Analysis of MEG BioMarkers To Determine Parkinson's And Related Diseases

2nd Draft December 15, 2016

Summary

The EEG data provided by UCH at Denver Neuroscience Department has been analyzed using the data science techniques most likely to achieve results. The small sample of 54 measurements was diagnosed for Parkinson's, Essential Tremor and Parkinson's Plus. The sample size is small but significant results are achieved using random forest classifiers and either raw or bootstrapped data is found[i]. The process is to classify classical Parkinson's versus Other Parkinson's and then classifying Other Parkinson's between Essential Tremor and Parkinson's Plus.[1] The results were found to be accurate within a 95% confidence interval. Within a 95% confidence interval the EEG data can be relied upon to predict the three classes of Parkinson's Disease('PD'). [2]

The bootstrapping method shows that a sample between 250 and 500 would yield predictions within a 99% confidence interval for both the precision and specificity.[3] This assumes that the classes would remain at around half of the data samples each. Imbalanced classes will need to be adjusted by bootstrapping.[4] Confirmation by a larger sample between 250 and 500 is recommended.

Samples

The sample size is 44 patients previously diagnosed for Parkinson's, Parkinson's Plus or Essential Tremor. There are two sets of measurements available for each sample: (1) the samples compared on two consolidated measurements with such consolidation performed by traditional techniques and (2) the samples compared on the entire set of measurements available from the technique ('all phenotypes') of over 500 measurements per sample. The analysis demonstrates that the two measurements sample set is an effective predictor while the multiple phenotypes sample shows promise in identifying relationships between the measurements that may lead to further areas of study.

Ensemble Procedure

The 'two measurements' data is scaled and passed directly in the classifiers. The 'all phenotypes' data is reduced through principal components analysis[ii] and used directly in the classifier. Both samples were bootstrapped up to 704 equivalent samples and passed to the classifiers. Consequently, four sample sets are analyzed for results: (1) 'two measurements' without bootstrapping, (2) 'two measurements' with bootstrapping, (3) 'all phenotypes' without bootstrapping and (4) 'all phenotypes' with bootstrapping. In both 'all phenotype' cases the data is analyzed for principal components and reduced to ten independent variables.

1 A hierarcheal ensemble using the 'One vs All' technique.
2 The pricision and specificity will be predicted within a 98% or better confidence interval.
3 Precision is measure of predicting positive and specificity is measure of predicting negative accurately.
4 Bootstrapping is resampling with replacement.

Two classifiers are applied to each of the four samples: (1) random forest classifier and (2) probabilistic Bayesian classifier. Random forests is a well established, *non-parametric*[5] technique commonly applied throughout data analysis (see endnotes for references in biostatistics). Probabilistic Bayesian Analysis is the newest Bayes technique founded on the statistical methods favored in medical studies combined with machine learning based estimators[iii].

Analysis Using Random Forest Classifier:

The most useful measure of predictive ability for random forest classifiers is the confusion matrix. This measures the predicted classes, i.e., PD vs Other PD and ET vs PP, against the known diagnosed classes.[iv] In both the raw and bootstrapped 'two component' samples the confusion matrix shows extremely high and confident predictive ability.

Two Measurements Sample

Results For Parkinsons' vs Other Parkinson's Using Raw Sample

Confusion Matrix		
Predicted	**Actual**	
	Negative	Positive
Negative	19	1
Positive	1	23

Confidence Interval: 95%
Precision: 95.4 %
Specificity: 95.0%
True Positive Rate: 95.8%
True Negative Rate: 95.0%

Results For Essential Tremor vs Parkinson's Plus Using Raw Sample

Confusion Matrix		
Predicted	**Actual**	
	Negative	Positive
Negative	9	1
Positive	1	9

Confidence Interval: 95%
Precision: 90.0 %
Specificity: 90.0%
True Positive Rate: 90.0%
True Negative Rate: 90.0%

5 Non-parametric means it relies on measure of entropy rather than statistical distributions

Results For Parkinsons' vs Other Parkinson's Using Bootstrap Training Sample

Training Sample

Confusion Matrix Bootstrap			
Predicted	**Actual**		
	Negative	Positive	Subtotal
Negative	276	0	276
Positive	0	428	428
Subtotal	276	428	704

The normal metrics are all 100.0% at a 99% confidence interval.

Results For Parkinsons' vs Other Parkinson's Using Bootstrap Test Sample

Test Sample

Confusion Matrix Bootstrap Test Sample			
Predicted	**Actual**		
	Negative	Positive	Subtotal
Negative	28	0	28
Positive	0	43	43
Subtotal	28	43	71

The normal metrics are all 100.0% at a 99% confidence interval.

Results For Essential Tremor vs Parkinson's Plus Using Bootstrap

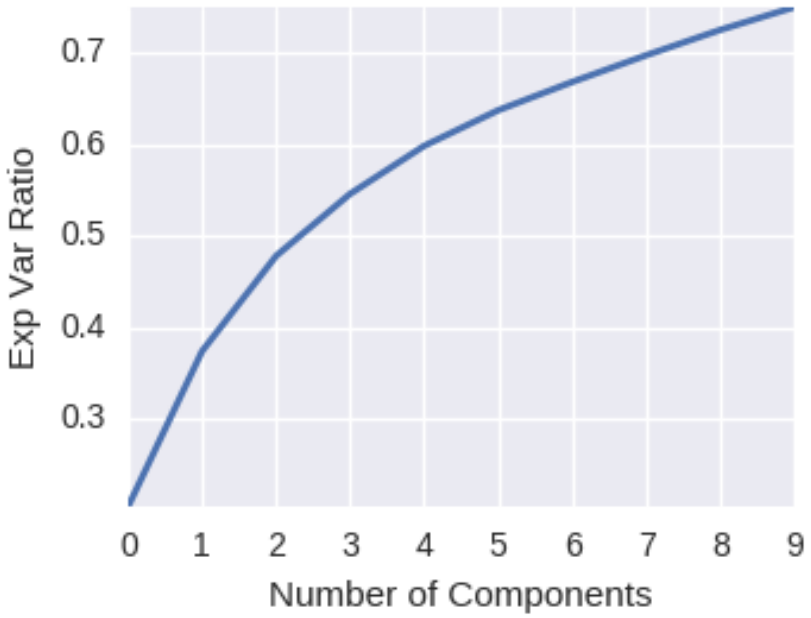
Confusion Matrix Bootstrap ET vs PP			
Predicted	**Actual**		
	Negative	Positive	Subtotal
Negative	357	0	357
Positive	0	283	283
Subtotal	357	283	640

The normal metrics are all 100.0% at a 99% confidence interval.

All_Phenotypes Sample

Principal Companents Analysis('PCA'): The data set for 'All Phenotypes' was reduced to ten major features using PCA techniques. It is found that these ten components explains over 75% of all the variation in the sample.

Comparison of Components And Explained Variance Ratio



The majority of the variation is explained by only 5 components. Whether or not these powerful components represent an important measurement can only be determined by the owners of this study who have the necessary neurological expertise. There may be an additional opportunity in this analysis.

This sample demonstrated the same high performance confusion matrix and metrics as the 'two measurements' random forest classifiers. The method will be retained in the program for future use but it has no immediate application in the random forest classifier.

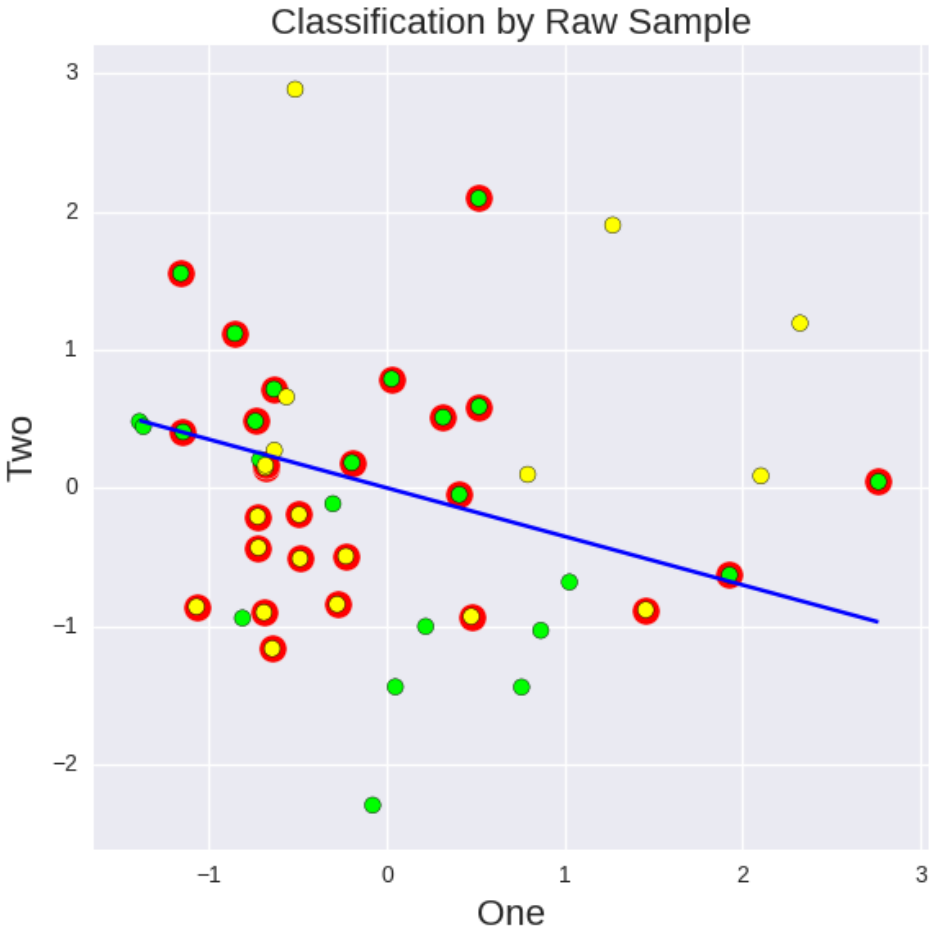
Analysis Using Logistic Probabilistic Bayesian Analysis

Two Measurements Sample

The Bayesian Analysis is not a well suited model for this sample. The raw sample data, i.e., 44 samples divided between Parkinson's and Other Parkinson's did not converge after 2,000 iterations. The results had a p value of less than .15 indicating a very low degree of confidence. A plot of the raw data predictions shows the distributions of the two classes to have considerable overlap based on these measurements.

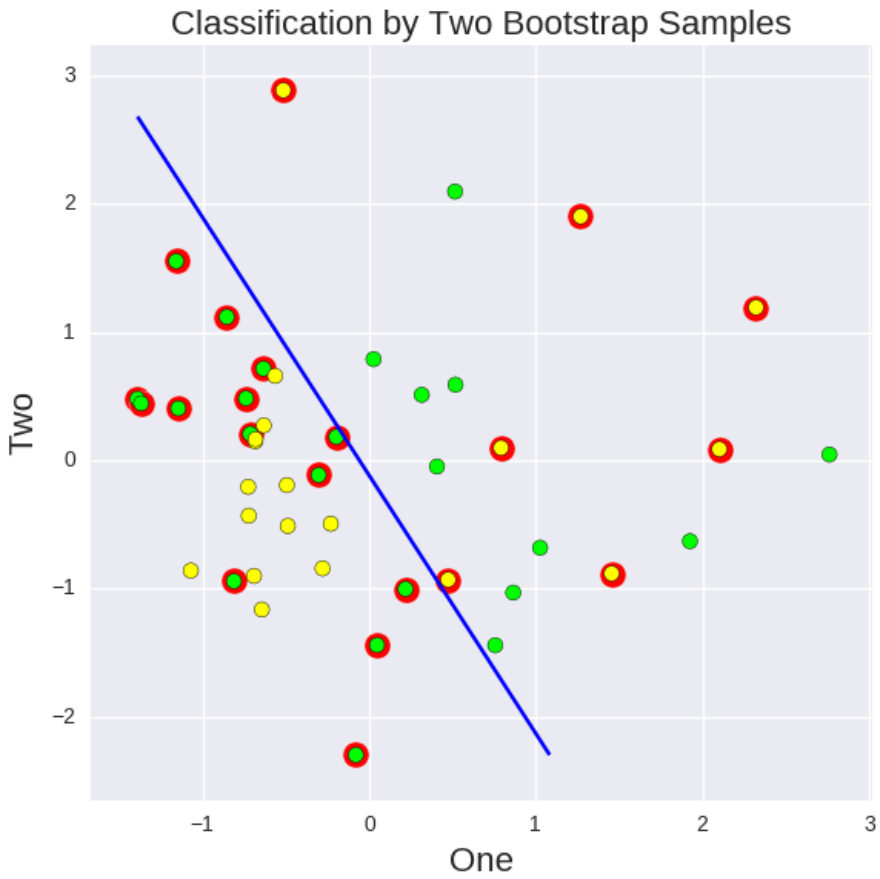
The lime points are Parkinson's and the yellow points are Other Parkinson's. A red perimeter indicates a misclassification on either side of the boundary blue line which is set at the mean of the predicted values.

Bayesian Classification On Raw Two Measurement Sample



The sample bootstrapped to 758 records begins to show some convergence and predictive ability. In order for the Bayesian Analysis on two measurements to be effective the sample must approach 2,500 records.

Bayesian Classification On Bootstrapped Two Measurement Sample



The convergence table for the beta coefficients demonstrates the statistical issue: the confidence interval on all of the expected values is too large.

Bayesian Analysis Bootstrapped Two Measurement Sample

Measurement	Estimate	Confidence Interval @ 95%	
Intercept	60.9	49.5	73.4
One	-48.2	-55.0	-42.4
Two	-185.6	-223.1	-147.4
p	0.5	0.0	1.0

All Phenotypes Sample

The all phenotypes sample was unable to complete a convergence after 4,000 iterations. The All Phenotype sample in raw or bootstrapped form is too large a computational scheme to converge even using multi-processing cores to enable a more powerful computational scheme.

Conclusions:

The Random Forest Classifier on the bootstrapped Parkinson's vs Other Parkinson's sample yielded nearly 100% precision on the sample and the test sample split. It yielded nearly 100% precision on the bootstrapped Essential Tremor vs Parkinson's Plus. The performance on the sample is compelling for further research.

This program is written in Python and Sklearn. They both can be implemented with an interactive menu using Python Bokeh. All three programs are open source and available Both programs are free as open source. The author would be willing to program such an application at the owner's request.

i Citations for using random forests in biostatistics
ii Citations for using PCA in bio statistics
iii Citations for Bayes in biostatistics
iv Citations and further explanation of confusion matrix reference