Schizophrenia Classification using MRI Scans

Final project report for Math 6450A Statistical Machine Learning

MCDONNELL Serena Man-Yon (SID: 20299217)
LO Yi-Su (SID: 20399988)
Department of Mathematics, HKUST

May 18, 2017

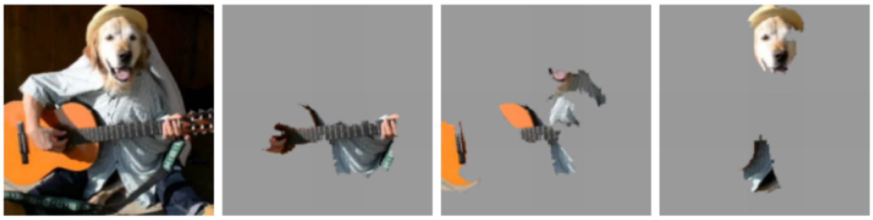
1 Motivation: Machine Learning and Medicine

Machine learning (ML) has been widely stated to have the potential to revolutionize medicine. By learning rules from data, a machine learning algorithm learns much like a young doctor does throughout the course of medical school and residency. The ability of ML algorithms to accurately provide medical diagnoses could revolutionize medicine as we see it today.

The implications are varied: while some argue we could eliminate the need for certain physicians with the use of ML, others say that it'll never catch on.

In order for machine learning to become effective and widely used in medical diagnoses, it'll be necessary to understand the logic behind the diagnosis, or classification. It's much easier to trust a computer that tells us a patient's tumour is malignant when we're told why, rather than just the plain fact. For example, Local Interpretable Model-Agnostic Explanations, or LIME, is a newly released project that aims to explain the predictions of any machine learning classifier. We'll need many more projects like LIME in order to make the medical community completely trust computers.

The elimination of medical doctors through the use of machine learning seems unlikely. What is likely, however, is that through the use of ML's diagnostic capabilities, physicians will have their time freed up to focus on other important aspects of work, such as comforting patients and forming relationships, which is important in medicine but often overlooked by busy physicians.



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Figure 1: Figure from the LIME paper.

Overall, as machine learning algorithms become more sophisticated, its use in our lives will become more prominent, and surely this will apply to the medical field as well. Whether this results in elimination of doctors, or a rejection of the technology by existing ones, remains to be seen.

2 Problem and data set

2.1 Introduction to the Kaggle competition

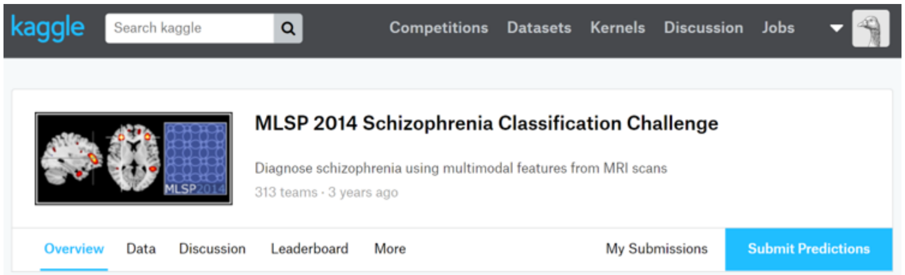
Kaggle is a website where competition "hosts" (companies, researchers) post their data, for individuals and teams throughout the world to produce the best models and submit their results. Once a competition closes, the later submissions might be accepted but the scores they receive won't be issued.

The host for the data set we have chosen is the IEEE International Workshop on Machine Learning for Signal Processing (MLSP). This workshop took place from September 21-24, 2014, in Reims, France.

In the MLSP 2014 Schizophrenia Classification Challenge competition (`https://www.kaggle.com/c/mlsp-2014-mri`), the task is to automatically diagnose subjects with schizophrenia based on features derived from MRI scans. Winners of the competition were offered the chance to attend the workshop in Reims.

2.2 Schizophrenia

Schizophrenia, derived from the Greek skhizein (to split) and phren (mind), is a mental disorder where the person affected thinks, and acts in abnormal ways, with behaviour often being described as "out of touch." Common symptoms



MLSP 2014 Schizophrenia Classification Challenge

Diagnose schizophrenia using multimodal features from MRI scans

313 teams · 3 years ago

Figure 2: A portion of the homepage for the Kaggle competition of interest. (`https://www.kaggle.com/c/mlsp-2014-mri`)

include delusions, thought and behaviour disorders, reduced feeling, and overall difficulty functioning in regular society. The onset of the disease usually occurs between ages 16 to 30, and is believed to be due to a combination of genetic and environmental factors. An imbalance in the chemical composition of the brain is also believed to lead to schizophrenia.

2.3 Magnetic Resonance Imaging

Certain medical examinations, such as *Magnetic Resonance Imaging* (MRI) have been used to identify subtle differences in the brain structures of schizophrenic individuals. Commonly affected areas are the frontal lobes, hippocampus, and temporal lobes.

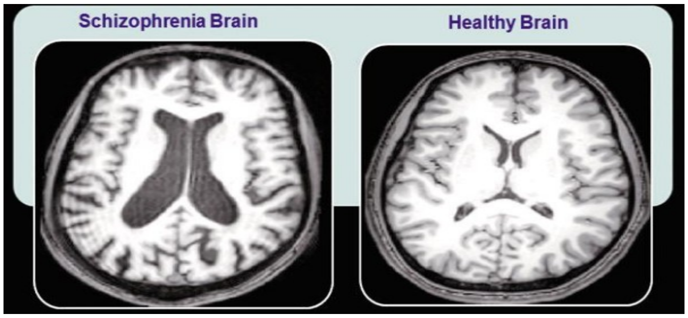


Figure 3: An MRI image of a schizophrenia vs healthy brain.

An MRI scanner uses magnetic fields, radio waves, and radio gradients to generate images of organs in the body. As MRI does include the use of x-rays, it is seen as a safer alternative to computed tomography, or CT/CAT scans.

MRI technology is often divided into structural and functional MRI. Structural MRI provides information on the structure, shape, and size of a body organ, while functional magnetic resonance imaging, or fMRI, uses MRI technology to measure changes in blood flow in the brain. When an area of the brain is in use, blood flow to that region is increased. In this way, fMRI presents a way of determining the relative usage of certain brain structures in schizophrenic patients, compared to unaffected individuals.

2.4 The Data Set

In the competition, both of the training and test data are supplied. The training data set is originated from 86 patients who are labeled with 'Schizophrenic' (40/86) or 'Healthy' (46/86). The test data set include the data of $119,748$ patients. To prevent competitors from cheating, the host added a lot of fake data into the test set and it leads to a much larger test set.

There are two types of features in both of the training and test data set: *functional network connectivity* (FNC) and *source based morphometry* (SBM). A fair bit of medical knowledge is required to fully understand what FNC and SBM mean exactly. In brief, FNC features are correlation values that describe the connection between different structures in the brain over time. SBM features, on the other hand, are standardized weights that describe the expression of brain maps, which are maps denoting independent structures in the brain composed of gray matter. In the brain, the majority of processing occurs in gray matter.

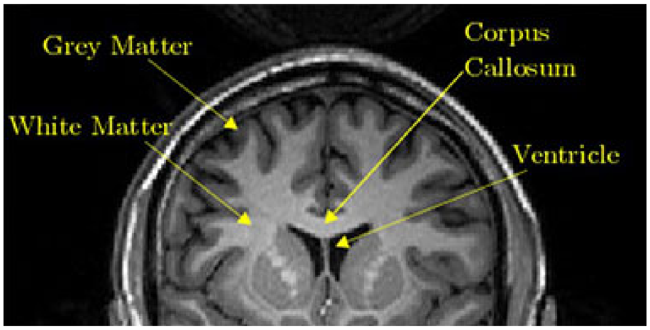


Figure 4: Gray vs white matter in the brain.

Source Based Morphometry (SBM)

SBM is a method used to determine the connection between brain structure and brain function. SBM loadings are coefficients that correspond to the weights of the brain maps obtained from independent component analysis (ICA), essentially giving information on how much processing occurs within a given region in the brain. ICA is similar to principal component analysis in that it finds a new basis in which to represent the data.

The ICA algorithm is as follows:

$$x = As,$$

where s $\in \Re^n$ is some data generated via n independent sources or components and A is the mixing matrix. Let

$$W = A^{-1}$$

be the unmixing matrix. We search for the matrix W so that we can recover sources or in our case, components of the brain, by computing $s = Wx$. Next, let

$$p(s) = \prod_i^n p_s(s_i),$$
$$p(x) = \prod_i^n p_s(w_i^T x)|W|,$$

where s_i is the source distribution, p_s is its density, $p(s)$ is the joint distribution of sources s, $p(x)$ is the density on $x = As = W^{-1}s$, and $p_s(s) = g'(s)$, where $g(s) = 1/(1 + e^{-s})$. For a given a training set $\{x^{(i)} : i = 1, ..., m\}$, the log likelihood is given by

$$l(W) = \sum_{i=1}^{m}(\sum_{j=1}^{n} \log g'(w_j^T x^{(i)}) + \log |W|).$$

The goal is to maximize this function in terms of W.

SBM loadings are found as follows:

1. Obtain GMD information for each subject.
2. Flatten GMD information into a *subject* x *voxel* matrix, where a voxel is the volume gray matter in a given region.
3. Apply spatial ICA to obtain two matrices: one is a *subject* x *component* matrix, and the second is a *component* x *voxel* matrix.

The first matrix obtained in step 3 is the one that we are given by Kaggle. Each component is called an SBM map. 32 components are chosen for the purposes of the competition, but 60 and 80 maps work well, too.

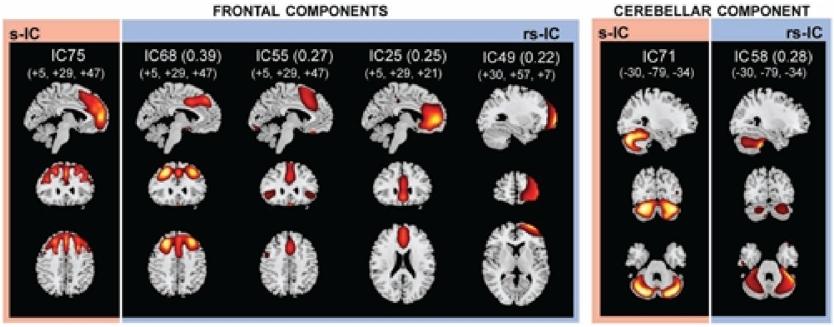


Figure 5: Example of independent components.

Functional Network Connectivity (FNC)

A functional network refers to areas in the brain that interact to perform certain functions. Functional networks can be determined through the use of fMRI scans, to look at blood-oxygen-level-dependent fluctuations in the brain. Scans are typically done on regions of high gray matter density (GMD), since the majority of processing in the brain occurs in gray matter.

More specifically, there is a great deal of temporal variability in fMRI scans, which is largely overlooked. Most approaches examine functional connectivity (FC), or activity between regions, as constant throughout an entire fMRI recording.

FNC values are found by:

1. Obtain GMD information for each subject.

2. Use ICA to decompose GMD information into functional networks, identify time courses of activity, and identify subject specific spatial maps. The functional networks were expected to show peak gray matter activity, and have time courses dominated by low frequency fluctuations

3. Estimate dynamic FC: use sliding window approach to obtain sample covariance matrix from time course matrix.

4. Apply k-means clustering to the windowed covariance matrix obtained in step 4.

From this, the temporal variability within fMRI recordings can be determined. In the data set, 378 FNC components were obtained. FNC highlights the connections between different structures in the brain, and how the activity of certain structures of the brain evolves over time in an fMRI scan.

3 Method

3.1 Overview of the algorithm

We briefly describe our algorithm for this problem in the following.

1. Load and preprocess the data.
2. (optional) Carry out the principle component analysis to reduce the dimension of data points.
3. Perform a model selection process for discovering appropriate parameters for the following training.
4. Use the support vector machine (SVM) approach with the Gaussian radial basis function kernel to train a classifier.
5. Make predictions for the test data set with the trained SVM classifier and transform the classification score into a real number in $[0, 1]$ as the probability of illness.

The detail for every step is given in the next subsection. Note that in the implementation, we adopt the popular `Python` package `scikit-learn`. A link for accessing our source code is provided in the Appendix B.

3.2 Ingredients of the algorithm

Data preprocessing

As mentioned in the last section, there are totally 32 SBM components and 378 FNC components obtained from each patient's MRI scans. After a little inspection, we find the ranges of all SBM and FNC values are $[-0.446, 1.641]$ and $[-1, 1]$, respectively. Since they are quite close to each other, we don't conduct any preprocessing procedure but put the two features into one set directly. The resulting training set contains data for 86 patients and every data point has dimension 410 (378 FNC + 32 SBM). Similarly, the test set comprises $119,748$ data points whose dimension are all 410.

On the other hand, we notice that the numbers of ill and healthy patients in the training set are at the same scale (40 vs. 46). In the case the sizes of different classes are not balanced, an introduction of additional strategy might be necessary.

Principle component analysis (PCA)

PCA is a traditional but effective procedure for reducing the dimension of data points. For a data set $\{x_i : i = 1, 2, \dots, n\}$, it solves

$$\min_{\mu,\beta_i,U} \sum_i \parallel x_i - (\mu + U\beta_i) \parallel,$$

where μ is the sample mean i.e. $\mu = \frac{1}{n}\sum_i x_i$, U consists of the *principle components* and $\sum_i \beta_i = 0$. In the case the number of principal components is assigned as k, it is equivalent to

$$\min_L \parallel X - L \parallel \qquad \text{s.t.} \qquad \text{rank}(L) \le k,$$

where X is the data matrix collecting all data points x_i.

Classifier training with support vector machine (SVM)

In a training data set $\{(x_i, y_i) : y_i = \pm 1, i = 1, 2, \dots, n\}$, the *hard-margin* SVM could be regarded as solving the optimization problem

$$\min_{\beta,b} \frac{1}{2}\beta^T\beta \qquad \text{subject to} \qquad y_i\, f(x_i) \ge 1,\ \forall\, i,$$

where $f(x_i) = \beta^T\phi(x_i) + b$. Once the optimizer $(\hat{\beta}, \hat{b})$ is determined, the value of $\hat{f}(x) = \hat{\beta}^T\phi(x) + \hat{b}$ presents the distance from any data point x to the hyperplane/decision boundary, as well as the *classification score* of x.

Usually, the data points might not be perfectly separated by any hyperplane. In this case we may adopt the *soft-margin* SVM which solves

$$\begin{aligned} &\min_{\beta,b,\xi_i} \frac{1}{2}\beta^T\beta + C\sum_i \xi_i^2 \\ &\text{subject to} \begin{cases} y_i\, f(x_i) \ge 1 - \xi_i, \\ \xi_i \ge 0, \end{cases} \quad \forall\, i, \end{aligned}$$

where ξ_i's serve as *slack variables.* C is often called the *box constraint* in which higher value of C allows a higher tolerance to the classification errors. By considering the Lagrange function of the problem

$$\mathcal{L}(\beta, b, \xi_i, \alpha_i, \mu_i) = \frac{1}{2}\beta^T\beta + C\sum_i \xi_i^2 - \sum_i \alpha_i(y_i f(x_i) - (1 - \xi_i)) - \sum_i \mu_i\xi_i$$

and the KKT condition $\nabla \mathcal{L} = 0$, we derive the dual problem

$$\begin{aligned} &\max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ &\text{subject to } \begin{cases} y_i \alpha_i = 0, \\ 0 \leq \alpha_i \leq C, \end{cases} \quad \forall\, i, \end{aligned}$$

where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the *kernel function* and the inequality of α_i explains the reason we call C a box constraint. In our algorithm, we adopt the popular *(Gaussian) radial basis function* (RBF) kernel

$$K(x_i, x_j) = \exp\left(-\gamma \parallel x_i - x_j \parallel^2\right),$$

where γ is the scaling parameter.

Model selection

As mentioned, there are two parameters for the SVM classifier training: box constraint C and γ in the RBF kernel. To tune these two parameters, we apply the techniques of *grid search* and *5-fold cross validation*:

1. Lay down a uniform mesh of $m \times m$ grid points on a certain C-γ domain.
2. Divide the training set into five subsets.
3. At every grid point (C_k, γ_k), $k = 1, 2, \ldots, m$, we train five SVM classifiers with parameters C_k and γ_k. Each classifier is trained in four subsets and evaluated with the error rate in the remaining one. The overall error rate of the five classifiers is taken as objective function value at (C_k, γ_k).
4. After totally $m \times m \times 5$ trainings, the grid point with minimum objective function value is chosen for the ultimate classifier training in the whole training set.

In our work, we empirically restrict the C-γ domain within $[10^{-1}, 10^4] \times [10^{-4}, 10]$. There are two trial values for the mesh size, $m = 8$ and $m = 16$.

Platting scaling

After the classifier training, we are able to classify the data points of the test data set with the obtained SVM classifier. However, because the competition demands the probabilities of illness as the prediction result, it is necessary to transform the classification score $f(x)$ into a real number between 0 and 1.

`Scikit-learn` provides a classical manner, *Platt scaling*, for this purpose. The transformation is completed by fitting a logistic regression model to the classification scores. To be more specific, the probability is obtained through the *Sigmoid function*

$$P_{A,B}\left[f(x)\right] = \frac{1}{1+\exp\left[Af(x)+B\right]},$$

where A and B are determined by the maximum likelihood problem

$$\min_{A,B} -\sum_{i=1}^{n}\left[t_i\log(p_i)+(1-t_i)\log(1-p_i)\right],$$

where $p_i = P_{A,B}(f(x_i))$ and

$$t_i = \begin{cases} \frac{N_+1}{N_++2}, & \text{if } y_i = 1, \\ \frac{1}{N_-+2}, & \text{if } y_i = -1, \end{cases}$$

N_+ and N_- represent the number of positives and negatives (i.e. ‘schizophrenic’ and ’healthy’ in our case) in the training set.

3.3 Result evaluation: Receiver operator characteristic (ROC) curve

As the last step, we submit the prediction results to Kaggle. According to the competition’s rule, the accuracy/score of every submission is measured by the *area under the ROC curve* (AUC). Once a probability threshold for the classification of positives (schizophrenic cases) and negatives (healthy cases) is determined, such as 0.5, the true positive rate (TPR) and false positive rate (FPR) can be computed through the formulas below. Note that the true positive and false positive are two of the four kinds of classification errors, see Table 1.

$$\text{TPR} = \frac{\text{number of samples which are actually positive and classified correctly}}{\text{number of samples which are actually positive}}$$

$$\text{FPR} = \frac{\text{number of samples which are actually negative but classified as positive}}{\text{number of samples which are actually negative}}$$

If one starts with the data point which has the lowest predicted probability and then add data points in the ascending order of probability, the TPR and FPR increases alternately as the number of data points increases. The ROC curve, whose y- and x-values are the TPR and FPR at a certain number of data points, is drawn as Figure 6 shows. As one can observe, a curve with larger AUC illustrates the TPR grows much faster than FPR, which suggests a good performance of the classification method.

		Truth	
		Positive	Negative
Prediction	Positive	True positive	False positive
	Negative	False negative	True negative

Table 1: Four types of classification errors.

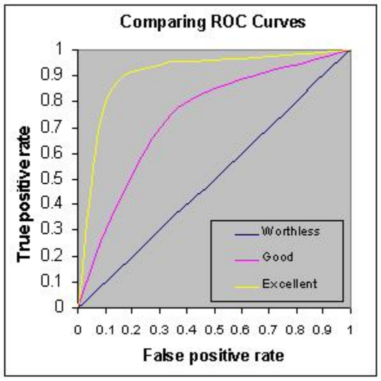


Figure 6: Example ROC curves. Source page: `http://gim.unmc.edu/dxtests/roc3.htm`

4 Result and Discussion

4.1 Result

Since Kaggle limits the number of submissions per day for every user, the competition host further split the test data set into *public* and *private* subsets. Every competition participant could submit unlimited number of results into the preliminary stage and receive public scores (i.e. the AUC score with regard to the public subset) immediately, then choose up to two submissions per day entering the assessment of private scores (i.e. the AUC score regarding the private subset). However, since the competition already closed, the Kaggle users who submit results now can receive both scores at the same time.

We summarize the scores and training time of our main results in the Table 2. For comparison, the Public Leaderboard and Private Leaderboard which ranked all 313 competitors by the scores they received, when the competition was ongoing, are available at `https://www.kaggle.com/c/mlsp-2014-mri/leaderboard` (both of the two Leaderboards are no longer updated now).

We briefly draw some conclusions from Table 2.

Model sel. mesh size	Item	Without PCA (d=410)	PCA (d=16)	PCA (d=32)	PCA (d=64)
–	Explained variance	100%	62%	82%	97%
8×8	Public score	**0.85714**$^{\#}$	0.70982	**0.90179***	0.86161
	Private score	**0.88718**$^{\#}$	0.79847	**0.84615***	0.81026
	Training time (sec)	8.608	4.990	6.773	7.319
16×16	Public score	0.79911	0.72678	0.83036	0.83036
	Private score	0.82051	0.80769	0.84103	0.80153
	Training time (sec)	10.821	5.870	7.928	9.871

Table 2: The classifier training time and the scores received from Kaggle under different algorithm settings.

- The scores with mark * are the best scores, on the average. If we were one of the competitors, the public score of it would rank 28th, while the private score would rank 104th on the Leaderboards. On the other hand, our second best scores with mark $^{\#}$ would rank 89th and 22nd, respectively.

- The use of 32 principal components not only reduces the dimension of data (and hence the training time) but also improves the accuracy of classification. In comparison, 16 components may not explain sufficient variance and 64 components might contain too many insignificant features, therefore result in lower scores.

- Finer mesh for model selection doesn't seem to help. 8×8 is quite enough in our case. But different grids i.e. different set of parameter candidates do influence the obtained classifier, as well as the prediction result.

4.2 Future of Machine Learning in Medical Research

Since the competition date three years ago, progress has been made in using ML for medical research. Convolutional neural networks (CNNs) can now be used for various tasks. They can classify white blood cell shape as polynuclear or mononuclear with 98% accuracy, diagnose an eye disease called diabetic retinopathy in just twenty seconds, and classify skin cancer as well as trained dermatologists. Machine learning in medicine certainly is a powerful tool, and is only becoming more popular with time.

Appendix A. References

NIH Schizophrenia
NIH MRI
Tracking Whole-Brain Connectivity Dynamics in the Resting State
Correspondence between structure and function in the human brain at rest.
CNN for WBC classification
ICA

Appendix B. Data and source code

All the `Python` source code, as well as the data sets from the Kaggle competition of interest, are available for accessing at `https://drive.google.com/drive/folders/0B3zEmYEa2Wc5NGZqbGNoc3ZPUzg?usp=sharing`.