Cortical Surface Thickness as a Classifier: Boosting for Autism Classification*

Vikas Singh[1], Lopamudra Mukherjee[2], and Moo K. Chung[1]

[1] Biostatistics and Medical Informatics, University of Wisconsin-Madison,
vsingh@biostat.wisc.edu, mkchung@wisc.edu
[2] Computer Science & Engineering, State University of New York at Buffalo
lm37@cse.buffalo.edu

Abstract. We study the problem of classifying an autistic group from controls using structural *image data alone*, a task that requires a clinical interview with a psychologist. Because of the highly convoluted brain surface topology, feature extraction poses the first obstacle. A clinically relevant measure called the cortical thickness has shown promise but yields a rather challenging learning problem – where the dimensionality of the distribution is extremely large and the training set is small. By observing that each point on the brain cortical surface may be treated as a "hypothesis", we propose a new algorithm for LPBoosting (with truncated neighborhoods) for this problem. In addition to learning a high quality classifier, our model incorporates topological priors into the classification framework directly – that two neighboring points on the cortical surface (hypothesis pairs) must have similar discriminative qualities. As a result, we obtain not just a label $\{+1, -1\}$ for test items, but also an indication of the "discriminative regions" on the cortical surface. We discuss the formulation and present interesting experimental results.

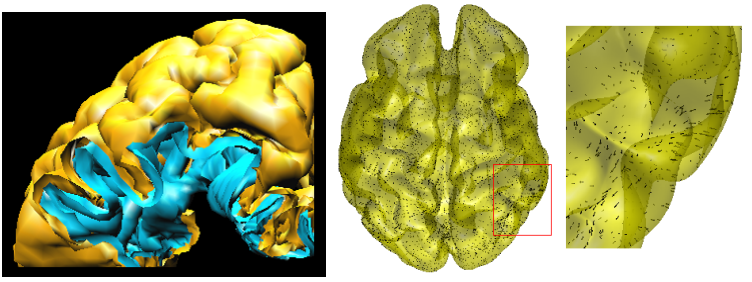
1 Introduction

Learning in biomedical imaging employs training samples provided in the form of image data, with given class labels. We must learn a classifier to assign the correct "label" (positive or negative) to an unseen (test) image. The label may be a pathology (presence or absence of a disease), such as in computer assisted diagnosis. The label may also be a *clinical population group*: for instance, in this paper, our goal is to classify an autistic group from controls – a task that requires an extensive clinical interview with an experienced psychologist [1]. But can we achieve the same objective based on *structural imaging data alone* – efficiently and reliably?

In order to answer the above question, the first difficulty relates to the choice of the features to extract from the images for learning and classification. Feature and shape descriptors (e.g., medial axis, SIFT) that work well across a variety of applications in classical computer vision yield a less than satisfactory performance (in classification) when applied to highly convoluted brain surfaces (their variations are useful in volume

* The first author was supported in part by funds from Dept. of Biostatistics and Medical Informatics, UW-Madison and UW Institute for Clinical and Translational Research (ICTR).

D. Metaxas et al. (Eds.): MICCAI 2008, Part I, LNCS 5241, pp. 999–1007, 2008.



(a) (b)

Fig. 1. (a) Cortical thickness illustration: the outer cortical surface (in yellow) and the inner cortical surface (in blue). The distance between the two surfaces is the cortical thickness. (b) Subsampled surface displacement vector field showing the displacement for one surface (first control subject) to match the other surface (second control subject), as an illustration. The red rectangle region is enlarged to show the displacement vector field (black arrows). Note that the segmentation and cortical thickness calculation were performed in native space.

registrations, however). Among alternatives explored in literature, a promising option is cortical thickness – this measures the distance between the *outer cortical surface* (the interface between gray matter and cerebrospinal fluid), and the *inner cortical surface* (the interface between gray and white matter), see Fig. 1(a). Some neuroanatomical studies [2,3] have reported using this measure for discriminating a clinical population from controls. In cortical thickness based discrimination [3], the image volume is first segmented into tissue types, a mesh representation of the cortical surface (CS) is derived (by triangulation), and the thickness values are calculated at mesh vertex points (in the native space). Then, the standard procedure is to feed such values into a two-sample T statistic at each mesh vertex. However, to account for correlated T statistics at neighboring mesh vertices, we must solve the *multiple comparison problem* [4]: unfortunately, computing the P-value for multiple comparisons is quite challenging. Secondly, such hypothesis driven approaches must satisfy distributional assumptions (e.g., the normality assumption on the cortical thickness values) which may not hold in practice[1], making such approaches error-prone and subsequent quantitative analysis problematic. Since the problem at hand is a classification problem, let us briefly explore the applicability of the powerful support vector machine framework. Given that cortical thickness has a reasonable clinical interpretation, we may consider using it as the measure of choice. Therefore, if a CS (of a single subject) is represented as a mesh with ~ 40000 points, the vectorial representation (say, x) of the cortical thickness lives in $\Re^{40000}$. However, the sizes of datasets in brain imaging literature are typically small (< 100) due to difficulty of recruiting volunteers and cost issues. Therefore, with a finite (and small) training sample $n < 100$, the high-dimensional feature space ($d \gg 100$) where the classifier is calculated is *almost empty* [5]. Hence, as noted in [6], in such cases SVM-based

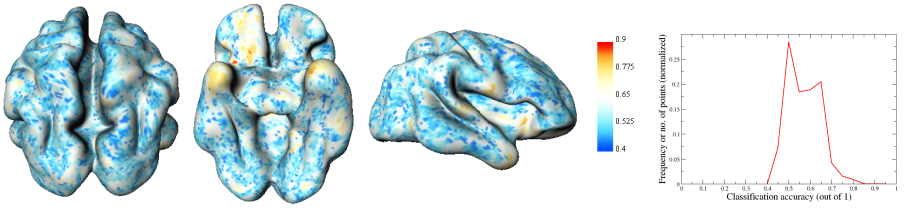
[1] The support of the normal distribution is $-\infty$ and ∞ but cortical thickness values are bounded in $[0\ mm, 6\ mm]$. Also, thickness values are defined on a mesh-vertex and cannot possibly be smooth and differentiable. It may also not be bell-shaped.

classifiers may perform well on training data but will generalize to test data poorly. While some authors have used SVM based methods for classification using brain image data [7], a pre-processing step (typically, a brute force dimensionality reduction using PCA) is used. This seems reasonable for simple shapes (e.g., hippocampus shape data used in [7]) but is too simplistic and immensely lossy for cortical thickness data, and more sophisticated classification tasks.

In this paper, we propose a novel approach for this classification problem. By viewing each set of point-wise correspondences in the training set as a "weak classifier", the training phase seeks to find their best weighted combination, given the correct labels for the training samples. Because the weak classifiers are inherently "spatial", we can exploit this relationship as *priors* within the LPboost framework [8] – which enables us to learn with a small dataset. The paper makes the following contributions: (1) In contrast to the statistical approaches, we do not need to test for the null hypothesis, we also do not compute P-values. Hence, we totally bypass the multiple comparisons problem; (2) In addition to $\sim 90\%$ accuracy, our model yields the goodness of each mesh point as a classifier. This has a physical interpretation – these are the discriminative points between autistic subjects and controls; (3) Because weak classifier pairs are related due to the CS mesh topology, we may derive the discriminative characteristics of "regions" by solving a modified model of LPBoost; (4) The model proposes an algorithm for classifying an autistic group from controls *using image data alone*.

2 Background

First, the raw MRI images must be processed to extract information for use in subsequent steps. We perform an intensity non-uniformity correction before tissue segmentation into three classes: cerebrospinal fluid (CSF), gray matter and white matter (segmentation was performed in native space). We then use a topology preserving deformable surface algorithm [9] to obtain the outer and inner cortical meshes. The details of tissue segmentation and the mesh construction are given in [9]. From the triangular mesh representation of the CS (vertices and triangles) and the spherical



(a) (b)

Fig. 2. (a) Unsupervised clustering on coordinates (correspondence sets) and classification accuracy (see color bar) of individual CS mesh points. (b) A histogram of accuracy. Note that given the clustering for each S_i (correspondence set), we know the color assignment (i.e., accuracy) for the cortical surface point i by comparing with ground-truth.

mapping, we may directly obtain the spherical harmonic representation (SPHARM) [3] of cortical thickness g and cortical surface coordinates $\mathbf{p}$ (specific details can be found in [10,3]). Briefly, SPHARM for surface coordinates is calculated as $\mathbf{p}(\theta,\varphi) = \sum_{l=0}^{k}\sum_{m=-l}^{l}\mathbf{p}_{lm}Y_{lm}(\theta,\varphi)$, where Y_{lm} is the spherical harmonic of degree l and order m, and θ, φ are the Euler angles that parameterize the CS. The Fourier coefficient vectors $\mathbf{p}_{lm}$ are estimated iteratively (low to high degree). The SPHARM representation for cortical thickness is $g(\theta,\varphi) = \sum_{l=0}^{k}\sum_{m=-l}^{l} g_{lm}Y_{lm}(\theta,\varphi)$. Here, we use degree $k = 42$ for the representation. In all, we have 1849 coefficients characterizing the cortical thickness. The surface coordinates $\mathbf{p}$ are represented similarly with 3×1849 coefficients. We may calculate the cortical thickness value at each point on the CS. Nonlinear surface correspondence (registration) may also be established using the spherical harmonic correspondence, see [10]. Fig. 1(b) shows the displacement vector field for warping one subject's CS on to another subject's CS, as an illustration. In §3, we turn our attention back to the classification problem.

3 Main Ideas

A useful observation in approaching our classification problem is the following. Because registration has been performed, we may pick a point i on a CS (and obtain the cortical thickness value), retrieve the correspondences for this point in the other $N-1$ cortical surfaces (with their cortical thickness values), the N thickness values define a *correspondence set*, S_i. We may analyze clusters in this set: by performing a maximum margin clustering (identifying two consecutive points in a sorted set with maximum separation) on this set with two classes. Based on this clustering on S_i, CSs in our dataset belong to one of two classes. A classification at this stage assigns a "$+1$" or "-1" label based on cluster membership of a single point on the CS. Fig. 2(a) shows the classification accuracy of each point i on the average CS. Figure 2(b) shows a histogram of the corresponding values. The classification accuracy is in $[44\%, 92.5\%]$. Not all points on the CS are good class discriminators but a subset (1.8%) of points perform well and have an accuracy of $> 80\%$. The key lies in selecting the subset automatically. Note that we calculate the accuracy by comparing the CS classification to ground-truth data.

3.1 Supervised Classification on the Coordinates

With the large variation in classification accuracy of individual CS points, using individual correspondence set clustering for classification does not seem to be a good idea. However, notice that we may consider each CS point (i.e., correspondence set) to be a "weak classifier" (or hypothesis), then, our goal is to combine a multitude of weak classifiers to obtain a discriminative classifier using training. A powerful machine learning method called "boosting" offers this capability.

Boosting was proposed in [11,12] and has since found applications in many areas including biomedical imaging [13]. A popular boosting algorithm is AdaBoost [14]. Adaboost adds weak classifiers to the ensemble in an iterative fashion, by adjusting the *weight* of the classifier, and the training samples w.r.t. classification accuracy (on the training set). The vector of learnt weights of the unrelated classifiers may then be used

with a new CS to determine class membership ($+1$ or -1). As the reader may have realized, a peculiar characteristic of the problem at hand is that the weak hypotheses are strongly correlated in a four or eight neighborhood sense. For example, two adjacent points on a CS surface (which correspond to two hypothesis) must be expected to have a similar discriminative power. Unfortunately, the solution from AdaBoost may not meet this requirement. Hence, calculated hypothesis weights (via boosting) have little physical interpretation in terms of the cortical surfaces. To address this difficulty, we will look at an alternate model for boosting proposed recently called LPBoost [8]. We will then analyze how it can be modified to include additional meta-information from our problem (e.g., point neighborhoods) in a natural manner.

Classifier Boosting using LPBoost. The method of LPBoost relies on applying the power of linear programming to boosting. Each weak classifier (divides the data set into $+1$ and -1. Rather than adding a new classifier (a combination of given weak classifiers) iteratively at each step, LPboost assumes all weak hypotheses are available. The labels generated by the weak classifiers are considered to be a new feature space, where the goal is to learn a function that minimizes the misclassification error and maximizes the maximum margin of separation. The model is given as [8]:

$$\begin{aligned} \text{(LPBoost)} \min & \sum_{j=1}^{n} a_j + C \sum_{i=1}^{m} \xi_i \\ s.t. & \sum_{j=1}^{n} y_i H_{ij} a_j + \xi_i \geq 1, \quad \forall i \in \{1, \cdots, m\}, \end{aligned} \tag{1}$$

where $a_j \geq 0$. In (1), $\mathbf{y} \in \Re^m$ denotes class-membership for training set items whose entries take values $\{+1, -1\}$. H_{ij} tabulates the response of the j-th hypothesis on the i-th item and $a \in \Re^n_+$ weighs the hypotheses. Like in SVMs, we must allow for a small margin of error (appropriately penalized) to minimize the effects of a few outliers – this gives the so-called "slack" as ξ and C is a regularizer. For better generalization, we penalize the 1-norm of a in the objective which also suppresses redundant features [6].

Boosting with Neighborhoods (truncated smoothness penalty). The previous model of LPBoost does not provide a way to incorporate additional relational information between classifiers. While this may suffice for other applications, in our case, the classifier corresponds to a mesh point and is spatially related to other classifiers. We may model the given cortical surface triangulation as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The points on the CS constitute $\mathcal{V}$ and adjacent points p_i, p_j on the CS mesh are neighbors in $\mathcal{G}$: $v_i, v_j \in \mathcal{V}$ have an edge $e_{ij} \in \mathcal{E}$, $\sim$ denotes neighborhood. Since the cortical thickness (and accuracy of classification) cannot change abruptly across neighboring mesh points, the weights assigned to the classifiers i.e., a entries for neighbors should also be smoothly varying. We propose a model with the following objective function to incorporate such neighborhood information with the same constraints as (1).

$$\text{(LPBoost-N)} \quad \min \quad \sum_{j=1}^{n} a_j + C \sum_{i=1}^{m} \xi_i + \sum_{v_j \sim v_{j'}} \gamma \| a_j - a_{j'} \| \tag{2}$$

where a_j is non-negative. We impose smoothness over the weights assigned to CS points, by penalizing the variation between weights assigned to neighbors v_j and $v_{j'}$, regularized by a parameter, γ. There is no penalty if v_j and $v_{j'}$ take the same weight. To

address the difficulty with the 1-norm in the second term, we can introduce an additional variable, $t_{jj'}$. It can be easily verified that the following is an equivalent model

$$\begin{aligned}
\text{(LPBoost-N}') \min & \sum_{j=1}^{n} a_j + C \sum_{i=1}^{m} \xi_i + \sum_{v_j \sim v_{j'}} \underbrace{\gamma'}_{\rho_{ij}\gamma} t_{jj'} \\
s.t. & \sum_{j=1}^{n} y_i H_{ij} a_j + \xi_i \geq 1 \quad \forall i \in \{1, \cdots, m\}, \\
& a_j - a_{j'} \leq t_{jj'},\ a_{j'} - a_j \leq t_{jj'} \quad \forall e_{jj'} \in \mathcal{E}, \\
& a_j \geq 0 \quad \forall j \in \{1, \cdots, n\}. \qquad (3)
\end{aligned}$$

A note on the third term, $\gamma' t_{jj'}$, in the objective in (3) is in order. Observe that in (2), if γ is constant in $\gamma \| a_j - a_{j'} \|$, we impose a smoothness penalty for all neighbors as a function of $\| a_j - a_{j'} \|$. This encourages the identification of smooth discriminative regions, but the term unnecessarily penalizes (j, j') pairs that lie on either side of the "regions" (analogous to edge pixels on the boundary of foreground/background in image segmentation). Ideally, a_j and $a_{j'}$ should *not* be similar, and the difference should not be penalized. To address this problem, we use a *truncated cost model* [15]: by imposing a smoothness penalty only if (j, j') had similar classification accuracy *on the training set*. This can be modeled using ρ_{ij} which is 1 if (j, j') had similar accuracy (within a threshold, t) and $\rho_{ij} = 0$ otherwise. Therefore, in the third term in (3), $\gamma' = \rho_{ij}\gamma$. To wrap up, the model in (3) is linear, and the requirement on a is only of non-negativity. So, (3) can be solved optimally in polytime.

4 Experimental Results

The experimental evaluation of the algorithm was designed to investigate the suitability of the framework in context of the following issues from §1: (1) Can we learn a classifier from training image data to reliably classify autism group and controls? If yes, what kind of accuracy can we hope for? (2) In addition to a binary class assignment, can we determine the *discriminative regions* that help us classify? This would be very useful information – the existence of such areas convey that the structural basis of autism is localized in brain regions, we may be able to better understand the structural connection to the functional deficit in autistic subjects. So, instead of trying to investigate every part of the brain, we may limit our investigation to these discriminative areas, possibly using more traditional hypothesis driven statistical inference.

Acquisition and Processing. We acquired three Tesla T_1-weighted MR brain image scans for 11 controls and 16 high functioning autistic subjects (27 subjects in all), see [1] for details. The autistic subjects were diagnosed by a certified psychologist, this was used as the "truth" classification, $\mathbf{y}$. The standard image processing steps from §2 were performed, and the weak hypotheses were generated for boosting. Boosting using our model in (3) was performed on the cortical thickness vectors using k-fold cross validation procedure ($k = 9$), and the mean of the results analyzed. Since cross validation experiments can be repeated $n!$ times (for n items), which grows rapidly, we

randomly permuted the data set and repeated the experiments 10 times and report on the mean for each case.

A large number of CS points have poor discriminative characteristics. Ideally, the boosting algorithm should ignore all such points, the weighted combination should include only the discriminative weak classifiers. However, when the number of hypotheses is large (with a relatively smaller training set), occasionally a few not-so-good hypotheses are assigned non-zero weights. Inclusion of such hypotheses reduces the generalization behavior of the boosted classifier. This can be partly mitigated in practice by moderately pruning the set of hypotheses and boosting only the better classifiers, i.e., classifiers that have lower training error for a particular training set. We performed this pruning step for classifiers by only including hypothesis with error (on training data) below a user specified threshold ($15\% - 45\%$).

The model in (3) has at most $27 + H + N$ variables and $27 + 2N$ constraints, where H is the maximum number of hypotheses (40392) and N is the number of neighbors ($< 6H$). Solving the model to obtain a solution using CPLEX took $\sim 20s$. To prune the set of hypothesis, we repeated each of the above experiments for hypotheses with $15\%-45\%$ training error. Each experiment was repeated for LPBoost without neighborhoods (1) and LPBoost-N$'$ with truncated neighborhoods (3). The regularizer, C, was set to 100. In all cases, we report on misclassification errors on the test data sets.

Fig. 3 compares the misclassification errors for LPBoost and LPBoost-N$'$ (truncated neighborhoods). We see that LPBoost-N$'$ outperforms LPBoost in all cases, with a mean error of $\sim$ 10% in cases where the maximum error in the included hypotheses is 25%. Also, analyzing misclassification error as a function of increasing the number of hypotheses considered (increasing training error) shows that the errors are relatively high when the size of the hypothesis set is very small. It improves as the size increases and plateaus between $25\% - 40\%$, with a slight deterioration as the set includes more hypotheses with very high error. In summary, by combining the discriminative power of cortical thickness at many CS points, we can classify the autism group from controls with $\sim 90\%$ accuracy.

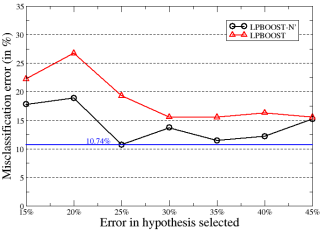


Fig. 3. Misclassification error for boosting on coordinates for LPBoost and LPBoost-N$'$

The entries of the weight vector, a, returned by our model are non-zero for a small subset of hypotheses, this gives a way of determining discriminative weak classifiers from a large set. Because of the spatial contiguity requirement in (3), we get contiguous discriminative regions on the cortical surface (see Fig. 4, regions in red). By incrementally including more hypotheses to boost (15% to 35%), most regions identified for fewer hypotheses exhibit an expansion, see Fig. 4 (left to right). It is also very encouraging to see that if we compare Fig. 4 with Fig. 2(a), the regions in Fig. 4 are a subset of the "good" regions in Fig. 2(a) (notice that since Fig. 2(a) corresponds to all 27 CS, it may serve as ground truth). It is expected that all high accuracy CS points in Fig. 2(a) are not selected by boosting because the model selects *only* a minimal subset of hypotheses needed to yield an accurate classifier. If desired, we may determine *all* discriminative regions by

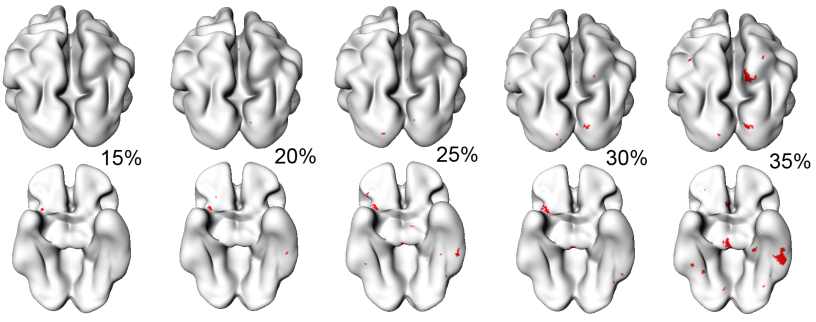


Fig. 4. The discriminating regions selected by the model (non-zero weights in a in (3)) in red corresponding to increasing set of hypotheses selected

repeatedly running the algorithm on a reduced set of hypotheses (by removing the regions already selected), we omit these results due to limited space.

5 Conclusions

We present a LPboosting based algorithm for classifying autistic subjects from controls based on cortical thickness. The model incorporates spatial priors – as a result we obtain discriminating regions on the cortical surface in addition to high classification accuracy ($\sim 90\%$) on test items. As future work, it will be interesting to see if improvements are possible by incorporating additional clinically relevant features (apart from cortical thickness) in the model. Given that in brain imaging, we often encounter datasets in high dimensions but with few data items, classifiers that learn from a wide spectrum of information may generalize better, and are desirable for robust classification systems.

References

1. Dalton, K.M., Nacewicz, B.M., Johnstone, T., Schaefer, H.S., Gernsbacher, M.A., Goldsmith, H.H., Alexander, A.L., Davidson, R.J.: Gaze fixation and the neural circuitry of face processing in autism. Nature Neuroscience 8, 519–526 (2005)
2. Kabani, N., Le Goualher, G., MacDonald, D., Evans, A.C.: Measurement of cortical thickness using an automated 3-d algorithm: a validation study. NeuroImage 13(2), 375–380 (2001)
3. Chung, M.K., Dalton, K.M., Shen, L., Evans, A.C., Davidson, R.J.: Weighted fourier representation and its application to quantifying the amount of gray matter. IEEE Trans. Med. Imaging 26, 566–581 (2007)
4. Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C.: A unified statistical approach for determining significant signals in images of cerebral activation. Human Brain Mapping 4, 58–73 (1996)
5. Hertz, J., Krogh, A., Palmer, R.G.: Introduction to the Theory of Neural Computation. Addison-Wesley, Reading (1991)
6. Bradley, P.S., Mangasarian, O.L.: Feature selection via concave minimization and support vector machines. In: International Conf. on Machine Learning, pp. 82–90 (1998)

7. Shen, L., Ford, J., Makedon, F., Saykin, A.: Surface-based approach for classification of 3d neuroanatomical structures. Intelligent Data Anal. 8, 519–542 (2004)
8. Demiriz, A., Bennett, K.P., Shawe-Taylor, J.: Linear programming boosting via column generation. Machine Learning 46(1-3), 225–254 (2002)
9. MacDonald, J.D., Kabani, N., Avis, D., Evans, A.C.: Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. NeuroImage 12, 340–356 (2000)
10. Chung, M.K., Hartley, R., Dalton, K.M., Davidson, R.J.: Encoding cortical surface by spherical harmonics. In: Statistica Sinica, Special Issue on Statistical Challenges and Advances in Brain Science (in press, 2008)
11. Schapire, R.E.: The strength of weak learnability. Machine Learning 5, 197–227 (1990)
12. Freund, Y.: Boosting a weak learning algorithm by majority. In: Proc. of Computational Learning Theory (1990)
13. Tu, Z., Zheng, S., Yuille, A.L., Reiss, A.L., Dutton, R.A., Lee, A.D., Galaburda, A.M., Dinov, I., Thompson, P., Toga, A.W.: Automated extraction of the cortical sulci based on a supervised learning approach. IEEE Trans. Med. Imaging 26(4), 541–552 (2007)
14. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Conf. on Computational Learning Theory, pp. 23–37 (1995)
15. Gupta, A., Tardos, E.: Constant factor approximation algorithms for a class of classification problems. In: ACM Symposium on Theory of Computing, pp. 652–658 (2000)