, e.g mono-variate methods like SPM. A multi-variate approach generally increases sensitivity at the price of loosing regional specificity (e.g. depicting local hypo-perfusion regions). However, in the approach presented in this paper compared to our earlier work [22], we use feature selection while trying to add spatial constraints to the classification.

The following section first discusses the pre-processing of the data; next, we describe our proposed mathematical programming formulation. Unlike the traditional SVM-like formulations, spatial information about the feature (voxels) locations is incorporated into the optimization problem. This leads to feature selection where the classifier depends on regions in the brain instead of isolated non-connected voxels. In Sect. 3, we present the data we used for our experiments. It consists of real brain SPECT images obtained from four different institutions. The results on the data are presented in Sect. 4 and discussed in Sect. 5.

1.1 Notation

We now describe the notation used in this paper. The notation $A \in R^{m \times n}$ will signify a real $m \times n$ matrix. For such a matrix, A' will denote the transpose of A and A_i will denote the i-th row of A. All vectors will be column vectors. For $x \in R^n$, $\|x\|_p$ denotes the p-norm, $p = 1, 2, \infty$. A vector of ones in a real space of arbitrary dimension will be denoted by e. Thus, for $e \in R^m$ and $y \in R^m$, $e'y$ is the sum of the components of y. A vector of zeros in a real space of arbitrary dimension will be denoted by 0. A *separating hyperplane*, with respect to two given point sets $\mathcal{A}$ and $\mathcal{B}$, is a plane that attempts to separate R^n into two halfspaces such that each open halfspace contains points mostly of $\mathcal{A}$ or $\mathcal{B}$.

2 Methods

2.1 Spatial normalization

In the classifier-based approach, we need the assumption that the same position in the volume coordinate system within different volumes corresponds to the same anatomical position. This makes it possible to do meaningful voxel-wise comparisons between images. However, this assumption is not met by the images without pre-processing: First of all, the subject which is being imaged, is not always positioned at the same position in the reference frame of the imaging device. This reference frame defines where e.g. the brain is positioned in the image. Secondly, the anatomy does not always have the same shape and size between different subjects. For example, the size and shape of the skull can already be largely different between subjects. This means that we need to spatially register the volumes. In recent years, many registration algorithms have been developed, we point the interested reader to the following reviews for more information [23–26]. In our application, we do not have detailed knowledge of the anatomy of our subjects as only HMPAO-SPECT images of the subjects were available. These images are so-called functional images. They only depict the regional blood flow of the subject. The regional cerebral blood flow provides us of course with some gross information about the anatomy, but only based on the fact that there is a relationship between the blood flow, and the underlying anatomy. Understanding this characteristic of HMPAO-SPECT images is fundamental for the choice of the registration method.

In the ideal case anatomical images of the subjects should be acquired as well. CT images, or even better MRI images would provide detailed insight in the anatomy. The registration of images of these modalities of different subjects with each other, would provide us with the transformations between the images based on the anatomy of the subjects. Subsequently, these transformations could be applied to the functional images. Due to practical clinical limitations, however, these modalities of images were not available. Hence, we tried to deduce the shape and size of the anatomy based on the functional images.

Because of the limited anatomical information available in the volumes, we chose to estimate affine transformations between the volumes and not use transformations with a larger number of degrees of freedom. We used the correlation ratio as the similarity measure [27] that we minimized using Powell optimization

[28]. To obtain a more robust result, we used the following procedure. First of all, we registered all volumes to a single volume, then we calculated a mean volume. This mean volume was first put on the mid-sagittal plane by registering it with a flipped version (see [29]). Subsequently, it was made to be symmetrical by taking the mean of itself with a flipped version. Finally, all volumes were matched to this volume.

2.2 Intensity normalization

HMPAO-SPECT imaging generates volumes that only give a relative measure of the blood flow. The blood flow measure is relative to the blood flow in other regions of the brain. Direct comparison, of the voxel intensities, between images, even different acquisitions of the same subject, is thus not possible without normalization of the intensities.

For all the experiments, we normalize the intensities by applying an affine transformation to the intensities. The transformation parameters are estimated on the training set of each experiment such that the intensities for each voxel position have zero mean and standard deviation of one for all the training subjects. We choose this very common data normalization since it provides numerical stability to the algorithms involved. In [30] it is shown that the relationship between the tracer activity and the blood flow is not completely linear. So the HMPAO-SPECT images show less contrast between high and low activity flow regions than would be expected of the regional blood flow. Even though this effect exists, we assumed it to be small, and not to influence the ratio between high and low blood flow as a function of the tracer concentration in the blood. Thus, for our work, we assumed a linear scaling of the intensities between the images to be sufficient.

It should be noted, that because we only have relative measurements, changes in the global blood flow might induce perturbing effects [31]. For example, if the blood flow doubles in one region, but only increases by a quarter in other region, we might be tempted to conclude that for this case, the blood flow in this latter region has decreased, when comparing to other images where these effects have not taken place. This might happen when comparing images of Alzheimer's disease patients with images of normal subjects. This measured reduction, even if it does not correspond to the physical reality, can of course still be used as a diagnostic sign.

2.3 Classification

Because the hypo-perfusion pattern for early AD is not very well defined we choose to develop a method where we do not use any explicit knowledge about the typical perfusion patterns. We use implicit knowledge about the perfusion patterns by using a database of images of AD patients and normal subjects. To separate the images we use a classifier using the voxel intensities as features and this database to train the classifier. Using the voxel intensities as features makes it possible not to introduce any particular knowledge about the exact location of the hypo-perfusion area(s).

Thus by using a database of images and the voxel intensities we circumvent the problem of the exact definition of the typical perfusion pattern for early AD. In general, the number of images available in the training databases is significantly smaller (<100) than the number of voxels (>1000). Thus, the number of features (voxels) is much larger than the number of samples (training images). The number of samples is considered to be small if it is about the same or smaller than the number of dimensions. In this case, we speak of almost empty spaces, the small sample size problem or the so-called curse of dimensionality. In classical pattern recognition, it is believed that no good generalization could be obtained for these cases when using the whole feature space [32]. Generalization is the capacity of a classifier to rightly classify a sample never seen before. In order to improve generalization of our final classifier, minimal feature dependency (small amount of features) of the classifier is desired.

2.3.1 The linear support vector machine

We consider the problem, depicted in Fig. 1, of classifying m points in the n-dimensional real space R^n, represented by the $m \times n$ matrix A, according to membership of each point A_i in the class $A+$ or $A-$ as specified by a given $m \times m$ diagonal matrix D with plus ones or minus ones along its diagonal. For this problem the standard support vector machine with a linear kernel [32] is given by the following quadratic program with parameter $\nu > 0$:

$$\begin{array}{ll} \min\limits_{(w,\gamma,y)\in R^{n+1+m}} & \nu e'y + \frac{1}{2}w'w \\ s.t. & D(Aw - e\gamma) + y \geq e \\ & y \geq 0. \end{array} \tag{1}$$

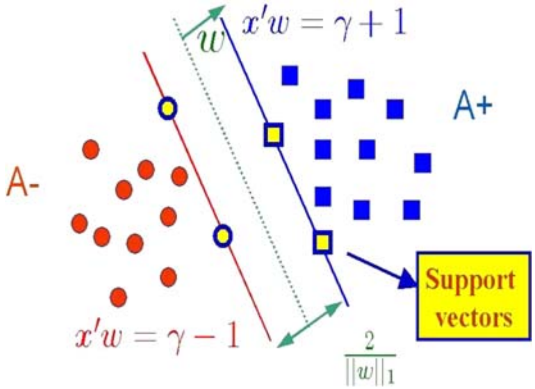

Fig. 1 The approximately bounding planes of Eq. (2) with a soft (i.e. with some error) margin $\frac{2}{\|w\|_1}$, and the plane $x'w = \gamma$ approximately separating $A+$ from $A-$ are represented by the red, green and blue lines. In this case, the support vectors are the points that lie on the bounding planes

Here, the plane $x'w = \gamma + 1$ bounds the class $A+$ points, while the plane $x'w = \gamma - 1$ bounds the class $A-$ points as follows:

$$\begin{aligned} A_i w &\geq \gamma + 1, \quad \text{for } D_{ii} = 1 \\ A_i w &\leq \gamma - 1, \quad \text{for } D_{ii} = -1. \end{aligned} \tag{2}$$

The linear separating surface is the plane $x'w = \gamma$ midway between the bounding planes (2). The quadratic term in (1) maximizes the distance or "margin" between the bounding planes. Maximizing the margin enhances the generalization capability of a support vector machine [32]. In order to make use of a faster linear programming based approach, instead of the standard quadratic programming formulation (1), we reformulate (1) by replacing the 2-norm by a 1-norm as follows [33]:

$$\begin{array}{ll} \min\limits_{(w,\gamma,y)\in R^{n+1+m}} & \nu e'y + \|w\|_1 = \nu \sum\limits_{i=1}^{m} y_i + \sum\limits_{j=1}^{n} |w_j| \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & y \geq 0. \end{array} \tag{3}$$

This SVM$\|\cdot\|_1$ reformulation in effect maximizes the margin, the distance between the two bounding planes of Fig. 1, using a different norm, the ∞-norm, and results with a margin in terms of the 1-norm, $\frac{2}{\|w\|_1}$, instead of $\frac{2}{\|w\|_2}$ [34]. The mathematical program (3) is easily converted to a linear program as follows:

$$\begin{array}{ll} \min\limits_{(w,\gamma,y,v)\in R^{n+1+m+n}} & \nu e'y + e'v = \nu \sum\limits_{i=1}^{m} y_i + \sum\limits_{j=1}^{n} v_j \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & v \geq w \geq -v \\ & y \geq 0, \end{array} \tag{4}$$

Empirical evidence [33] indicates that the 1-norm formulation has the advantage of generating very sparse solutions. This results in the normal w to the separating plane $x'w = \gamma$ having many zero components, which implies that many input space features do not play a role in determining the linear classifier. This makes this approach suitable for feature selection in classification problems. We note that in addition to the conventional interpretation of smaller ν as emphasizing a larger margin between the bounding planes (2), a smaller ν here also results in a sparse solution. The "right" value of ν is determined by a tuning procedure where the performance is adjusted to the desired compromise between the classification performance and the sparseness of the solution. Next, we will revisit some regularization theory results that would motivate the SVM-like formulation we are proposing in this paper.

2.4 Regularization theory and SVMs

Let $f : \Re^n \rightarrow \Re$ with $f(x) = w'x - \gamma$ the our prediction or classification function. Then, Formulation (4) and support vector machine (SVM) formulations in general can be seen as a particular case of regularization networks [35] where the functional $R_{\text{reg}}[f] = R_{\text{emp}} + \lambda G(Pf)$ that is often referred as the regularized risk, is minimized. $R_{\text{reg}}[f]$ is equal to the empirical risk functional $R_{\text{emp}}[f]$ plus a regularization term $G(Pf)$ that is usually defined as $\|Pf\|^2$. $\lambda = \frac{1}{\nu}$ is the regularization parameter and P is a called the regularization operator. P maps the the classifier function f into some dot product space [36]. For example, in the case of SVMs, the type of regularization and the class of functions that form the basis for the prediction function are intimately related. The SVM algorithm is equivalent to minimizing $R_{\text{reg}}[f]$ on the family of functions $f(x) = \sum_i \alpha_i k(x_i, x) + b$ provided that the kernel k is chosen as a Green's function of $P * P$ [36]. For example, in Formulation (4) the regularization term is $G(Pf) = \|w\|_1$. and $K(x_i, x_j) = x_i' x_j$ (the linear kernel). Our proposed formulation also minimize the regularized risk $R_{\text{reg}}[f]$ but for a very specific linear regularization operator P that encodes prior information (in the form of spatial information) about the classification task at hand.

2.4.1 The contiguous linear SVM (CSVM)

There are two drawbacks related to standard SVM formulations, especially when they are applied to imaging classification problems. The first drawback is related to the fact that little or no spatial information about the imaging problem is incorporated into the optimization problem to solve, discarding very valuable information that could lead to better and more robust classifiers. In the case of imaging problems where the features are related to voxel/pixel intensities a relation can be predefined among the voxels using spatial information or previous knowledge about the problem. The second drawback is related to the interpretability of the results. In several applications a feature selection scheme is implemented not only to get sparse models but also to determine which of the input features are relevant for the classification task, leading to insights about the problem in question. For example in the problem that we are addressing in this article it is easier to interpret a final classifier depending on contiguous voxels defining regions than a subset of independent voxels with no apparent connection among them. Our goal in this paper is to incorporate spatial information about every voxel into the optimization problem in a manner that the final obtained hyperplane classifier depends on regions or clusters of features rather than on isolated voxels. Let us consider a similarity function r that defines binary relations among any two features (f_i, f_j) of any given training datapoint. Let R be a matrix such that:

$$R_{ij} = r(f_i, f_j) \in \{0, 1\}, i, j \in \{1, \ldots, n\}$$

We define now, $\hat{R} = R - I_{n \times n}$, $\hat{R}$ is the symmetric adjacency matrix of an undirected graph representing the relation among the features according to the relation function r. R is a pseudo-adjacency matrix of a graph where every node has a self-loop. For most problems in real life R is based on local relations and therefore it is a very sparse matrix (see e.g. Fig. 3). The function r could be defined in a more general way, where instead of a binary relations it can be a similarity function or any other kind of function encoding extra information about the features or the datapoints in the training set. Figure 2 shows three examples of graphs representing possible relations among features. For example, Graph (a) represents a

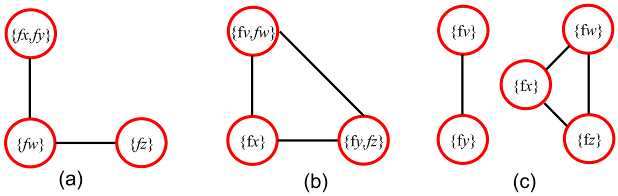

Fig. 2 Three examples of graphs representing different relations among features

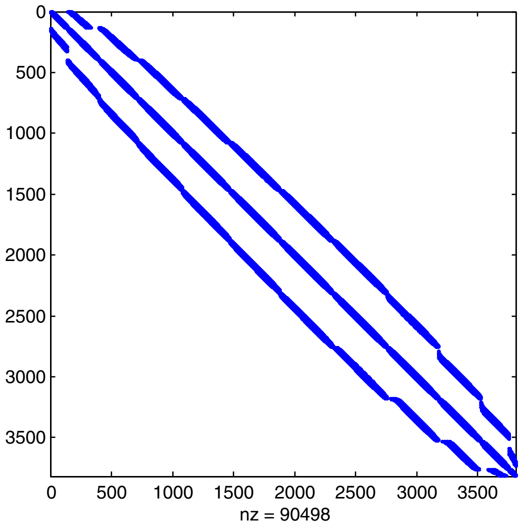

Fig. 3 The sparse adjacency matrix R for the mask defining the 26-closest neighbors of each voxel

dataset where f_x and f_y are directly related to feature f_w and feature f_w is related to feature f_z , note that features f_x and f_y are indirectly related (by transitivity) to feature f_z, even when there is not an explicit edge in the graph connecting them. Graph (b) represents a graph where every feature is related to each other and graph (c) represents a graph that is not fully connected , where the features are clustered (two clusters) according to the relations among them.

In our specific case, we choose the relation r to be defined by a $3 \times 3 \times 3$ mask defining the 26-closest neighbors of each voxel. Note that this very local simple mask allows to encode the sense of contiguity among voxels in a global sense across the whole volume. This mask size was chosen because it provided excellent results while maintaining the sparsity of the relation r A very simple but effective way to incorporate this extra information about the features into the 1-norm SVM formulation (4) is to use the relationship matrix R as a regularization operator and then minimize the the regularized risk:

$$R_{\text{reg}}[f] = R_{\text{emp}} + \frac{1}{\nu} \left\| R^{-1} w \right\|_1 \tag{5}$$

This can be formulated as the following linear programming problem:

$$\begin{array}{ll}\min\limits_{(w,\gamma,y,v)\in R^{n+1+m+n}} & \nu e'y + e'v = \nu\sum\limits_{i=1}^{m} y_i + \sum\limits_{j=1}^{n} v_j \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & Rv \geq w \geq -Rv \\ & y \geq 0, \end{array} \tag{6}$$

At a solution of problem (4), v is the absolute value $|w|$ of w. This fact follows from the constraints $v \geq w \geq -v$ which imply that $v_i \geq |w_i|, i = 1\ldots, n$. Hence at optimality, $v = |w|$, otherwise the objective function can be strictly decreased without changing any variable except v. In this new formulation (4) we have at optimality that $Pv = |w|$, this is

$$|w_i| = \sum_{j=1}^{n} R_{ij}vj = \sum_{\{j|ri,j=1\}} R_{ij}vj \tag{7}$$

In other words, this means that the magnitude of the weight w_i of the related feature i, not only depends on itself but it also depends on all the features j that are related to i according to the relation function r. Moreover, R can be interpreted as a covariance matrix such that the prior over the vector of weights w is given by $P(w) = \propto \exp(\|R^{-1}w\|_1)$.

3 Materials

3.1 Subjects

The images we used for our experiments were taken from a concurrent study investigating the use of SPECT as a diagnostic tool for the early onset of AD. A detailed description of this data can be found in [37]. Subjects of four different centers, Edinburgh (Scotland), Nice (France), Genoa (Italy), and Cologne (Germany) were included for this study. In total 158 subjects participated, including 99 patients with AD, 28 patients suffering from depression (not used in this article), and 31 healthy volunteers. An example of this data is seen in Fig. 4. Confirmation of Alzheimer's disease was obtained by clinical follow-up. There was no statistically significant age difference between the AD patients and the healthy subjects. For technical acquisition related reasons images of 7 AD subjects had to be excluded.

3.1.1 Pre-processing

Applying the registration procedure as described above results in images of 128 by 128 by 89 voxels, with a voxelsize of 1.71 mm by 1.71 mm by 1.88 mm for all four centers. The SPECT images have an effective resolution of about 7 mm full-width at half-maximum (FWHM). Therefore, we can subsequently subsample the images a factor of two in each dimension by taking the average value over the subsampled areas without loosing much information. We only use the voxel

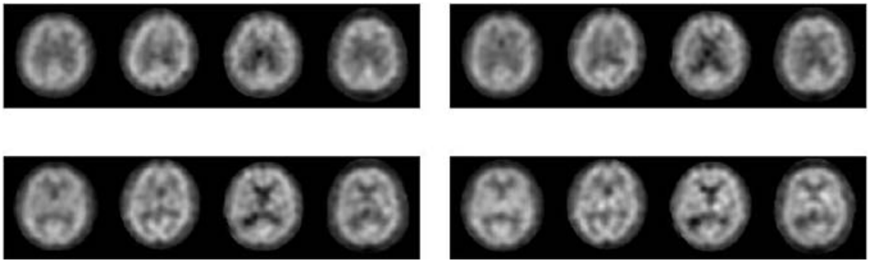

Fig. 4 Examples of four volumes from Cologne after intensity and spatial normalization. In each column the first two small images show two normal subjects, the last two images show slices of AD subjects. The sets of slices are ordered from left to right and from top to bottom. Strong hypo-perfusion can be seen for the first AD patient, whereas the hypo-perfusion is more subtle for the second patient

intensities for the voxels in the part of the brain that has been imaged for all subjects. Applying this procedure results in 3816 features per subject available for classification/feature selection.

3.1.2 Experts

All real images were rated in four categories (very probable, probably, probably not and very unlikely to have AD) by 16 European expert nuclear medicine physicians. The possible ratings were as follows: very probably Alzheimer's disease, probably Alzheimer's disease, probably not Alzheimer's disease and very unlikely Alzheimer's disease. To be able to compare the data from the experts with that of the automatic methods, we considered the first two ratings as positive and the other two as negative.

4 Experiments

In all of our experiments, we divided the data into two disjoint training and testing sets. The idea is to tune the parameters in our model only using data from the training set, once the final model is fixed, it is tested in the unseen testing set. We used leave-one-out cross validation to tune the model parameter ν of the contiguous SVM. Performance of our Contiguous SVM algorithm, in terms of generalization ability, is compared with a Fisher's Linear Discriminant (FLD) classifier as previously presented in [22]. The FLD algorithm used here is based on the FLD mathematical programming formulation introduced by Mika et al. [38]. For solving all the optimization problems involved in this paper we used the widely used commercial solver CPLEX 6.5 [39]. Next, we outline the results of our comparative testing. Two set of experiments were performed:

1. We randomly divided the 123 cases into 90 training examples and 33 testing examples, the goal of this experiment is to approximately measure the generalization capability of our proposed classifier.
2. In order to test the generalization performance of our approach across institutions, we divided the data into two different subsets according to the institution

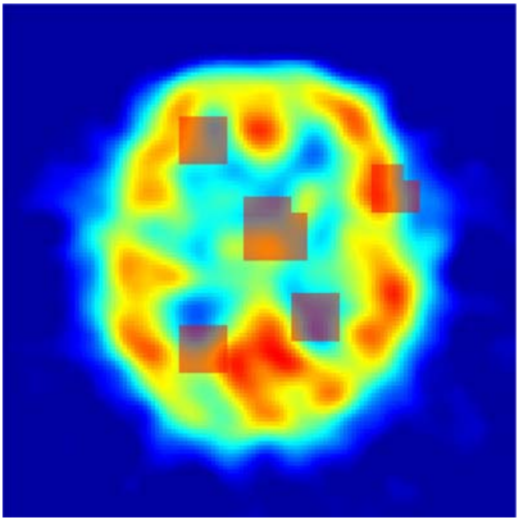

Fig. 5 A single axial image showing the regions picked by the algorithm overlayed on an image of an Alzheimer's disease patient SPECT image

where they were collected. The training set consists of 68 cases coming from Genoa (34 cases) and Cologne (34 cases) and the testing set consists of 55 cases coming from Edinburgh (28 cases) and Nice (27 cases).

The first experiment resulted in a selection of 253 features grouped in 7 connected areas. Figure 5 shows part of the selected features (a subset that can easily be visualized in 2D). Most selected groups of features are in the ventricles. This is consistent with the general atrophy of the brain observed in Alzheimer's disease patients which enlarges the ventricles relative to the other parts of the brain. This result shows the potential of the proposed approach at selecting meaningful grouped features which can be interpreted more easily than traditional feature selection approaches. The experts had an average sensitivity of 56.6% and a specificity of 82.4% for all 123 cases. In the SPM approach, we use SPM at a significance level of 0.1 at the cluster level. We consider each image where some significant clusters were found to be a positive result, this leads to a sensitivity of 55.9% and a specificity of 77.4% for SPM. Our classification approach as shown in Tables 1 and 4 outperforms both the experts and the SPM approach. Results in Table 4 show that even if the performance decreases on the training set due to

Table 1 Results for the first experiment for 90 training cases and 33 testing cases randomly sampled among the different institutions

	CSVM Sensitivity Specificity (%)	FLD Sensitivity Specificity (%)
Training	86.7	**88.7**
	80.0	65.0
Testing	**84.4**	82.0
	90.9	87.5

Note.The training results are based on leave-one out.

Table 2 Results for the second experiment

	CSVM Sensitivity Specificity (%)	FLD Sensitivity Specificity (%)
Training	**86.2**	84.6
	68.0	62.5
Testing	**72.5**	45.0
	93.0	**100.0**

Note. The classifier was trained on the data from Genoa (34 cases) and Cologne (34 cases), and tested on the data from Edinburgh (28 cases) and Nice (27 cases). The training results are based on leave-one out

differences in the way the images were aqcuired at the different institutions the contiguous SVM approach still shows good generalization capabilities.

5 Conclusion

Based on the experiments described in this article, we conclude that our automatic approach to the classification of images performs at least as well as human observers. In general, our contiguous support vector machine is more sensitive and more specific. One would need more data, especially of control subjects to be able to state that automatic methods always significantly outperform human observers in clinical practice.

We have shown that classification of images using the voxel values as features outperforms the local SPM approach. We have shown that classification without using any specific knowledge related to the pathology is possible.

The approach we propose in this article gives only a global decision based on a specific image. However, only providing global information might not be sufficient for clinicians. Therefore, we proposed a method that might do useful feature selection which might provide useful information to the clinician, at least at the group level. A trained classifier represents the group of images it was trained on, it does not show which areas where discriminative for any specific single image. Further research should focus on how to obtain subject specific local information while still retaining the advantage of a global approach.

We did not study the influence of the registration and intensity normalization steps. We used the registration methods that were readily available. Grova et al., and Grova [40, 41] carried out large-scale simulations of SPECT images, not for Alzheimer's disease but for epileptic patients. In future work one could test the influence of different registration and intensity normalization methods on the the classification approaches.

For future work one also might want to try the presented approach for differential diagnosis (other dementias versus Alzheimer's disease) which might be an even more important clinical issue. ROC analysis of the classifier as well as of the experts will be useful to better compare performances. This will also provide means to handle the differences in operating points for the different experts (e.g. some experts are more specific while others are more sensitive). Also an interesting future direction would be to extend the Contiguous SVM formulation, where a relation among datapoints is considered instead of a relation among the features. This approach can potentially be used for a general semi-supervised SVM approach where only some of the labels for the training data are available.

Acknowledgements We are indebted to Prof. Klaus Ebmeier, Prof. Nicholas Ayache and Prof. Jacques Darcourt for their large contributions to this work.

References

1. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984) Mental and clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of the Department of Health and Human Services Task Force on Alzheimer's disease. Neurology 34(7):939–944
2. Folstein MF, Folstein SE, McHugh PR (1975) "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12(3):189–198
3. Gosche KM, Mortimer JA, Smith CD, Markesbery WR, Snowdon DA (2002) Hippocampal volume as an index of Alzheimer neuropathology: findings from the Nun Study. Neurol 58(10):1476–1482
4. Goethals I, van de Wiele C, Slosman D, Dierckx R (2002) Brain SPET perfusion in early Alzheimer's disease: where to look? Eur J Nucl Med 29(8):975–978
5. Claus JJ, van Harskamp F, Breteler MM, Krenning EP, de Koning I, van der Cammen TJ, Hofman A, Hasan D (1994) The diagnostic value of SPECT with Tc 99m HMPAO in Alzheimer's disease: a population-based study. Neurol 44(3):454–461
6. Messa C, Perani D, Lucignani G, Zenorini A, Zito F, Rizzo G, Grassi F, Del Sole A, Franceschi M, Gilardi MC, et al (1994) High-resolution technetium-99m-HMPAO SPECT in patients with probable Alzheimer's disease: comparison with fluorine-18-FDG PET. J Nucl Med 35(2):210–216
7. Jobst KA, Smith AD, Barker CS, Wear A, King EM, Smith A, Anslow PA, Molyneux AJ, Shepstone BJ, Soper N, et al (1992) Association of atrophy of the medial temporal lobe with reduced blood flow in the posterior parietotemporal cortex in patients with a clinical and pathological diagnosis of Alzheimer's disease. J Neurol Neurosurg Psychiatry 55(3):190–194
8. Talbot PR, Lloyd JJ, Snowden JS, Neary D, Testa HJ (1998) A clinical role for 99mTc-HMPAO SPECT in the investigation of dementia? J Neurol Neurosurg Psychiatry 64(3):306–313
9. Van Gool WA, Walstra GJ, Teunisse S, Van der Zant FM, Weinstein HC, Van Royen EA (1995) Diagnosing Alzheimer's disease in elderly, mildly demented patients: the impact of routine single photon emission computed tomography. J Neurol 242(6):401–405
10. McMurdo ME, Grant DJ, Kennedy NS, Gilchrist J, Findlay D, McLennan JM (1994) The value of HMPAO SPECT scanning in the diagnosis of early Alzheimer's disease in patients attending a memory clinic. Nucl Med Commun 15(6):405–409
11. Kogure D, Matsuda H, Ohnishi T, Asada T, Uno M, Kunihiro T, Nakano S, Takasaki M (2000) Longitudinal evaluation of early Alzheimer's disease using brain perfusion SPECT. J Nucl Med 41(7):1155–1162
12. Minoshima S, Giordani B, Berent S, Frey KA, Foster NL, Kuhl DE (1997) Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. Ann Neurol 42(1):85–94
13. Ishii K, Sasaki M, Yamaji S, Sakamoto S, Kitagaki H, Mori E (1997) Demonstration of decreased posterior cingulate perfusion in mild Alzheimer's disease by means of H215O positron emission tomography. Eur J Nucl Med 24(6):670–673
14. Ibanez V, Pietrini P, Alexander GE, Furey ML, Teichberg D, Rajapakse JC, Rapoport SI, Schapiro MB, Horwitz B (1998) Regional glucose metabolic abnormalities are not the result of atrophy in Alzheimer's disease. Neurol 50(6):1585–1593
15. Braak H, Braak E (1997) Diagnostic criteria for neuropathologic assessment of Alzheimer's disease. Neurobiol Aging 18(4):S85–S88
16. Bobinski M, de Leon MJ, Convit A, De Santi S, Wegiel J, Tarshish CY, Saint Louis LA, Wisniewski HM (1999) MRI of entorhinal cortex in mild Alzheimer's disease. Lancet 353(9146):38–40

17. Rodriguez G, Vitali P, Calvini P, Bordoni C, Girtler N, Taddei G, Mariani G, Nobili F (2000) Hippocampal perfusion in mild Alzheimer's disease. Psychiatry Res 100(2):65–74
18. Dawson MR, Dobbs A, Hooper HR, McEwan AJ, Triscott J, Cooney J (1994) Artificial neural networks that use single-photon emission tomography to identify patients with probable Alzheimer's disease. Eur J Nucl Med 21(12):1303–1311
19. Hamilton D, O'Mahony D, Coffey J, Murphy J, O'Hare N, Freyne P, Walsh B, Coakley D (1997) Classification of mild Alzheimer's disease by artificial neural network analysis of SPET data. Nucl Med Commun 18(9):805–810
20. Frackowiak RSJ, Friston KJ, Frith CD, Dolan R (1997) Human brain function. Academic Press
21. Ashburner J, Friston K, Holmes A, Poline J-B (1999) Statistical parametric mapping, SPM'99. The Welcome Department of Cognitive Neurol, Institute of Neurol, University College London. Freely available at: http://www.fil.ion.ucl.ac.uk/spm
22. Stoeckel J, Malandain G, Migneco O, Koulibaly PM, Robert P, Ayache N, Darcourt J (2001) Classification of SPECT images of normal subjects versus images of Alzheimer's disease patients. In: Niessen WJ, Viergever MA (eds) Proceedings of the 4th international conference on medical image computing and computer-assisted intervention (MICCAI'01), Utrecht, The Netherlands. Lecture notes in computer science, vol 2208, pp 666–674
23. Brown LG (1992) A survey of image registration techniques. ACM Comput Surveys 24(4):325–376
24. van den Elsen PA, Pol EJD, Viergever MA (1993) Medical image matching—a review with classification. IEEE Eng Med Biol 12(4):26–39
25. Maintz JBA, Viergever MA (1998) A survey of medical image registration. Med Image Anal 2(1):1–37
26. Hill DL, Batchelor PG, Holden M, Hawkes DJ (2001) Medical image registration. Phys Med Biol 46(3):R1–R45
27. Roche A, Malandain G, Pennec X, Ayache N (1998) The correlation ratio as a new similarity metric for multimodal image registration. In: Wells WM, Colchester ACF, Delp S (eds) Medical image computing and computer-assisted intervention (MICCAI'98), Boston, USA. Lecture notes in computer science, vol 1496, pp 1115–1124
28. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1997) Numerical recipes. The art of scientific computing, 2nd edn. Cambridge University Press
29. Prima S, Ourselin S, Ayache N (2002) Computation of the mid-sagittal plane in 3D brain images. IEEE Trans Med Image 21(2):122–138
30. Yonekura Y, Nishizawa S, Mukai T, Fujita T, Fukuyama H, Ishikawa M, Kikuchi H, Konishi J, Andersen AR, Lassen NA (1988) SPECT with [99mTc]-d,l-hexamethyl-propylene amine oxime (HM-PAO) compared with regional cerebral blood flow measured by PET: effects of linearization. J Cereb Blood Flow Metab 8(6):S82–S89
31. Schmidt K (2002) Against: can ROI methodology/normalised tissue activities be used instead of absolute blood flow measurements in the brain? Eur J Nucl Med 29(7):953–956
32. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
33. Bradley PS, Mangasarian OL (1998) Feature selection via concave minimization and support vector machines. In: Shavlik J (ed) Machine learning proceedings of the fifteenth international conference(ICML '98), San Francisco, CA. Morgan Kaufmann, pp 82–90. Available at: ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps
34. Mangasarian OL (1999) Arbitrary-norm separating plane. Oper Res Lett 24:15–23. Available at: ftp://ftp.cs.wisc.edu/math-prog/tech-reports/97-07r.ps
35. Evgeniou T, Pontil M, Poggio T (2000) Regularization networks and support vector machines. Adv Comput Math 13:1–50
36. Smola A, Bartlett PL, Schölkopf B, Schuurmann J (eds) (2000) Advances in large margin classifiers. MIT Press, Cambridge, MA
37. Soonawala D, Amin T, Ebmeier KP, Steele JD, Dougall NJ, Best J, Migneco O, Nobili F, Scheidhauer K (2002) Statistical parametric mapping of (99m)Tc-HMPAO-SPECT images for the diagnosis of Alzheimer's disease: normalizing to cerebellar tracer uptake. Neuroimage 17(3):1193–1202
38. Mika S, Rätsch G, Müller K-R (2000) A mathematical programming approach to the kernel fisher algorithm. NIPS, pp 591–597
39. ILOG CPLEX Division, 889 Alder Avenue, Incline Village, Nevada. CPLEX Optimizer, 2004.

40. Grova C, Janin P, Biraben A, Buvat I, Benali H, Bernard AM, Scarabin JM, Gibaud B (2001) Validation of MRI/SPECT similarity based registration methods using realistic simulations of normal and pathological data. In: Niessen WJ, Viergever MA (eds) Medical image computing and computer-assisted intervention (MICCAI'01). Lecture notes in computer science, vol 2208, pp 257–282
41. Grova C (2002) Simulations réalistes de donées de tomographie monophotonique (TEMP) pour l'évaluation de methodes de recalage TEMP/IRM utilisant des mesures de statistiques de similarité: application dans le contexte de la fusion de données en épilepsie. PhD thesis, Université de Rennes I

Glenn Fung received a B.S. degree in pure mathematics from Universidad Lisandro Alvarado in Barquisimeto, Venezuela, then earned an M.S. in applied mathematics from Universidad Simon Bolivar, Caracas, Venezuela, where later he worked as an assistant professor for 2 years. Later, he earned an M.S. degree and a Ph.D. degree in computer sciences from the University of Wisconsin-Madison. His main interests are optimization approaches to machine learning and data mining, with emphasis in support vector machines. In the summer of 2003, he joined the computer aided diagnosis group at Siemens, Medical Solutions in Malvern, PA, where he has been applying machine learning techniques to solve challenging problems that arise in the medical domain. His recent papers are available at www.cs.wisc.edu/ gfung.

Jonathan Stoeckel received a B.E. degree from Xi'an Jiao Tong University, Xi'an, China, in 1993 and an M.E. degree from Shanghai Jiao Tong University, Shanghai, China, in 1996. From 1997 to 1998, he did research work in the Data Mining Group at the School of Computing and Information Technology, Griffith University, Brisbane, Australia. He is currently a Ph.D. student at the Department of Computer Science, Dartmouth College, USA. His research interests include data mining, multimedia, database and software engineering.