Classifying Cognitive Tasks to fMRI Data Using Machine Learning Techniques

David R. Hardoon
University of Southampton
School of Electronics and Computer Science
ISIS Research Group
Southampton, UK SO17 1BJ
E-mail: drh@ecs.soton.ac.uk

Larry M. Manevitz
University of Haifa
Department of Computer Science
Haifa, Israel 31905
E-mail: manevitz@cs.haifa.ac.il

***Abstract*— Using Machine Learning Tools (Neural Networks and Support Vector Machines) we show how raw fMRI brain scan data can be correctly assigned to cognitive tasks.**

We describe experiments classifying visual and motor tasks using one-class and two-class labeling for training. No a priori knowledge (e.g. of anatomy or physiology) is needed for the system to work.

These results further suggest that feature reduction techniques may allow for the automatic location of brain areas correlated with specific cognitive tasks - possibly even when the needed features are not localized.

I. Introduction

Functional magnetic resonance imaging (fMRI) [1] is an imaging technique which can be used in principle to map different sensor, motor and cognitive functions to specific regions in the brain. fMRI allows the carrying out of specific non-invasive studies within a given subject while providing an important insight to the neural basis brain processes (Figure 1 shows a slice overlying of fMRI scans). Neurons, which are the basic functional unit of the brain, consume a higher level of oxygen when active, hence blood with a higher level of oxygenation is supplied to those active neurons. fMRI makes indirect use of this effect by detecting areas of the brain which have an elevated consumption of oxygen. This effect can be used to identify areas of the brain associated with specific functions.

The current methodology used to identify such regions is to compare, using various mathematical techniques [2], [3], the elevation of oxygen consumption during a task with that used during a resting state.

We are interested here in the *inverse* problem - given the entire fMRI data, to classify the cognitive task the subject was engaging in. This is a challenging task since, amongst other things, the subject may be engaged in a variety of tasks; the dimension of data (i.e. the number of pixels in the fMRI) is enormous; and without more information, the signal to noise ration may be quite poor. (That is, probably most dimensions are irrelevant to a classification.)

One can think of this as a standard classification problem; and in this context one can use either clustering, one-class or two-class techniques.

From the standard two-class perspective, [4] applied machine learning techniqes to this problem, when considering the classification of the cognitive state of a human subject. [5] presented initial work on this problem using Kernel Canonical Correlation Analysis (KCCA) [6], [7]. Thus, in order to determine the elevation of oxygen consumption during a task, images acquired during a resting state are required for the second class. In order to keep the alternation between activity, a reference time-course is needed, where the resting and active states are embedded. A commonly used reference time-course is the square-wave time-course as plotted in Figure 2.

[8] first considered the problem of identifying fMRI scans that have only been acquired during the “active” state, i.e. scans acquired during the duration when the human subject has performed the given task. In machine learning terminology, this is called ”one-class” classification, because the learning method is trained solely with positive information. The basic intuitions are that, if available, two-class classification should perform better; although not always [9]. However, as is the case under consideration here, often we have some reasonable sampling of the positive examples; i.e. the distribution of positive examples can be estimated;

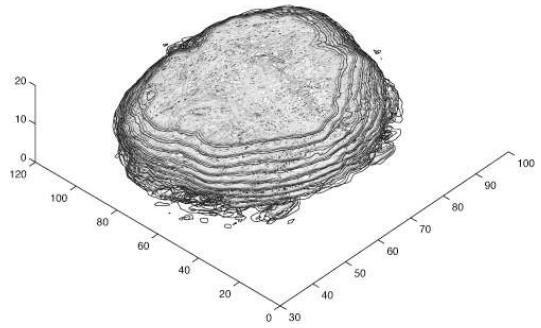

Fig. 1. Overlaid fMRI slices of the brain.

while the negative examples are either non-existent or episodic; i.e. not necessarily representative.

Obtaining good results under this assumption is known to be quite challenging [10], [11], [12], [13]; nonetheless it is often the most realistic assumption. Moreover, although not addressed in this paper, in principle one can imagine combining the one-class approaches with clustering methods which would allow the development of classification without any a priori labeling. We hope to address this issue in a later paper.

For the fMRI classification described above, this problem is particularly non-trivial as we expect the data to be of very high dimension and extremely noisy, as the brain concurrently works on many given tasks. It is also quite natural to assume that there is only representative data of the task of interest; and not necessarily representative data of the negation of this task thus making the one-class learning techniques appropriate.

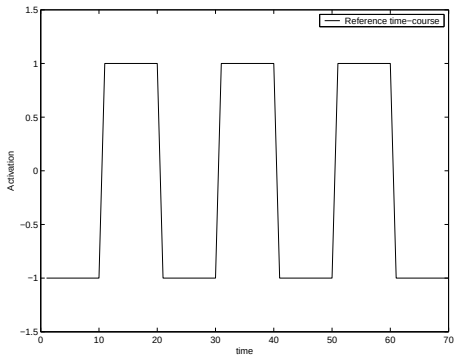

Fig. 2. The commonly used square-wave reference time-course.

In this work, we investigated both one-class and two-class learning regarding data involving both motor tasks (where we imagine important features to be in the motor cortex) and visual tasks (where we imagine important features to be in the visual cortex). These data sets were generously provided by Ora Friman (currently of Harvard Medical School) for the motor task and by Rafi Malach (of the Weizmann Institute) for the visual task.

An earlier report on this work (involving part of the one-class results only) was reported in [8].

The paper is organized as follows: Section II describes the methods and results obtained using one-class methodologies. Section III does the same for the two-class case. Section IV discusses possible extensions of our work, and mentions how it might be useful for the opposite problem of identifying brain features related to specific cognitive tasks. Section V summarizes our conclusions.

II. One-Class Results

We used two major one-class learning techniques - ”bottleneck” or compression neural networks [10] and a common version of the one-class Support Vector Machine (SVM) [14], [10] on brain slice data obtained from fMRI obtained while a subject is doing a simple motor (”finger lifting”) task. We point out that we use the entire brain slice, with no pre-filtering - i.e. the data is the entire slice, labeled with the task. [1] In addition, since we use data where there was, in fact, two-class labeling, we use this to illustrate the difference in the two methodologies, and how much classification ability is lost.

A. One-Class Methods

We use two techniques for the one-class approach. The first one is the compression neural network method [15], [16], [10]. We apply a design of a feed-forward neural network where in order to accommodate the usage of only positive examples we use a “bottleneck”. A bottleneck feed-forward network has the assumption that the images are represented in a m dimensional space where we choose a three level network with m inputs, m outputs and k neurons on the hidden level, where

[1] In early simulations because of computational limitations, we manually reduced the brain to one quadrant, where the motor cortex is known to lie. This reduction increased the efficacy of the methods presented here, for example, lifting the classification of the compression neural network for the motor data. This suggests that further research in feature reduction will improve all of the results.

$m > k$. Figure 3 gives a graphical example of the bottleneck network. This network is then trained using the standard back-propagation to learn the identity function on the sample example [10]. Thus the architecture of the bottleneck neural network used, is that of a feed-forward one with three layers, an input, hidden and output layer. All the neurons used were standard sigmoids and initial weights were chosen as small random values. We have used the standard back-propagation in the Neural Networks Toolbox in Matlab, where we have trained for 20 epochs which we observed avoids overfitting.

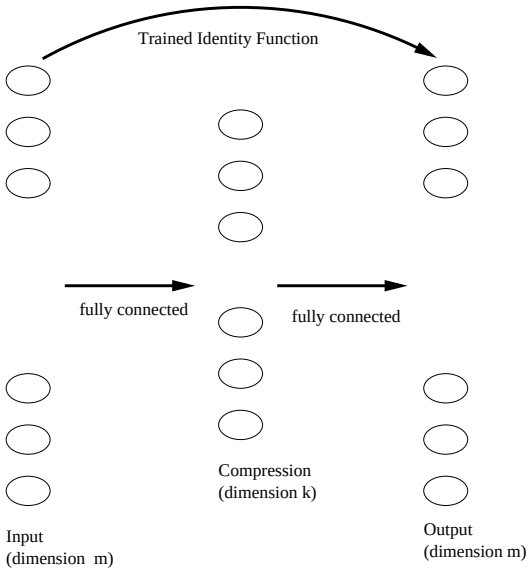

Fig. 3. **Bottleneck NN Architecture**

[10] showed that much thought is needed for selecting a good threshold procedure. [16] has suggested a heuristic approach to the threshold selection using only the positive information. This is done by training the network for some predetermined number of epochs and to relax the maximal error obtained by some percentage. [10] have tested this approach with poor results, and have suggested a similar method that is opposite to [16]. Instead of relaxing the maximal error obtained on the training they tighten the threshold by an amount heuristically related to the percentage of near zero vectors in the training set. In this work we suggest a different approach. Experimentally we found that the error during training exhibits a behaviour of having two spikes of high error, whereas following the second spike the error reduces to near zero. We thus take the threshold as the value of the error following the second spike. In figure 4 the error during training and average error on testing is plotted. We are able to observe that the average error on testing is roughly the same as the value following the second spike.

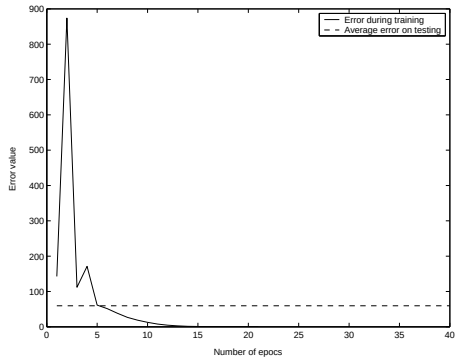

Fig. 4. **Error during training and average testing error.**

The second method used is the one-class Support Vector Machine (SVM) method [14]. Under this method, instead of separating positive and negative samples in the kernel feature space, as in standard (two-class) SVM, the origin is the only negative sample and therefore the method separates the positive samples from the origin via using relaxation parameters in SVM. We use the OSU-SVM 3.00 package[2] for Matlab for the one-class SVM experiments. Full details regarding the approaches can be found in [14], [17].

B. Two-Class methods

To discuss the difference in performance between one-class and two-class methods on the motor task, we also used a regular two-class SVM on this data. We used the OSU-SVM 3.00 package with the default settings for two-class SVM experiments.

C. Motor Task Experimental Protocols

The fRMI scans are of a volunteer[3] flexing their index finger on the right hand inside a MR-scanner while 12 image slices of the brain were obtained from a T2*-weighted MR scanner. Figure 5 gives an example to the extracted slices from the brain. The time-course reference of the flexing, as plotted in Figure 2, is built from the subject performing a sequence of 20 total actions and rests consisting of rest, flex, rest, ... flex (an example for the rest, flex sequence on the MRI images is given in Figure 6). Two hundred fMRI scans

[2]OSU SVMs Toolbox http://www.ece.osu.edu/∼maj/osu_svm/

[3]Provided by Ola Friman [18].

are taken over this sequence; ten for each action and rest. The individual fMRI images are dicom[4] format of size 128×128. Each image is labelled as either 1 (active) or -1 (inactive). The labelling was done manually at the time of the scans.

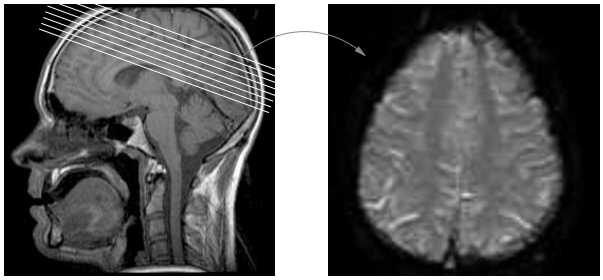

Fig. 5. Extracted slices of the brain.

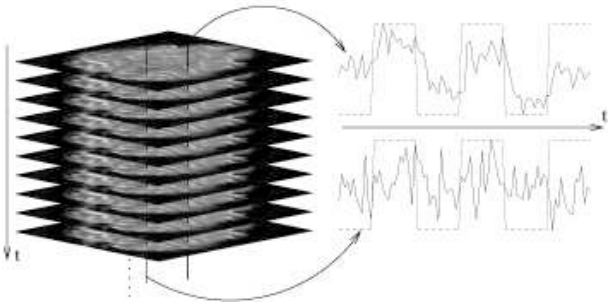

Fig. 6. The time-sequence with the images sequence.

Thus, in our data we have 100 positive and 100 negative images for each of the 12 slices. For the bottleneck neural network 80 positive samples were chosen randomly and presented for training and 40 samples, consisting of the remaining 20 positive and 20 random negative samples, were used for testing. This experiment was redone with ten independent random runs. The limitation to 20 negative samples out of a possible 100 was chosen to keep the testing fair between the positive and negative classes. We manually cropped the non-brain background from the scans; resulting in a slightly different input/output size for each slice of about $8,300$ inputs and outputs.

The compression percentage arising from the bottleneck was chosen by experimenting with different possible values. Table I shows some typical results. A uniform compression of about 60% gave the best results. for the hidden layer. The irrelevant (non-brain) image data was cropped for each slice resulting in a slightly different input/output size for the network for each slice.

TABLE I

BOTTLENECK COMPRESSION COMPARISON

Method	Result on slices	Compression
BN - NN	$56.19\% \pm 1.26\%$	60%
BN - NN	$56.02\% \pm 0.89\%$	70%
BN - NN	$54.79\% \pm 0.90\%$	80%

Thus a typical network had an architecture of about $8,300$ (input level) $\times$ about $2,500$ (compression level) $\times$ $8,300$ (output level). The network was trained to the identity using 20 epochs on the above chosen data. Following training the network was used as a classification filter, with an input value being classified as positive if the error level was lower or equal to a threshold as defined in the previous section and classified as negative. We used the same protocol in a one-class SVM.

Additionally, we used the two-class SVM where we randomly selected 160 training images and the remaining 40 for testing. This was also repeated 10 times.

D. Motor Task Experimental Results

We performed this experiment twice; by running another fMRI session on the same individual performing the same task. We report the results of each session separately.

The obtained results are an average over all the slices. Each slice was averaged over 10 repeats where in each repeat a random split of training testing was selected. Both SVM classifiers were used in their default setting as set by the OSU-SVM 3.00 package with a linear kernel with $C = 1$ and a radial based (RBF) kernel with $\gamma = 1$, the one-class SVM was used with $\nu = 0.5$. In addition, the two-class SVM was used with the unnormalised data as we have experimentally found that when the data was normalised, as with the other methods, with the two-class SVM the overall results were significantly worse.

1) Session 1: In Table II the SVM results for the linear and RBF kernel are presented, we are able to observe that while the one-class SVM performs better with the RBF kernel, the two-class SVM is better with the linear kernel.

[4]For information regarding dicom see http://medical.nema.org/

TABLE II

SVM RESULTS.

Method	Linear kernel	RBF kernel
One-class SVM	$49.12\% \pm 0.86\%$	$59.18\% \pm 1.47\%$
Two-class SVM	$68.06\% \pm 2.10\%$	$44.70\% \pm 1.12\%$

In Table III we compare the one-class to two-class techniques for the motor task. As initially expected we are able to observe that the two-class approach outperforms those of the one-class. The one-class SVM is slightly better then the bottleneck compression NN. We further analyse the statistics of the methods i.e. the

TABLE III

METHODS SUCCESS RESULTS.

Method	Result on slices
BN - NN	$56.19\% \pm 1.26\%$
One-class SVM	$59.18\% \pm 1.47\%$
Two-class SVM	$68.06\% \pm 2.10\%$

separation of the classified samples to their true classes. In Table IV[5] we compute and show the statistics of the fMRI images of the Positive samples that were classified as positive, denoted as true-positive, and the positive samples that were classified as negative, denoted as false-negative. While in Table V the statistics of the negative fMRI images samples that were classified as negative, denoted as true-negative, and those that were classified as positive, denoted as false-positive, are presented. We observe in Table IV that the compression NN is able to find a higher rate of true-positive fMRI images then the one-class SVM and the two-class methods even though they have obtained an higher overall success rate. In Table V we observe that the two-class methods perform better then the one-class. This is expected as the one-class methods make no use of the negative samples and eminently will have a lower ability in classifying it.

TABLE IV

METHODS STATISTICS - POSITIVE TESTING SAMPLES

Method	True-Positive	False-Negative	std
BN - NN	78.96%	21.04%	$\pm 3.15\%$
One-class SVM	72.83%	27.17%	$\pm 1.98\%$
Two-class SVM	71.55%	28.45%	$\pm 3.21\%$

[5]std stands for Standard Deviation

TABLE V

METHODS STATISTICS - NEGATIVE TESTING SAMPLES

Method	True-Negative	False-Positive	std
BN - NN	33.42%	66.58%	$\pm 3.45\%$
One-class SVM	39.25%	60.75%	$\pm 3.25\%$
Two-class SVM	65.64%	54.46%	$\pm 3.02\%$

2) Second Session: Corroborating our results by running the same experiments on another fMRI session of the same individual performing the same task as described above. The experiments have been run with the same configurations of the compression NN and one/two-class SVM. Table VI show the success rate in correctly classifying the fMRI scan of the second session. We find that compression NN is slightly better then the one-class SVM by $\approx 4\%$. Tables VII and VII give the statistics of the positive and negative testing samples. We are able to observe that even though the one-class SVM is able with a higher rate to correctly classify the positive scans, its ability to distinguish the negative from the positive is much lower then the compression NN.

TABLE VI

METHODS SUCCESS RESULTS ON SECOND SESSION.

Method	Result on slices
BN - NN	$58.92\% \pm 2.03\%$
One-class SVM	$54.81\% \pm 1.18\%$
Two-class SVM	$69.56\% \pm 4.12\%$

TABLE VII

METHODS STATISTICS (SECOND SESSION) - POSITIVE TESTING SAMPLES

Method	True-Positive	False-Negative	std
BN - NN	72.96%	27.04%	$\pm 4.06\%$
One-class SVM	84.96%	15.04%	$\pm 2.04\%$
Two-class SVM	72.49%	27.51%	$\pm 3.59\%$

TABLE VIII

METHODS STATISTICS (SECOND SESSION) - NEGATIVE TESTING SAMPLES

Method	True-Negative	False-Positive	std
BN - NN	44.88%	55.12%	$\pm 3.82\%$
One-class SVM	24.67%	75.33%	$\pm 3.24\%$
Two-class SVM	67.51%	32.49%	$\pm 2.17\%$

III. Two-Class Results on a Visual Task

In this section we present initial work done on a more complicated visual task where fMRI scans of 4 volunteers[6] watching five different categories of images while 58 image slices of their brain were taken in the MRI machine. The categories are of; Faces, Houses, Patterns, Objects and Blank. The different category images were displayed in alternating order, 7 repetitions for 3 time points each. Altogether 21 time points (images) per slice. The blank scene was shown to the volunteer in the start of the experiment for 6 time points and in-between repetitions and alternations of the main categories for 2 time points (a total of 56 time points). The over all time point length of MRI scans is 147. The individual fMRI images are dicom format of size 40×46. (In the data available part of the brain was not scanned.) Unlike the motor task described in section II, we used *all* of the slices together as one data point. Thus the dimension of a data point is in principle about $106,000$. [7]

We did two separate analyses of the data; once training between a specific category and blank for a specific subject; and once combining all three subjects into one data set and training between the specific categories and blank. All training was for specific categories versus blank. We used one subject, "A", to find the global SVM penalty parameter C. We then used this parameter for the other subjects and did not use the data of "A" again.

For the first analysis case, we had 21 positive labels and 63 negative ones for each subject; while for the second case we had 63 positive labels and 189 negative ones.

Each analysis was rerun 10 times with a random permutation of the training-testing split.

The results for the first analysis can be seen in Table IX while the results for the second analysis can be seen in Table X.

The results show a success rate of about 90% for each category trained for separate individuals and close to the same rate for the combined analysis.

IV. Discussion

A. One-Class Results and Methods

The classification results trained with either the Bottleneck Neural Network or the One Class SVM are on the one hand, substantially above random, and thus show that these methods can indeed be trained to find the information for these tasks.

On the other hand, the results (about 60% accuracy) are not yet sufficient for practical application. Since, for many tasks, it is unreasonable to expect to have the neat negative examples, as we had in the Motor Class protocol, it is important to find ways to leverage these results. Moreover, if we in fact were to take arbitrary negative slices (NOT from the protocol), one should expect that the advantage of the two-class would decline.

Looking at the results of the NN and the One-Class SVM, it is striking that they are quite successful at learning the positive class; but not as successful as ruling out the negative one.

B. Feature Selection, Applications to Brain Mapping and Other Future Work

It is important to emphasize that the two cognitive tasks were done in different ways. The Motor Task data were analyzed with separate slices; i.e. not as the full three dimensional brain.

This was done because of limitations of both data and computational ability at the time. The results reported are thus the average results over the different slices. In juxtaposition to that, the visual task used the entire three dimensional brain.

Looking over (not reported here) the data from the slices, there is a big variance between the results from the separate slices. This is to be expected, since it is quite possible that some of the slices have very few features that are in fact related to the task. Thus those levels should be only slightly above random choice, which is in fact what was observed. One could alleviate this by reporting only the maximal result and assume that as a by-product the machine learning is also picking out the appropriate levels.

However, there is no reason to assume that the features are in fact located in a specific slice; or certainly in the slices that were available.

This situation suggests using the machine learning to narrow in on the significant features. There are several ways to do this; one method currently under development will remove areas of the brain; redo the learning and then see if there is a loss in the results. If there is not, then that part of the brain, with all of its features can be safely eliminated. A binary search can then be used to "focus in" on the areas which are pertinent. This method can be combined with different focusing strategies and we will

[6] Provided by Rafael Malekh [19], [20]

[7] In actual fact, on the data supplied, part of the brain was not scanned, so the actual dimension used was about 53,000.

TABLE IX

SEPARATE INDIVIDUALS - SVM PARAMETERS SET BY SUBJECT *A*

	Face	Pattern	House	Object
Subject B	83.21% ± 7.53%	87.49% ± 4.20%	81.78% ± 5.17%	79.28% ± 5.78%
Subject C	86.78% ± 5.06%	92.13% ± 4.39%	91.06% ± 3.46%	89.99% ± 6.89%
Subject D	97.13% ± 2.82%	93.92% ± 4.77%	94.63% ± 5.39%	97.13% ± 2.82%

TABLE X

COMBINED INDIVIDUALS - SVM PARAMETERS SET BY SUBJECT *A*

	Face	Pattern	House	Object
B & C & D (combined)	86.00% ± 2.05%	89.50% ± 2.50%	88.40% ± 2.83%	89.30% ± 2.90%

hope to report on various experiments in this direction shortly.

We hope that such a method of eliminating features will allow a substantial boosting of the results.

In addition, we envisage the possibility of using such a search to discover appropriate areas pertinent to various cognitive tasks - that is, we hope in this way to also use the machine learning tools on the *opposite* task, automatically locating areas of the brain related to specific cognitive tasks. Note that, in principle, such areas do not need to be spatially compact; which no current techniques can find.

C. Future Work

- We feel that further investigation on automated feature reduction might be fruitful.
- Comparison of the same individual across sessions.
- Compare training between distinct active lables in the visual task.
- Further comparison of training across individuals.
- Technically, we feel that further work can be done in the threshold selection for the compression neural network.

V. CONCLUSIONS

- One class classification can be done, even with the ”noisy” data and even with the full slices of the brain scan.
- Comparable results (about 58% accuracy) were obtained under both one-class SVM and Compression-Based Neural Network techniques.
- Two class classification on Visual data using standard SVM techniques results in close to 90% accuracy.
- We have proposed methods to bootstrap our results which we will apply in future work.

ACKNOWLEDGMENT

We thank Ola Friman of the Harvard Medical School and Rafael Malach of the Weizmann Institute for their very generous sharing of their data. Ola Friman also provided Figures 5 and 6). We also are indebted to Sharon Gilaie-Dotan and Hagar Gelbard of Malach's laboratory for many patient explanations regarding the details of the data.

This collaboration was supported by the *Caesarea Rothschild Foundation*, by the *HIACS* Research Center and the *Neurocomputation Laboratory* of the University of Haifa. The first author is funded by the European project LAVA num. IST-2001-34405 and the PASCAL network of excellent num. IST-2002-506778.

REFERENCES

[1] A. Parry and P. M. Matthews, “Function magnetic resonance imaging (fmri): A “window” into the brain,” 2002.

[2] A. McIntosh, F. Bookstein, J. Haxby, and C. Grady, “Spatial pattern analysis of functional brain images using partial least square,” 1996. [Online]. Available: http://citeseer.nj.nec.com/mcintosh96spatial.html

[3] G. Aguirre, E. Zarahn, and M. D’Esposito, “The variability of human, BOLD hemodynamic responses,” *NeuroImage*, vol. 8, no. 4, pp. 360–369, 1998.

[4] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman, “Learning to decode cognitive states from brain images,” *Machine Learning*, vol. 1-2, pp. 145–175, 2004.

[5] D. R. Hardoon, J. Shawe-Taylor, and O. Friman, “KCCA for fmri analysis,” in *Proceedings of Medical Image Understanding and Analysis*, London, UK, 2004.

[6] F. Bach and M. Jordan, “Kernel independent component analysis,” *Journal of Machine Leaning Research*, vol. 3, pp. 1–48, 2002.

[7] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, pp. 2639–2664, 2004.
[8] D. R. Hardoon and L. M. Manevitz, "fmri analysis via one-class machine learning techniques," in *Proceedings of IJCAI-05*, 2005, p. to appear.
[9] N. Japkowicz, "Are we better off without counter examples?" in *Proceedings of the First International ICSC Congress on Computational Intelligence Methods and Applications*, 1999, pp. 242–248.
[10] L. Manevitz and M. Yousef, "One-class svms for document classification," *Journal of Machine Learning Research 2*, pp. 139–154, 2001.
[11] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *SIGKDD Explorations*, vol. 6(1), pp. 40–49, 2004.
[12] M. Yousef, "Document classification using positive examples only," Ph.D. dissertation, University of Haifa, 2000.
[13] H. Schwenk and M. Milgram, "Transformation invariant autoassociation with application to handwritten character recognition," *Advances in neural information processing systems*, vol. 7, pp. 991–998, 1995.
[14] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," Microsoft Research., Technical Report 99-87, 1999.
[15] G. W. Cottrell, P. Munro, and D. Zipser, "Image compression by back propagation: an example of extensional programming," *Advances in Cognitive Science*, vol. 3, 1988.
[16] N. Japkowicz, C. Myers, and M. A. Gluck, "A novelty detection approach to classification," *IJCAI*, pp. 518–523, 1995.
[17] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
[18] O. Friman, "Adaptive analysis of functional mri data," Ph.D. dissertation, Linkoping Studies in Science and Technology, 2003.
[19] I. Levy, U. Hasson, G. Avidan, T. Hendler, and R. Malach, "Center-periphery organization of humand object areas," *Nature Neuroscience*, vol. 4(5), pp. 533–539, 2001.
[20] U. Hasson, M. Harel, I. Levy, and R. Malach, "Large-scale mirror-symmetry organization of human occipito-temporal objects areas," *Neuron*, vol. 37, pp. 1027–1041, 2003.