Diagnosis of Fragile X and Turner's Syndrome

Kevin Miller (kjmiller@stanford.edu)

Dilli Raj Paudel (drpaudel@stanford.edu)

CS 229 Machine Learning Final Project

1. Introduction

Fragile X and Turner's syndrome are genetic disorders that affect a person's development and cause changes in brain structure. Though both of them affect the X chromosomes, Turner's syndrome is common only among girls. We are using data from the Center for Integrated Brain Sciences Research (CIBSR) covering brain images, behavioral scores and demographics. The brain imaging data have been preprocessed using FreeSurfer, a software toolkit that uses MRI data to reproduce the brain's cortical surface. Machine Learning techniques have been applied to raw data with varying degrees of success, but this is the first time they are being used on FreeSurfer data for these diseases.

2. Problem Statement

A typical question is what brain structure changes characterize the different populations. The problem is complicated by the fact that brain structure size depends on head size, and also normally changes during childhood and adolescence. The brain region results were computed with FreeSurfer, and the datasets include many more individual brain regions than have yet been studied.

Among the datasets are 60 subjects scanned at two time points, thus longitudinal growth changes can be measured in these subjects. Not much is known yet about how longitudinal changes in brain structure may correlate with changes in behavioral measurements.

In this project we tried to answer the following questions:

1. Can we tell a Turner's subject from control and a Fragile X patient from control?
2. Can we separate Fragile X and Control subjects (girls only) from one another and from controls?
3. Can we classify a particular subject as having Fragile X, Turner's or other Developmental Delay or a healthy control?

3. General Approach

We used SVM based algorithms, primarily those found in the liblinear and LIBSVM library. The algorithms included L1-regularized L2-loss SVM, L2-regularized L2-loss SVM, L2-regularized L1-loss SVM and L2-regularized Logistic Regression and Gaussian Kernel SVM. For multivariate classification, we used the multiclass option in liblinear and the Gaussian-kernel multi-class classification algorithm from LIBSVM.

4. Turner's Syndrome Classification

The Turner's dataset contained 57 data points after cleaning the missing data. Each data point had 425 features. We ran rankfeatures using KL Divergence, TSanalysis (written by Paul Mazaika from CIBSR) and L1 regularized SVM to come up with 205 most relevant features. We used these features to run a backward search on the dataset to get the best set of features on the basis of minimum generalization error. Similarly, we ran a forward search on the dataset using the complete set of (425) features. To compute the generalization error, we ran 3-fold cross-validation a thousand times in order to deal with noise.

Generalization Error for Turner's Syndrome Data Using Different Algorithms

Algorithm	Backward	Backward (No IQ)	Forward	Forward (No IQ)
Gaussian	6.5479%	10.703%	8.3001%	17.837%
L2RL2L	3.3166%	2.0772%	5.2325%	17.21%
L2RL1L	4.4965%	4.6552%	7.3015%	13.524%
L1RL2L	4.3992%	7.9177%	5.9533%	12.42%
L2RLR	***1.8875%***	4.6085%	8.3077%	16.903%

Best Result Using L2 regularized Logistic Regression and Backward Search to get the best set of features:

RH Parahippocampal Area
RH Insula Area
RH BA6 Volume
PRIorPIQ
RH Parsopercularis Area
LH V1 Volume
RH BA1 Mean Curvature
RH Superior Parietal Area
LH BA4p Area
LH BA44 Mean Curvature

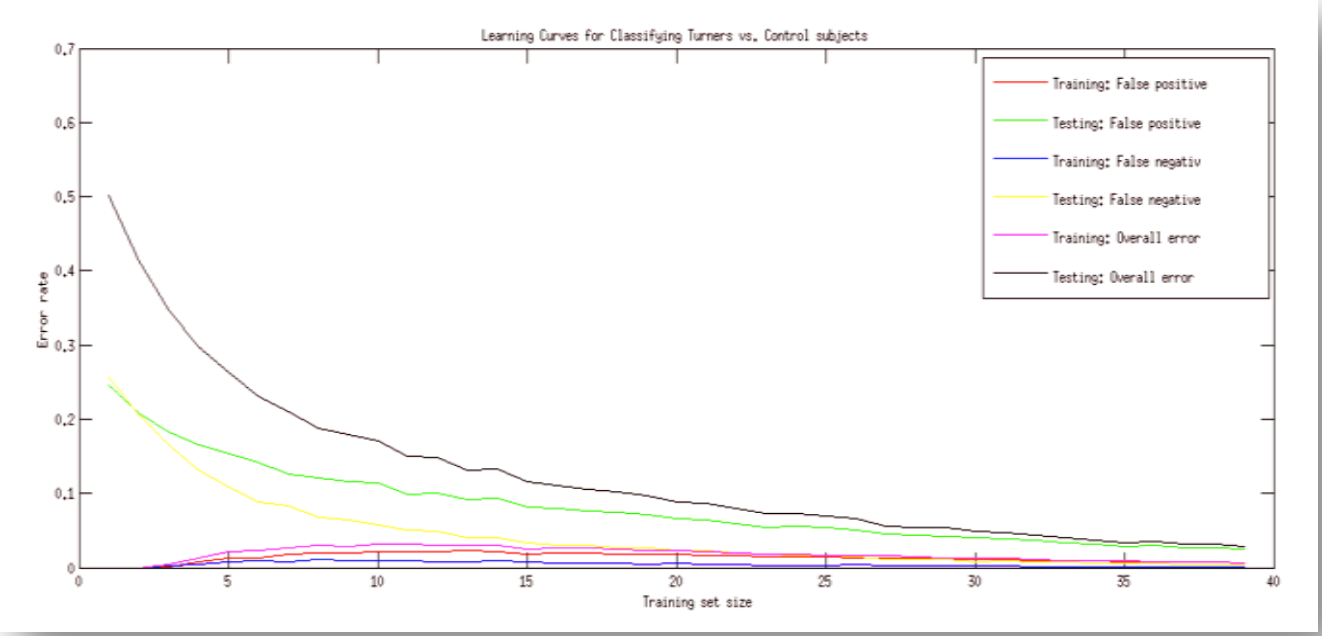

Figure 1: Learning Curves for Turner's syndrome Data Classification (Turner's vs. Controls)

5. Fragile X Classification

The Fragile X dataset contained 243 data points 183 from time point 1 and 60 from time point 2. We only ran our algorithms on time point 1 data. Each data point had 318 features, 276 of which were common with Turner's syndrome. We used these 276 features to run backward and forward searches. We came up with the following best set of features on the basis of the generalization error. Here also, we ran 3-fold cross validation 1000 times using the best set of features and plotted the learning curves as shown in Figure 2.

Best Results Using L2 regularized Logistic Regression and Backward Search to get the best set of features:

Gender
RH Lateral Occipital Volume
LH Rostral Anterior Cingulate Thickness
LH Posterior Cingulate Thickness
RH Transverse Temporal Thickness
RH Transverse Temporal Area
RH Precentral Area
Left Ventral DC
RH Parsopercularis Volume
LH Transverse Temporal Thickness
LH Lateral Occipital Thickness
RH Parahippocampal Thickness
RH Parstriangularis Area
LH Fusiform Meancurv

Generalization Error for Fragile X Syndrome Data Using Different Algorithms

Algorithm	Backward	Forward
Gaussian	41.176%	6.4595%
L2RL2L	2.6705%	4.2376%
L2RL1L	2.7814%	3.6805%
L1RL2L	2.511%	4.0592%
L2RLR	***2.48%***	4.454%

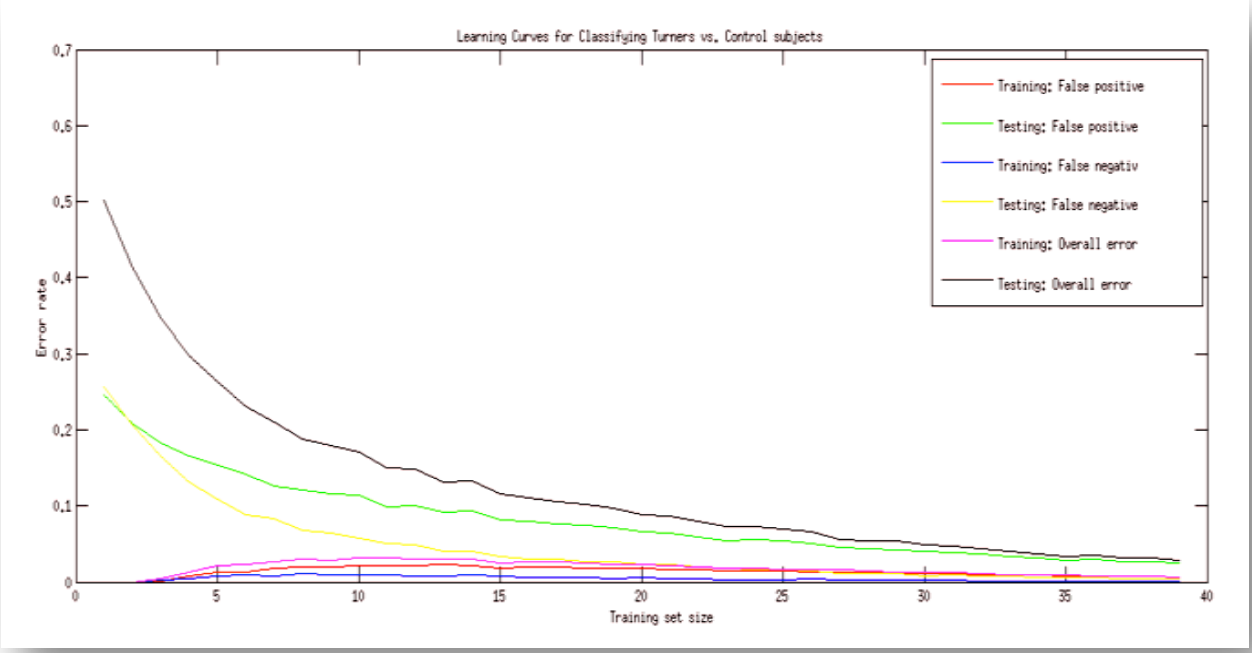

Figure 2: Learning Curves for Fragile X Syndrome Data (Fragile X vs. Control)

6. Fragile X vs Turner's

We used the data points involving girls only to see if we could do multiclass classification on Fragile X, Turner's and controls. The generalization error rates we got were 12.793% forward search and 15.103% backward search when using 3-fold cross validation with 1000 repetitions. When using the LIBSVM Gaussian kernel for multiclass we managed get generalization errors of 11.702% for backward search and 13.759% for forward search using 3-fold cross-validation with 10 repetitions.

7. Further Work

We could try to regress out the Intra-Cranial Volume (ICV) and/or the age to cut down variance. Analyzing the residuals of the features between the two time points would probably give us some insight into changes in behavior. We could also try using some unsupervised learning algorithms to check what structure the data has.

8. Acknowledgements

We would like to thank Paul Mazaika from CIBSR for providing us the data and the TSAnalysis program and Catie Chang, one of our TA's for referring us to Paul. We would also like to thank Professor Alan Reiss and Paul (again) from CIBSR for providing up feedback on our initial results.

9. References

Bray, Signe, Catie Chang and Fumiko Hoeft.

Applications of multivariate pattern classification analyses in developmental neuroimaging of healthy and clinical populations. Frontiers in Human Neuroscience (2009): 1-12.

Walter, E., P. K. Mazaika and A. L. Reiss.

Insights into Brain Development from Neurogenetic Syndromes: Evidence from Fragile X Syndrome, Williams Syndrome, Turner Syndrome and Velocardiofacial Syndrome. Neuroscience (2009): 257-271.

Huang, Tzu-Kuo

Multi-class classification (and probability output) via error-correcting codes LIBSVM Tools Page

http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#multi_class_classification_and_probability_output_via_error_correcting_codes