NeuroImage 56 (2011) 809–813

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Machine learning classification with confidence: Application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression

Ilia Nouretdinov [a,b], Sergi G. Costafreda [a], Alexander Gammerman [b], Alexey Chervonenkis [b], Vladimir Vovk [b], Vladimir Vapnik [b], Cynthia H.Y. Fu [a,*]

[a] Institute of Psychiatry, King's College London, London, UK
[b] Computer Learning Research Centre, Department of Computer Science, Royal Holloway, University of London, London, UK

ARTICLE INFO

Article history:
Received 30 November 2009
Revised 28 April 2010
Accepted 11 May 2010
Available online 17 May 2010

ABSTRACT

There is rapidly accumulating evidence that the application of machine learning classification to neuroimaging measurements may be valuable for the development of diagnostic and prognostic prediction tools in psychiatry. However, current methods do not produce a measure of the reliability of the predictions. Knowing the risk of the error associated with a given prediction is essential for the development of neuroimaging-based clinical tools. We propose a general probabilistic classification method to produce measures of confidence for magnetic resonance imaging (MRI) data. We describe the application of transductive conformal predictor (TCP) to MRI images. TCP generates the most likely prediction and a valid measure of confidence, as well as the set of all possible predictions for a given confidence level. We present the theoretical motivation for TCP, and we have applied TCP to structural and functional MRI data in patients and healthy controls to investigate diagnostic and prognostic prediction in depression. We verify that TCP predictions are as accurate as those obtained with more standard machine learning methods, such as support vector machine, while providing the additional benefit of a valid measure of confidence for each prediction.

Introduction

Using support vector machine algorithms (Vapnik, 1995), we have been investigating potential neuroimaging markers for psychiatric disorders (Fu et al., 2008a; Marquand et al., 2008; Costafreda et al., 2009a,b). As a diagnostic marker of depression, we found 86% accuracy in identifying individual patients from the functional MRI pattern of brain activity to sad faces (Fu et al., 2008a), while the neural features of verbal working memory, as expected, showed reduced diagnostic accuracy (Marquand et al., 2008). To investigate predictive markers of clinical response, MRI data were acquired in patients experiencing an acute depressive episode who were not taking any medication and before they had begun therapy. In prospective studies, patients then received treatment with antidepressant medication (Fu et al., 2004, 2007; Walsh et al., 2007) or individual psychotherapy, i.e. cognitive behavioral therapy (Fu et al., 2008b). We found that structural MRI features were highly predictive of an individual patient's full clinical response to antidepressant medication with an accuracy of 89% (Costafreda et al., 2009a), while functional MRI responses showed the greatest predictive potential for cognitive behavioral therapy with an accuracy of 79% (Costafreda et al., 2009b). Machine learning analysis of MRI data has also demonstrated high diagnostic accuracy in other disorders, such as Alzheimer's disease (Klöppel et al., 2008) and schizophrenia (Davatzikos et al., 2005). Together, these findings point to the potential of applying machine learning methods to achieve clinically useful diagnostic and prognostic neurobiomarkers based on the pattern of brain activity and structure in psychiatric disorders.

However, a significant obstacle in advancing machine learning algorithms in clinical practice lies in the type of output that they typically provide. Most algorithms only generate a categorical classification, e.g. a "yes" or "no" diagnosis. An essential requirement for clinical applications of machine learning predictions is a measure of the quality of the predictions, also referred to as the confidence of the classification output (Klöppel et al., 2008). While classical statistical methods may produce confidence levels, they are usually applicable to low-dimensional data and include specific assumptions. In the present study, we have adapted a general probabilistic classification method to establish measures of confidence for neuroimaging data.

We propose to apply transductive conformal predictors (hereby referred to as TCP or confidence machines) to generate confidence measures for neuroimaging-based predictions. We have summarized the description of the method presented in Vovk et al. (2005) and Gammerman and Vovk (2007). TCP is a novel method which has been successfully applied to the clinical diagnosis of cancer from proteomics data (Gammerman et al., 2008). TCP may be built using various machine learning algorithms including support vector machines and

* Corresponding author. Institute of Psychiatry, 103 Denmark Hill, P074, London SE5 8AF, UK. Fax: +44 207 848 0783.
E-mail address: cynthia.fu@kcl.ac.uk (C.H.Y. Fu).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.05.023

is thus a natural extension of our work to date (Fu et al., 2008a; Marquand et al., 2008; Costafreda et al., 2009a,b).

In the TCP approach, confidence measures are based solely on the randomness assumption, i.e. the assumption that the training and test examples are produced independently and identically from the same distribution, which is also referred to as the *iid* assumption. The premise of TCP is twofold: to try every possible prediction label as a candidate for a new example, such as each of two possible diagnostic labels: patient and healthy control, and to measure how well the resulting sequence of training and test examples *conforms* to the randomness assumption. When we assign the correct label to a new example, the randomness assumption is still satisfied in the resulting sequence. However, if we assign an incorrect label to a new example, the randomness assumption is no longer satisfied and the resulting sequence will appear "strange" (or "atypical" within the randomness assumption or simply "non-random").

The TCP algorithm measures the "strangeness" of the data sequence and computes a probability for each possible label which reflects how well each one conforms to the randomness assumption. If the probability for a given label is low, then the randomness assumption was wrong or a rare event has occurred, which leads to the rejection of that label. Therefore, TCP does not require assumptions other than iid, unlike statistical methods that compute probabilistic predictions based on specific parametric models which are often complemented by a prior distribution on the parameters. Specifically, TCP assigns an individual confidence to each new example that is equal to one minus the minimal significance level at which all prediction labels but one are rejected under the iid assumption, and the remaining label is the announced label. TCP yields accurate and reliable predictions that are complemented by quantitative measures of their quality. These measures can also be used to generate predictions at a desired confidence level which are well-calibrated, in the sense of controlling the probability of issuing erroneous predictions (Vovk, 2002).

In the present study, we present the theoretical motivation of TCP, and we have applied TCP to structural and functional MRI data in patients and healthy individuals to investigate diagnostic and prognostic prediction in depression. For diagnostic prediction, we employed a functional MRI dataset of implicit processing of sad facial expressions in patients with depression and matched healthy controls (Fu et al., 2004, 2008a). For prognostic prediction, we examined the structural MRI scans of these patients while they were in an acute depressive episode, prior to the initiation of any treatment, and clinical response was assessed prospectively following 8 weeks of treatment with an antidepressant medication (Fu et al., 2004; Costafreda et al., 2009a). We sought to develop an algorithm that: (1) provides classification for MRI images from a database, and (2) generates a confidence estimate for each classification output.

Materials and methods

Given a standard machine-learning problem: a *training set* of *examples* $(x_1,y_1),\ldots,(x1,y1)_n$ every example $z_i=(x_i,y_i)$ consists of its *object* x_i and its *label* y_i. We are also given a test object x_n, while the actual label y_n is withheld from us. Our goal is to predict the label y_n.

Confidence of predictions is obtained under the general independent and identically distributed (iid) assumption or randomness assumption (Vovk et al., 2005; Gammerman and Vovk, 2007). There is a stochastic mechanism that generates the example/label pairs $z_i = (x_i, y_i)$ independently of each other. Generally, high confidence in a prediction of y_n means that all alternative prediction labels are excluded under the iid assumption.

Transductive conformal predictor

Suppose, we have a non-conformity (strangeness) measure:

$$\alpha_i = A(z_i, \{z_1, \ldots, z_n\})$$

defined to be a function A of a finite set $\{z_1,\ldots,z_n\}$ and its element z_i (in this work it is included into the set for convenience), which is intuitively a measure of disagreement between them.

A non-conformity measure could be based on a standard method of prediction e.g. Nearest Neighbors, support vector machine (SVM), or a more general method. The output of TCP is valid for any used non-conformity measure. However, the quality of prediction will depend on the choice of this measure. If it is inappropriate, the confidence in prediction will be never high. The following algorithm summarizes the TCP approach. We use enumeration: *1,…, l* for the training examples and *n* for the test example. One can suppose $n = l + 1$ unless the same training set is being used for more than one test example. If there are several testing examples for a given training set, this algorithm can be repeated for each test example. Here we use the notation #{S} to refer to the cardinality, or number of elements, of a set S. Therefore in the following $\#\{i = 1,\ldots,l+1 : \alpha_i \geq \alpha_{l+1}\}$ refers to the number of elements in the set consisting of the training examples plus the test example with a strangeness measure α_i greater than or equal to the strangeness measure α_{l+1} of the testing example.

Algorithm 1. Transductive conformal predictor

Input: training data: $(x_1, y_1),\ldots,(x_l,y_l)$
Input: testing example: x_n
Input (optional): significance level: γ
Input: non-conformity measure: A
$z_1 = (x_1, y_1), \ldots, z_l = (x_l, y_l)$
for y in set Y of possible labels **do**
  $z_{l+1} = (x_n, y)$
  for i in $1,\ldots,l+1$ **do**
    $\alpha_i = A(z_i, \{z_1, \ldots, z_l, z_{l+1}\})$
  end for
  $p(y) = \frac{\#\{i = 1,\ldots,l+1 : \alpha_i \geq \alpha_{l+1}\}}{l+1}$
end for
Output (optional): prediction set $R_n^\gamma = \{y : p(y) > \gamma\}$
Output: single prediction $\hat{y}_n = \arg\max_y \{p(y)\}$
Output: confidence $conf(\hat{y}_n) = 1 - \max_{y \neq \hat{y}_n}\{p(y)\}$

Here $p(y)$ is the *p-value* associated with a hypothetical completion $y_n = y$. The *p*-values generated by TCP form a valid randomness test (in the sense that $P\{p \leq \gamma\} \leq \gamma$), under the iid assumption.

A standard method of packaging the prediction results for pattern recognition (and the only method for regression) is to choose a "significance level" $\gamma < 1$ and output the $(1-\gamma)$ *prediction set (region)*:

$$R_n^\gamma = \{y : p(y) > \gamma\}$$

The validity property implies that

$$y_n \notin R_n^\gamma$$

with probability at most γ under the iid assumption.

A "cumulative" (not requiring input γ) characteristic of the prediction is *confidence* defined as 1- the second largest *p*-value (which is the same as 1- the smaller *p*-value in a two-class problem). The minimal significance level at which the prediction region is a one-element or empty set (prediction is certain) is the complement of confidence to 1.

Non-conformity measures

A non-conformity measure is a function

$$\alpha_i = A(z_i, \{z_1, \ldots, z_n\})$$

where z_i is a pair (x_i,y_i). It is a measure of the "disagreement" between an element z_i and its set $\{z_1,\ldots,z_n\}$; in other words, it measures how

untypical this element is in comparison to the other elements in the set.

Non-conformity measures are usually based on an underlying method of prediction. For instance, a usual non-conformity measure can be based on the k-Nearest-Neighbours algorithm, by defining the non-conformity measure α_i which answers the question 'how many of k nearest neighbors of x_i in $\{x_1,\ldots,x_n\}$ have a label other than y_i'?

For MRI data, due to the high-dimensionality of the data, the classic approach for classification consists of two steps: (1) selection of features (attributes) according to some criterion (e.g. t-test), followed by (2) application of a pattern recognition method (e.g. SVM) to the data which is restricted to the selected features. In the present study, we define non-conformity measure A directly based on this approach, and we present an alternative method, non-conformity measure B, which dispenses with the SVM step by generating a non-conformity measure directly based on the output of the t-test.

Non-conformity measure A: t-test and SVM

We are interested in the relative non-conformity of the elements $z_1=(x_1,y_1),\ldots,z_n=(x_n,y_n)$ where examples x_i are m-dimensional vectors $x_i=(x_i(1),\ldots,x_i(m))$, while labels y_i are binary-valued.

For any feature $j=1,\ldots,m$ we can calculate the p-value from the two-sample t-test. t-Test checks whether two sample sets, $U_j=\{x_i(j)|y_i=0\}$ and $V_j=\{x_i(j)|y_i=1\}$, come from distributions with equal means. We then excluded all the features (voxels) j such that the p-value of the t-test was larger than a given threshold θ, and ran linear SVM training of two classes using only the remaining dimensions (significant for t-test at level θ). The SVM algorithm assigned an α_i value to each example, in which α_i is positive for support vectors and $\alpha_i=0$ for non-support vectors. As it does not depend on the order of examples, it can be used as the measure of non-conformity assigned to an example z_i.

Non-conformity measure B: t-test with additional spatial filtering without SVM

An element is considered atypical if its addition to the set prevents separability of two classes. This gives the motivation for an alternative non-conformity measure: non-conformity α_i of z_i is the number of voxels that lead to significant separability after exclusion of example z_i from the set.

In other words, for each $i=1,\ldots,n$ we need to take the remaining set $z_1,\ldots,z_{i-1},z_{i+1},\ldots,z_n$, consider $U_j^{-i}=\{x_h(j)|y_h=0,\ h\neq i\}$ and $V_j^{-i}=\{x_h(j)|y_h=1,\ h\neq i\}$ for each $j=1,\ldots,m$ and calculate the number α_i of voxels j such that U_j^{-i} and V_j^{-i} are significantly different for t-test at level Θ. Moreover, for MRI data, each voxel has 26 neighbours in the space. To reduce the number of occasional and isolated voxels, which from *a priori* information we suppose are likely to be spurious findings and therefore without predictive value, we calculated only those voxels which fit into this set together with all their neighbours.

Benchmark clinical MRI datasets

We have analyzed two datasets of functional and structural measurements which have demonstrated statistically significant classification of diagnostic and prognostic predictions in depression (Fu et al., 2008a; Costafreda et al., 2009a). Briefly, the functional diagnosis (FDx) dataset contained functional MRI data from 19 patients with depression and 19 matched healthy controls (Fu et al., 2004). Subjects underwent functional MRI scans while observing faces with varying levels of sadness. The inputs for the classifier were the subject-level contrast maps reflecting blood oxygenation level dependent (BOLD) signal changes relative to baseline (crosshair fixation) during the observation of standardised faces expressing increasing levels of sadness. As a demonstration, we present the results for low (Task 1) and high (Task 2) intensities of sadness. The FDx data set consisted of 38 examples and 2×197484 features (197484 intracerebral voxels of dimensions $2\times2\times2$ mm by 2 tasks). 19 examples were controls (label $y_i=1$), and 19 examples were patients (label $y_i=2$).

The structural prognosis (SPx) dataset consisted of structural MRI data in 18 patients with depression who subsequently received treatment with antidepressant medication. Following 8 weeks of treatment, 9 of the 18 patients achieved a full clinical remission based on standard clinical response criteria (Fu et al., 2004). The aim of the classification was to predict clinical response based on grey matter density as measured by voxel-based morphometry (Costafreda et al., 2009a), with each example consisting of 342,002 intracerebral $1\times1\times1$ mm voxel measurements and the associated label (responder or non-responder).

Classification results

We tested conformal predictions on the datasets and used cross-validation, whereby at each iteration one subject from each group was selected for testing (testing set) and the model was trained on the remaining subjects (training set). We determined the accuracy of their single predictions with the two different versions of the algorithm.

In the algorithm, single-valued prediction is done by largest p-value, so y_l is 2 if $p_l(2)\geq p_l(1)$ and 1 otherwise. An example with number $l=1,\ldots,n$ is misclassified if either $y_l=2$ and $p_l(1)>p_l(2)$, or $y_l=1$ and $p_l(2)\geq p_l(1)$. The threshold for the t-test was fixed at $\theta=0.005$, which is a threshold that has demonstrated its utility empirically in our previous work on neuroimaging-based classification (Costafreda et al., 2009a).

High classification accuracies were generated for diagnosis and prognosis from the functional and structural MRI datasets with both non-conformity measures A and B (Table 1). As a diagnostic marker of depression, BOLD responses to sad faces showed high classification accuracy with both measures A (comparable to: Fu et al., 2008a) and B (Supplementary Table A). TCP analysis revealed that increased BOLD responses activity in the dorsal anterior cingulate, superior frontal and precentral cortices during implicit processing of high-intensity sad faces classified patients with depression, while increased activity in the insula, angular cortex and cerebellum predicted healthy control status (Fig. 1). As a prognostic marker of clinical response, greater grey matter density in the anterior and posterior cingulate cortices predicted a full clinical response to treatment with antidepressant medication, analysed with non-conformity measure A (comparable to: Costafreda et al., 2009a) and B (Fig. 2, Supplementary Table B).

To use our method and build a region predictor for a given significance level γ (or equivalently, a confidence level 1-γ), one needs to output the set of all labels y with $p(y)>\gamma$. This region prediction (unless it is empty) will always predict the label with the highest p-value, and it may also hedge this prediction by including the alternative label if its p-value is greater than γ. We refer to *certain* predictions, as opposed to genuinely hedged predictions, as those in

Table 1
Accuracy results for the Diagnosis and Prognosis datasets analysed with TCP and non-conformity measures A (t-test and SVM) and B (t-test with additional spatial filtering).

Dataset	Task	Non-conformity measures	Accuracy	Specificity	Sensitivity
Diagnosis	1	A	76.3%	68.4%	84.2%
		B	73.7%	78.9%	68.4%
	2	A	81.6%	78.9%	84.2%
		B	86.9%	84.2%	89.4%
Prognosis	n/a	A	77.8%	77.8%	77.8%
		B	83.3%	88.9%	77.8%

The Diagnosis dataset consisted of functional MRI tasks displaying sad facial expressions of low intensity (Task 1) and high intensity (Task 2) in patients and healthy controls. The Prognosis dataset contained structural MRI of acutely depressed patients with depression who subsequently achieved a full or partial clinical response following treatment.

which the output of the region predictor contains only one label. Region predictions are valid. For example, at the confidence level of 90%, we would expect the accuracy of prediction to be on average at least 90%. Accordingly, region prediction achieved the expected levels of accuracy for given significance levels in our dataset. As can be seen from Table 2, increasing the confidence level results in a reduced number of certain predictions. An extreme is reached for the structural MRI dataset in which at the 95% confidence level the algorithm could not produce any certain predictions. Such result is due to the small sample sizes of the present structural dataset.

Table 2
Accuracy of conformal predictions at selected confidence levels based on functional and structural MRI data with TCP.

	Confidence Level		
	0.80	0.90	0.95
Diagnosis			
Functional MRI Task 1 ($n=38$)			
Certain Predictions	31	22	6
Percentage of Correct Predictions	81.6%	89.5%	97.4%
Functional MRI Task 2 ($n=38$)			
Certain predictions	36	20	9
Percentage of correct predictions	86.1%	92.1%	97.4%
Prognosis			
Structural MRI ($n=18$)			
Certain predictions	13	4	0
Percentage of correct predictions	82.4%	94.4%	100%

Certain predictions are those for which the output of the region predictor contained only one label.

Discussion

We have applied TCP to functional and structural MRI data in order to generate diagnostic and prognostic decisions at the individual level. We have found that TCP is as accurate as the usual "forced" predictions in our benchmark datasets. Moreover, the advantage of TCP for psychiatric classification is that they provide measures of confidence which are given to each diagnostic or prognostic decision and thus the risk of an erroneous clinical decision is known for a given *individual*. We have demonstrated the reliability of these confidence levels despite a relatively low number of examples.

Determination of the risk of error is essential for clinical applications of machine learning predictions. Furthermore, the risk itself may have various utilizations. For example, if predictions for every example are required, simple prediction along with its credibility (largest p-value) and confidence (one–2nd largest p-value) can be produced, so that the risk of the error is known, a means of "hedging" prediction. The confidence measure can be directly interpreted as confidence in a given diagnosis or prognosis. Alternatively, the risk of error may be controlled by predetermining an acceptable level of confidence for a given clinical decision. In other words, if the prediction must be certain or "solid," we can set the confidence level and output predictions to an appointed minimum significance level. This feature may be particularly useful in situations in which the cost of a clinical error is high.

In the present study, we have considered only two-class discrimination problems (i.e. depressed patients vs. healthy controls and responders vs. non-responders). However, clinically useful diagnostic and prognostic tools often require discrimination across a range of potential outputs which may be considered a multi-class discrimination problem, for example to distinguish between unipolar depression vs. bipolar depression vs. healthy controls. TCP is a method that was originally designed to provide probabilistic outcomes for such multi-class problems and can be directly applied to datasets containing comparable neuroimaging measurements from a range of psychiatric disorders. Moreover, the risk of an erroneous prediction can be controlled for within such multi-class problems. As a region predictor, TCP can generate the most likely diagnoses when there is not enough information to make a single one, which is often more akin to real-life diagnostic dilemmas. The TCP algorithm can also be modified to take into account prior probabilities of diagnosis or situations in which the cost of a false positive error is different from that of false negative errors by using label-specific weights in the strangeness measure. We can achieve this by replacing $\alpha_i = A(z_i, \{z_1,\ldots,z_n\})$ with its weighted version: $\alpha_i = A(z_i, \{z_1,\ldots,z_n\})/w(y_i)$ where $w(y)$ depends on the importance of class y. For example, in the case of two labels (1 and 2), if both have equal prior probability then we could use $w(1) = w(2) = 0.5$, while if label 1 is twice as likely as label 2 based on prior information, then we could apply $w(1) = 2/3$ and $w(2) = 1/3$. Weights can also be similarly used to reflect unequal costs of errors.

Non-conformity measures may be built using many powerful machine learning algorithms, including support vector machine, and are thus complementary to current prediction methods commonly used in neuroimaging studies. The present paper also suggests that conformal prediction can be applied without SVM, by using a simple non-conformity or "strangeness" measure adapted to the nature of the data, which in our example involved the introduction of spatial information in measure B. Furthermore, the application of TCP is not limited to between group classification as TCP may also be applied within individual subjects in studies of "mind reading" to determine the confidence of the predicted cognitive state.

A limitation of the present study though is the small sample sizes of the test datasets. Low sample sizes directly impact in the precision attainable for the p-values and confidence outputs. It would be desirable to utilize the method presented in this paper in larger samples, allowing us to further empirically verify with increased precision the confidence associated with each prediction.

In summary, inclusion of a measure of confidence is essential for the development of the clinical application of machine learning predictions in psychiatric disorders. We have proposed a general

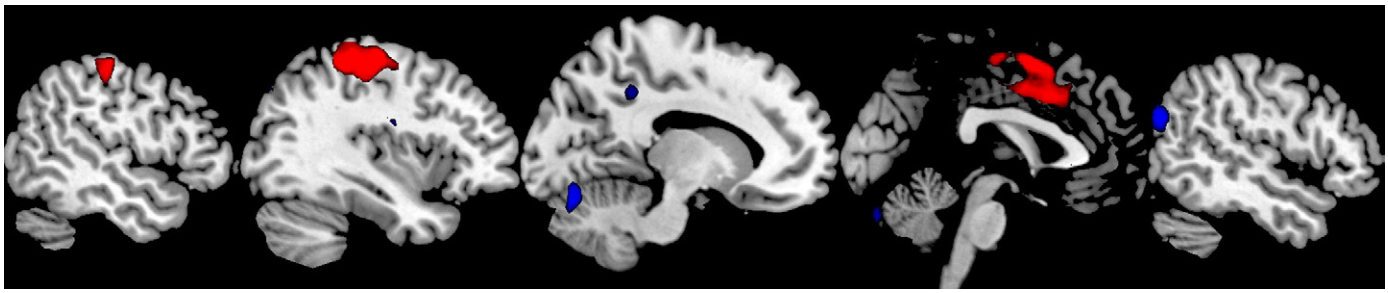

Fig. 1. Diagnostic classification of depression from functional MRI BOLD responses during implicit processing of sad facial expressions of high intensity. Using TCP, increased BOLD responses in the dorsal anterior cingulate, left precentral and left superior frontal gyrus (colored in red) were predictive of a diagnosis of depression, while higher BOLD responses in the left insular cortex, cerebellum and right angular gyrus (colored in blue) reduced the odds of such diagnosis. Representative sagittal views are presented in MNI space from $x = 48, -36, -12.0$ and $+48$.

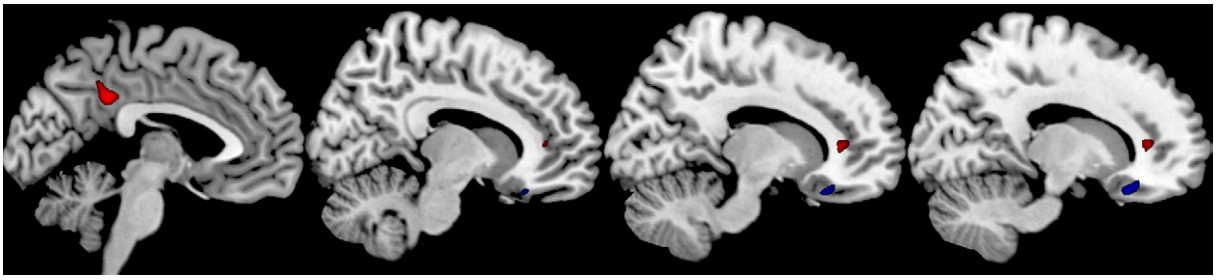

Fig. 2. Prediction of clinical response to antidepressant medication from grey matter density acquired from structural MRI. From the TCP analysis, greater grey matter density in the anterior and posterior cingulate cortices (colored in red) increased the probability of a full clinical response to the antidepressant medication fluoxetine, while greater grey matter density in the orbitofrontal cortex (colored in blue) increased the probability of having residual symptoms following treatment with antidepressant medication. Representative sagittal views are presented in MNI space at $x = -4$, 10, 12 and 14.

probabilistic classification method to produce measures of confidence for MRI data. Transductive conformal predictors generated significant diagnostic and prognostic classifications from structural and functional MRI in patients with depression and healthy controls. TCP predictions were not only robust but also provided a valid measure of confidence for each classification output.

Acknowledgments

The authors would like to acknowledge the support of a Medical Research Council (UK) Discipline Hopping Grant.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2010.05.023.

References

Costafreda, S.G., Chu, C., Ashburner, J., Fu, C.H., 2009a. Prognostic and diagnostic potential of the structural neuroanatomy of depression. PLoS One 4, e6353.

Costafreda, S.G., Khanna, A., Mourao-Miranda, J., Fu, C.H., 2009b. Neural correlates of sad faces predict clinical remission to cognitive behavioural therapy in depression. Neuroreport 20, 637–641.

Davatzikos, C., Shen, D., Gur, R.C., Wu, X., Liu, D., Fan, Y., Hughett, P., Turetsky, B.I., Gur, R. E., 2005. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. Arch. Gen. Psychiatry 62, 1218–1227.

Fu, C.H.Y., Williams, S.C.R., Cleare, A.J., Brammer, M.J., Walsh, N.D., Kim, J., Andrew, C., Pich, E.M., Williams, P.M., Reed, L.J., Mitterschiffthaler, M.T., Suckling, J., Bullmore, E. T., 2004. Antidepressant treatment attenuates the neural response to sad faces in major depression: a prospective, event-related functional MRI study. Arch. Gen. Psychiatry 61, 877–889.

Fu, C.H.Y., Williams, S.C.R., Brammer, M.J., Suckling, J., Kim, J., Cleare, A.J., Walsh, N.D., Mitterschiffthaler, M.T., Andrew, C., Pich, E.M., Bullmore, E.T., 2007. Neural responses to happy facial expressions in major depression following antidepressant treatment. Am. J. Psychiatry 164, 599–607.

Fu, C.H.Y., Mourao-Miranda, J., Costafreda, S.G., Khanna, A., Marquand, A.F., Williams, S. C.R., Brammer, M.J., 2008a. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. Biol. Psychiatry 63, 656–662.

Fu, C.H.Y., Williams, S.C.R., Cleare, A.J., Scott, J., Mitterschiffthaler, M.T., Walsh, N.D., Donaldson, C., Suckling, J., Andrew, C., Steiner, H., Murray, R.M., 2008b. Neural responses to sad facial expressions in major depression following cognitive behavior therapy. Biol. Psychiatry 64, 505–512.

Gammerman, A., Vovk, V., 2007. Hedging predictions in machine learning: the second computer journal lecture. Comput. J. 50, 173–177.

Gammerman, A., Nouretdinov, I., Burford, B., Chervonenkis, A., Vovk, V., Luo, Z., 2008. Clinical mass spectrometry proteomic diagnosis by conformal predictors. Stat. Appl. Genet. Mol. Biol. 7, 13.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131, 681–689.

Marquand, A.F., Mourao-Miranda, J., Brammer, M.J., Cleare, A.J., Fu, C.H.Y., 2008. Neuroanatomy of verbal working memory as a diagnostic biomarker for depression. Neuroreport 19, 1507–1511.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Vovk, V., 2002. On-line confidence machines are well-calibrated. Proceedings of the 43rd Annual Symposium on Foundations of Computer Science: IEEE Comput. , pp. 187–196.

Vovk, V., Gammerman, A., Shafer, G., 2005. Algorithmic learning in a random world. Springer Verlag, New York.

Walsh, N.D., Williams, S.C.R., Brammer, M.J., Bullmore, E.T., Kim, J., Suckling, J., Mitterschiffthaler, M.T., Cleare, A.J., Merlo Pich, E., Mehta, M., Fu, C.H.Y., 2007. A longitudinal fMRI study of verbal working memory in depression following antidepressant therapy. Biol. Psychiatry 62, 1236–1243.