Using Multiparametric Data with Missing Features for Learning Patterns of Pathology

Madhura Ingahalikar¹, William A. Parker¹, Luke Bloy¹,
Timothy P.L. Roberts², and Ragini Verma¹

¹ Section of Biomedical Image Analysis, University of Pennsylvania,
Philadelphia, PA, USA

{Madhura.Ingalhalikar,Ragini.Verma}@uphs.upenn.edu

² Lurie Family Foundations MEG Imaging Center, Department of Radiology,
Children's Hospital of Philadelphia, Philadelphia, PA, USA*

Abstract. The paper presents a method for learning multimodal classifiers from datasets in which not all subjects have data from all modalities. Usually, subjects with a severe form of pathology are the ones failing to satisfactorily complete the study, especially when it consists of multiple imaging modalities. A classifier capable of handling subjects with unequal numbers of modalities prevents discarding any subjects, as is traditionally done, thereby broadening the scope of the classifier to more severe pathology. It also allows design of the classifier to include as much of the available information as possible and facilitates testing of subjects with missing modalities over the constructed classifier. The presented method employs an ensemble based approach where several subsets of complete data are formed and trained using individual classifiers. The output from these classifiers is fused using a weighted aggregation step giving an optimal probabilistic score for each subject. The method is applied to a spatio-temporal dataset for autism spectrum disorders (ASD)(96 patients with ASD and 42 typically developing controls) that consists of functional features from magnetoencephalography (MEG) and structural connectivity features from diffusion tensor imaging (DTI). A clear distinction between ASD and controls is obtained with an average 5-fold accuracy of 83.3% and testing accuracy of 88.4%. The fusion classifier performance is superior to the classification achieved using single modalities as well as multimodal classifier using only complete data (78.3%). The presented multimodal classifier framework is applicable to all modality combinations.

1 Introduction

Pattern classification techniques are generating increasing interest in the neuroimaging community as they are powerful in learning the patterns of pathology

* The authors would like to acknowledge support from the NIH grants: MH092862, MH079938 and DC008871.

from a population, assign a probabilistic score to each subject which characterizes pathology on an individual basis and aid in assessing treatment in conjunction with other clinical scores [1,2]. The earlier single modality [1,2] studies have given way to multimodality classifiers that can potentially aid in exploring additional dimensions of pathology patterns and provide a rich multiparametric signature or profile with increased diagnostic accuracy [3,4]. However, none of these studies account for a challenging problem plaguing clinical studies, that data from some modalities could be missing, as the subjects do not complete the entire study due to enhanced pathological severity or scanner issues and noise that force the studies to partially discard the data. Removing subjects with incomplete datasets from the study (as is the approach adopted by the traditional multimodal classification studies) reduces the already small sample size and diminishes the information content in the dataset, thus making the classifier decision unreliable since it does not account for the pathology patterns from the subjects who were unable to complete the clinical study due to their more severe form of pathology. Further, it limits the dimensionality of the multimodal approach since the probability of a subject being excluded increases with the number of modalities attempted.

In statistical theory, missing value problems are addressed using various strategies based on the patterns in missing data [5]. For randomly missing data, imputation techniques that substitute or fill in the missing items are commonly used [5,6]. Imputation methods can include substitutions like mean or median of the feature or multiple substitutions which replace each missing value with a set of plausible ones, reflecting the underlying uncertainty in the data [6]. Thus, the multiple imputation technique is considered as one of the effective methods to handle partial data. Other well-established strategies to deal with missing data involve model based procedures like expectation maximization (EM) which potentially recover unknown values from similar samples. Finally, simple decision tree classifiers have also been utilized as they avoid the missing data completely [6].

However, most of the above methods perform substitution in some way or the other, which usually interpolates and may create spurious data and thus cannot be completely trusted when the percentage of missing data is high ($\approx 30\%$ and above). Moreover, if the missing data is associated with an extreme in the pathologic condition, which is widely true, interpolation would become extrapolation making the data highly unreliable. Finally, these methods just attempt to fill in and thus do not directly take the classification problem into consideration [7]. Recent machine learning literature has shown ensemble classifiers to be an effective way to accommodate sparse data by using weak classifiers and then boosting the performance by combining the output [7], [8].

In this paper we present an ensemble based classification framework that has the potential to handle spatio-temporal multimodal data with a high percentage of missing values in a petite sample size. The method learns patterns of pathology on different subsets created from the original data and aggregates the output from all the classifiers using a weighting strategy, giving an optimal probabilistic score for each subject. The method considers subjects with complete or incomplete data

in training the classifier as well as in testing, without filling in the missing values. We apply this method on a population with Autism Spectrum Disorder (ASD) and typically developing controls (TDC), where the pathology can be investigated by creating spatio-temporal classifiers that utilize the spatial features from diffusion tensor imaging (DTI) of white matter and the temporal features computed from magnetoencephalography (MEG) data increasing the classification accuracy over the single modality classifiers.

2 Methods

We use ensemble classifiers, i.e. a pool of classifiers, each trained using a subset of the original dataset [7]. The output from all the classifiers is fused together to boost the overall performance. This fusion is a weighted aggregation based on the classification accuracy of each classifier as well as the similarity between the features used in the training and the testing set. We successfully demonstrate the applicability of this method on a spatio-temporal dataset derived from MEG features and DTI features.

2.1 Classification of a Dataset with Missing Values

Consider a dataset with n subjects $(x_1, x_2, \dots, x_n)$ and m features. Each subset (S) is defined by a collection of subjects ($< n$) that have complete data for a specific set of features $s (s \leq m)$. Subsets are formed such that the total number of subsets t encompass all the subjects and the features from the original dataset. The decision of how many subsets to create depends upon what features are missing in the training and testing samples. In the case where all the features have some part missing, $2^m - 1$ subsets can be created.

A classifier model (e.g. LDA, SVM etc) is trained over each subset resulting in t outputs. Thus, for a subject x_i , the classification output can be formulated as $(O(x_i, S_1), \dots, O(x_i, S_t))$ resulting from t subsets. At the next stage, all the outputs from individual classifiers are combined to boost the overall performance. While fusing the outputs it is important to note that some subsets could be more valuable than others, depending on the subject under testing. Therefore, in the aggregation stage the final output for a subject is given by the weighted combination of the subset outputs [7]. For example, for a subject x_{test} under testing the final output is given by 1.

$$O_{agg}(x_{test}) = \frac{1}{\sum_i \frac{1}{\varphi_{i, x_{test}}}} \sum_{i=1}^t \frac{1}{\varphi_{i, x_{test}}} O(x_{test}, S_i) \quad (1)$$

In 1, $\varphi_{i, x_{test}}$ is the expected error of the classifier (given by equation 2) for subset S_i which depends on the general accuracy of the classifier for S_i as well as the number of features used in the classifier.

$$\varphi_{i, x_{test}} = \frac{1}{\eta_{x_{test}}} \sum_{x \in training} d(x, x_{test}) \varphi_{i, x} \quad (2)$$

where

$$d(x, x_{test}) = \sqrt{\|x - x_{test}\|^2 f(s_x^T K s_{x_{test}})} \quad (3)$$

$$\eta_{x_{test}} = \sum_{x \in training} d(x, x_{test}) \quad (4)$$

$$\varphi_{i,x} = (O(x, S_i) - Y(x))^2 f(s^T K s_x) \quad (5)$$

The training accuracy of each classifier is accounted via $\varphi_{i,x}$ in equation 2. $\varphi_{i,x}$ is the expected error for a subject under training given by equation 5, where $Y(x)$ is the known label and $O(x, S_i)$ is the output of subset S_i for a training subject x . The $d(x, x_{test})$ term in equation 2 takes into account the distance between two samples. The similarity between the feature space of the subset and the subject is given by function f in equations 3 and 5 where K is a diagonal matrix that weighs the features based on their information content [7]. Here we define f as $f(v) = 1/v$, a non-increasing function, accounting for the similarity term as well as the feature ranking.

2.2 Classification Using MEG+DTI Features

Feature Extraction. In this study, we consider 3 categories of features. These features are associated primarily with language impairment in ASD, but the framework is applicable to any set of MEG-DTI features. The features used in this study are: (i) the latency of auditory evoked neuromagnetic field 100ms component called M100 [9], (ii) the latency of magnetic mismatch field (MMF), which is a response component reflecting detection of ‘change’ in the auditory stream [10], and (iii) Fractional anisotropy (FA) and mean diffusivity (MD) measures from 37 ROI’s created by a normalized cuts clustering method [11] in WM areas of brain associated with language (figure 1), providing 74 values.

Feature Ranking in DTI. In the subsets that contain DTI as a feature, we perform feature ranking on the 74 DTI attributes. This step provides us with attributes that most contribute to the patient-control classification and aids in minimizing the classification error. To find a compact discriminatory subset of features, we choose a ranking and selection method known as the signal-to-noise (s2n) ratio coefficient filter [12]. For the j^{th} DTI attribute vector $d_{ti,j}$ and class labels Y , the signal to noise ratio is given by equation 6. In this equation, $\mu(y+)$ and $\mu(y-)$ are the mean values while $\sigma(y+)$ and $\sigma(y-)$ are the variances for class $y+$ and $y-$ respectively. Based on these s2n coefficients, the features are ranked and a subset of the top ranked features is implemented in the classification.

$$s2n(d_{ti,j}, Y) = \frac{\mu(y+) - \mu(y-)}{\sigma(y+) + \sigma(y-)} \quad (6)$$

Training, Cross Validation and Testing. We have created a generalized framework; therefore, any kind of classifier (e.g. SVM, QDA etc.) would work in

this ensemble setup. We implement a simple linear discriminant analysis (LDA) classifier on each of our subsets as the aim here is to demonstrate that incorporating missing data from our multimodal features aids in classification.

Since our missing data is spread over all three features, we use total of 7 subset combinations (M100, MMF, DTI, M100+DTI, MMF+DTI, MMF+M100 and MMF+M100+DTI). If a subset contains DTI as a feature, then s2n ranking is performed only on the DTI features. The number of top ranked DTI features to be retained in the classification process is based on the minimum cross validation error computed from the error plot that is constructed by using different number of features in the cross validation [12]. The ranking matrix K defined in section 2.1 is set to identity, implying that M100, MMF and DTI provide equal information.

We compute a probabilistic abnormality score for each subject using a 5-fold cross validation on the training data which is permuted 100 times for generalization of the folds. Finally, the trained classifier framework is applied to test data with missing values.

3 Results

Dataset and Preprocessing. Our dataset consisted of 138 subjects (42 TD and 96 ASD), out of which 55 subjects had complete data while others had some feature missing (60.1% subjects with partial data). 30% of subjects were missing MMF, 15.7% were missing M100 and 38% were missing DTI. We randomly picked 112 subjects (51 complete and 61 subjects with partial data, making 54.4% missing data) for training and the other 26 (4 complete and 22 partial data) as test data.

The MEG recordings were performed using a CTF 275-channel biomagnetometer with the following protocol: (i) binaural auditory presentation of brief sinusoidal tone stimuli at 45dB SL. M100 latency was determined from the source modeled peak of the stimulus-locked average of 100 trials of each token (ii) binaural auditory presentation of interleaved standard and deviant tone and vowel tokens (/a/, /u/). Mismatch field (MMF) latency was determined from the subtraction of superior temporal gyrus (STG) source-modeled responses for each token as deviant vs. standard. The DTI data were acquired on Siemens 3T

Fig. 1. Figure displays the ROI's in which mean FA and MD were used in the classifier. These ROI's were computed using normalized cuts algorithm in the areas associated with language.

Fig. 2. Bar chart comparing the accuracies of 4 different classifiers. The performance of the fusion method becomes superior with more missing data. Multimodal classifiers perform better than single modality classifiers in all the cases. The x-axis displays the percentage of missing data starting with 51 complete subjects and then randomly adding 15 subjects with missing data in each case until all the subjects with missing data were added (total 112 subjects) in the last case.

VerioTM scanner using the Stejskal Tanner diffusion weighted imaging sequence (2mm isotropic resolution) with $b=1000 \text{ mm/s}^2$ and 30 gradient directions.

The mean scalar features (FA and MD) from DTI were computed in each of the 37 ROIs (figure 1). For this initially, language related WM ROI's (superior temporal white matter (STWM), inferior and superior longitudinal fasciculi (SLF, ILF), inferior fronto-occipital fasciculus (IFOF)) were derived from a standard atlas called EVE [13]. The size of these ROI's is large enough to smooth out the effect of mean FA and MD. Therefore, we implemented a normalized cuts algorithm [11] that was based on a variance threshold computed over the DTI images to divide these ROI's into smaller regions with homogeneous WM.

Training, Cross Validation and Testing. Using 5-fold cross validation, we computed the classifier score for each subject in training. For our dataset, the top 15% DTI features were employed based on their optimal performance in the cross-validation of DTI classifier. Figure 2 shows a bar chart of training accuracies for classifiers with only DTI, only MEG (MMF and M100), DTI and MEG with deletion of incomplete data and DTI and MEG with the fusion approach. When only the subjects with complete data ($N=51$) were considered, the classifier performed with only 78.3% accuracy, while when using the whole sample ($N=112$), including those with missing data, the accuracy increased to 83.3%.

Out of the 7 subset classifiers, other than the classifier with all 3 features, the individual MMF classifier (77.7%) and the MMF+DTI classifier (79.4%) added to the overall discriminative power. Individual DTI classifier performed with an accuracy of $\approx 75\%$, higher than the individual M100 classifier ($\approx 70\%$), suggesting that combining modalities performed better (note that the number of subjects in

Fig. 3. (a) ROC curve for classifier with complete data (51 subjects) and ROC for classifier with 54.4% missing data (112 subjects). (b) ROI's that we frequently selected by the DTI feature selection technique in the 5-fold classification.

each classifier were different). The ensemble framework gains from the diversity of the subset classifiers, boosting the overall performance.

Figure 3(a) displays the receiver operating characteristic (ROC) curves for the cases with complete data (51 subjects) and missing data based (112 subjects) on the 5-fold validation. The area under curve for the first case was 0.73 that increased to 0.82 in the second case. Finally, we classified the 26 subjects reserved as testing samples, on the fusion classifier trained on the 112 subjects. The testing accuracy was 88.5% where 100% a (4/4) on subjects with complete data and 86.4% (19/22) subjects with partial/missing data were classified correctly.

All the brain regions employed to extract DTI features are specific to language impairment in ASD. The feature selection ranks these regions in each of the cross-validation loop suggesting that the selected regions have more discriminative power. The most frequently selected features (in the 5-fold validation) via s_{2n} ranking scheme are shown in figure 3(b). These mainly include right SLF and right and left STWM. Other regions like left SLF were also involved in the 5-fold classification, but were not selected very frequently.

4 Conclusion

We have presented a classification technique that can build classifiers on multi-modal data with missing modalities/features. We applied it to a problem involving spatio-temporal (MEG and DTI) data with features that were associated with language abnormalities in ASD. When multiple modalities are utilized within a classification framework, the aggregate output for each subject can potentially define a more comprehensive quantification of pathology than when used individually. Such an approach necessitates a means for handling subjects with incomplete data, such as presented here. Our ensemble approach demonstrated superior performance when compared with those utilizing the smaller complete samples, suggesting that utilizing the subjects with incomplete data was advantageous to

correctly learning the patterns of difference. The internal DTI feature ranking pointed out the regions that were responsible for language impairment in ASD while the subset classifiers in the ensemble provided insight into the relative contributions of DTI and MEG to classification.

In large studies that involve multimodality imaging data together with psychological scores, genomic data etc., a high percentage of missing data is expected. Our generalized framework can be readily applied to such problems and will have a high impact.

References

1. Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C.: Compare: classification of morphological patterns using adaptive regional elements. *IEEE Trans. Med. Imaging* 26(1), 93–105 (2007)
2. Ingallhalikar, M., Parker, D., Bloy, L., Roberts, T.P.L., Verma, R.: Diffusion based abnormality markers of pathology: toward learned diagnostic prediction of asd. *Neuroimage* 57(3), 918–927 (2011)
3. Batmanghelich, N., Dong, A., Taskar, B., Davatzikos, C.: Regularized Tensor Factorization for Multi-Modality Medical Image Classification. In: Fichtinger, G., Martel, A., Peters, T. (eds.) *MICCAI 2011, Part III*. LNCS, vol. 6893, pp. 17–24. Springer, Heidelberg (2011)
4. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Initiative, A.D.N.: Multi-modal classification of alzheimer's disease and mild cognitive impairment. *Neuroimage* 55(3), 856–867 (2011)
5. Little, R., Rubin, D.: *Statistical Analysis with Missing Data*. John Wiley (2002)
6. Garcia-Laencina, P.J., Sancho-Gomez, J.L., Figueiras-Vidal, A.R.: Pattern classification with missing data: a review. *Neural Comput. and Application* 19(2), 263–282 (2009)
7. Ghannad-Rezaie, M., Soltanian-Zadeh, H., Ying, H., Dong, M.: Selection-fusion approach for classification of datasets with missing values. *Pattern Recognit.* 43(6), 2340–2350 (2010)
8. Wang, C., Liao, X., Carin, L.: Classification of incomplete data using dirichlet process priors. *Journal of Machine Learning Research* 11, 3269–3311 (2010)
9. Cardy, J.E.O., Flagg, E.J., Roberts, W., Brian, J., Roberts, T.P.L.: Magnetoencephalography identifies rapid temporal processing deficit in autism and language impairment. *Neuroreport* 16(4), 329–332 (2005)
10. Roberts, T.P.L., et al.: Meg detection of delayed auditory evoked responses in autism spectrum disorders: towards an imaging biomarker for autism. *Autism Res.* 3(1), 8–18 (2010)
11. Bloy, L., Ingallhalikar, M., Eavani, H., Schultz, R.T., Roberts, T.P.L., Verma, R.: White matter atlas generation using hardi based automated parcellation. *Neuroimage* (August 2011)
12. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182 (2003)
13. Oishi, K., et al.: Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and alzheimer's disease participants. *Neuroimage* 46(2), 486–499 (2009)