THE 10TH ANNUAL MLSP COMPETITION: THIRD PLACE

Karolis Koncevičius

Institute of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania.

ABSTRACT

The goal of the MLSP 2014 competition was to automatically detect subjects with schizophrenia based on multimodal features derived from magnetic resonance imaging data. This report summarizes the 3rd place solution with the final ROC area score of 0.91282.

***Index Terms*—** MRI, Schizophrenia, Classification, Distance Weighted Discrimination.

1. INTRODUCTION

Participants of the MLSP 2014 Schizophrenia Classification Challenge were asked to diagnose subjects with schizophrenia based on multimodal features derived from their brain magnetic resonance imaging (MRI) scans. The full dataset had 46 cases (patients affected by schizophrenia) and 40 controls (unaffected individuals) with two sets of features: 32 source-based morphometric (SBM) loadings and 378 Functional Network Connectivity (FNC) features. Submissions were judged based on area under the ROC curve.

The data made available for this task had a peculiar property of having more features than available samples. Such situations are often called "High Dimensional Small Sample Size Data" (HDLSS) [1] in the literature and present a lot of challenges in both model selection and error estimation [2]. To overcome these difficulties I utilized the Distance Weighted Discrimination (DWD) [3] method which was designed to deal with HDLSS settings.

2. METHODS

Distance weighted discrimination was used as a single base classifier. The solution presented in this paper uses all of the available features, runs ten-fold cross-validation to determine the penalty parameter and then fits the DWD model.

2.1. Distance Weighted Discrimination

HDLSS data occupy only a subspace of the whole feature space. The result of this is the existence of linear projection vectors that have the (so called) data piling property [4]. This means that it is possible to project all the samples of both classes onto two distinct points: one for each class. These kind of projections will classify the dataset at hand perfectly, but we should not expect this result to generalize well since the particular configuration of sample points in the feature space can be determined by randomness of sampling.

The DWD approach is to take all of the distances from samples to the separating hyperplane into account. A simple way of allowing all the samples to influence the separation boundary is to minimize the sum of the inverse distances. This gives high significance to those points that are close to the hyperplane with little impact from points that are farther away:

$$\min_{r,w,\beta,\xi} \sum_{i=1} (1/r_i) + Ce'\xi$$

$$r = YX'w + \beta Y + \xi, \quad r \geq 0, \xi \geq 0$$

Here e is a vector of ones, w and β are parameters of the hyperplane, X is the data matrix, Y is a diagonal matrix with class-labels in the diagonal (1 or -1), C is misclassification penalty cost and ξ is the perturbation vector which allows some of the points to be misclassified. In turn r is the vector of distances from a sample point to the separating hyperplane.

In this case the solution that has more points farther away from the separating hyperplane is preferable over the same margin-size solution with data points piled close to the hyperplane.

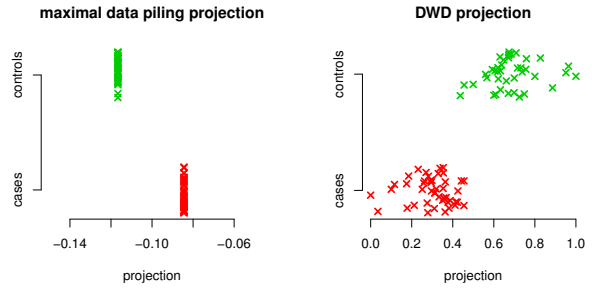

Fig. 1. maximal data piling projection showing complete data piling (left) and DWD projection (right).

Figure 1 provides a simple illustration of potential for data piling in HDLSS settings. It is possible to project all

978-1-4799-3694-6/14/$31.00

the samples of one class onto a single point. DWD avoids such projections by letting all of the points to have a distance-weighted influence on the separating hyperplane while other methods may be influenced by them. As one example consider Fisher's linear discriminant which tries to maximize between-class scatter and minimize within-class scatter[5].

2.2. Feature Selection

All of the available features were used for classification.

A number of unsupervised dimensionality reduction methods were also explored: removing features with low variance, doing Principal Component Analysis (PCA), removing highly correlated features. But all these approaches resulted in a lower area under the ROC curve scores.

2.3. Implementation

The DWD method uses second-order cone programming (SOCP) optimization. It has implementations in matlab and R. Here the R implementation was used [6].

The R version of DWD had one parameter C - the penalty cost associated with misclassification. Authors stated that they don't have clear guidelines for selection of this parameter and that it is a possible object of further studies [3].

In this entry cross-validation was used to determine the value of C. Several values (C=1; 5; 10; 50; 100; 300; 500; 1000) were chosen and 100 iterations of 10-fold cross validation were performed for each of them. Results showed that lower C values had lower classification accuracies and the roc area reached it's maximum at $C = 300$. After that point it saturated and remained unchanged up to $C = 1000$. Therefore $C = 300$ was selected for the final model.

3. RESULTS AND DISCUSSION

Arguably the hardest part in this competition was not overfitting. With a small amount of available samples it becomes hard to track the true error of misclassification. Cross validation may be non-reliable [7] especially if used multiple times on the same data with different models. If no a-priori information is available then simple and highly regularized models become the method of choice [8].

I tried following this philosophy of simple and regularized methods by selecting a few linear classifiers and making sure not to over-adapt to the testing set by trying too many variants of the same method. Among the tried classifiers were regularized linear discriminant analysis, support vector machines and distance weighted discrimination.

The winning DWD method presented in this paper achieved area under the ROC curve scores of 0.9590, 0.8125 and 0.91282 on internal cross-validation, initial test set and final test set respectively indicating quite a stable performance.

4. REFERENCES

[1] Hallm Peter, J. S. Marron, and Amnon Neeman, "Geometric representation of high dimension, low sample size data," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, vol. 67, pp. 427–444, June 2005.

[2] Šarunas Raudys and Anil K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, March 1991.

[3] J. S. Marron, Michael Todd, and Jeongyoun Ahn, "Distance weighted discrimination," *Journal of the American Statistical Association*, vol. 102, pp. 1267–1271, 2007.

[4] Jeongyoun Ahn and J. S. Marron, "The maximal data piling direction for discrimination," *Biometrika*, vol. 97, pp. 254–259, 2010.

[5] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[6] Hanwen Huang, Xiaosun Lu, Yufeng Liu, Perry Haaland, and J.S. Marron, "R/dwd: distance-weighted discrimination for classification, visualization and batch adjustment," *Bioinformatics*, vol. 28, pp. 1182–1183, April 2012.

[7] UM Braga-Neto and ER Dougherty, "Is cross-validation valid for small-sample microarray classification?," *Bioinformatics*, vol. 20, pp. 374–380, 2004.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.