RA Report: Application of SVM to Landscapes of ABIDE Dataset

Avani Sharma (u1065441)

Spring 2017

1 Introduction

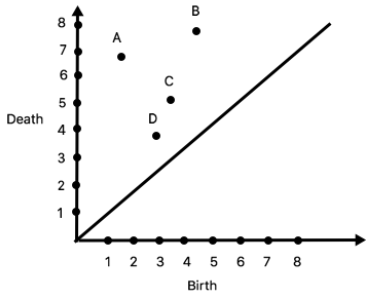
This Report corresponds to my research work in Spring 2017. We worked on classification of functional brain networks using topological features. The classification was mainly related to segregating autistic patients from control patients. My focus was mainly on using persistent landscapes in order to extract the features. Data sets were moulded according to them. Later on SVM (Support Vector Machine was) used for classification. We also used combination of base line and landscape features in order to classify the subjects (data points) and we could see there was 1% to 2% improvement over baseline. We also worked upon other data set where again the task was to classify subjects based on gender. We tried using Multi Dimensional Scaling. Again the plots did not show much segregation but yeah could get a brief idea of how points lie in 3 dimensional space. We are also working on using landscapes and support vector machine to see if we can segregate points based on gender. Remaining sections entail the details and step wise approach, that we followed for this research.

2 Brief explanation of Concepts

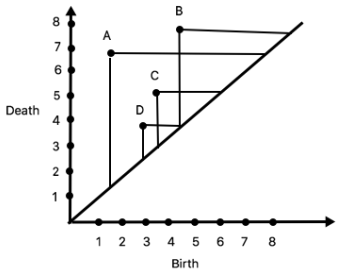
1. **Landscapes**

   For classification of brain networks we are using another set of topolological features called persistence landscapes. Like persistence barcodes and persistence diagrams, persistence landscapes also give us topological summary of given data. To get persistence landscapes from persistence diagrams, for each (birth,death) point draw the lines starting from it, parallel to x-axis and y axis respectively, touching the diagonal as shown in Figure 1 (b). After this persistence landscapes can be simply obtained from persistence diagrams by tilting the diagonal such that it becomes x-axis as depicted in Figure 1 (c). Now in new coordinate system we get piece wise linear functions know as persistence landscapes which form envelopes as shown in Figure 1 (d) . Each (birth,death) point is mapped to ((birth+death)/2,(death-birth)/2). The outer most envelop is called order 1 landscape, As we moved inside, we get lower order landscapes as shown in Figure 1 (c) and Figure 1 (d) . The largest persistence point (death-birth) corresponds to order 1 landscape. Points with same birth and death values have 0 value for landscape function because they lie on x-axis. This can be easily visualized in Figure 1 (d) .

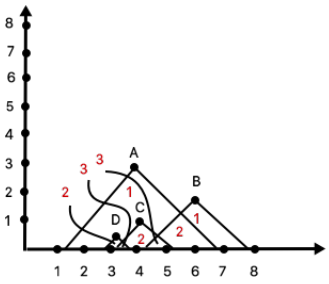
2. **Multi Dimensional Scaling (MDS)** Multi Dimensional Scaling is a way of visualizing, how different data points lie in the space in lower dimensions. It is a quick test that can tell major differences between the data points of different classes. It is a non linear dimensionality reduction technique. It takes input as distance matrix and uses that information to place all data points in space such that the between points distances are are well preserved. Points can be visualized in 2D or 3D space however, if sole aim is dimensionality reduction and not visualization the dimension of points may increase beyond 2 or 3. Generally scatter plot is used for visualization of data points in space. Details about MDS can be found in [5]



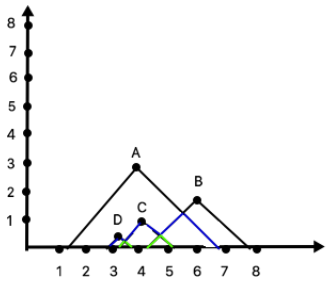
(a) Persistence Diagram



(b) Draw Horizontal and Vertical Lines



(c) Persistence Landscapes (red denotes order)



(d) Persistence Landscapes (black = order1, blue = order2, green = order3)

Figure 1: Persistence Landscapes

3 Data Sets

1. We are using the data collected from 3 of the ABIDE sites: UCLA, NYU and USM following [1]. We filtered the subjects using IQ ($\geq$ 80) and age (mean +/- 1*std within group). The age filter was applied to ASD and control groups separately. Therefore, the age range for control subjects was : (11.92, 23.24) and for ASD subjects it was (11.25, 22.37). After this, we eliminated subjects which had missing nodes (due to our preprocessing) Finally, we were left with 175 subjects, 88 control and 87 ASD. All the classification experiments are being run on this data set.

2. For the first dataset listed above, we were trying to classify these subjects using topological features (persistence diagrams and landscapes) and machine learning. However we couldn't get significant improvement over baseline as data set was quite noisy, so we shifted to Human Connectome Dataset. Data consists of 838 control subjects. There is time series data, correlation matrices, and behaviour metrics for these subjects. Specifically we are looking at correlation matrices of 838 subjects, and we want to classify those subjects according behaviour metric gender (male, female).

4 Approach

1. The approach that we are using here is extracting topological features and using those features as input to machine learning algorithm (support vector machines) for classification. Among topological features we are trying to generate persistence diagrams and landscapes for each subject. Persistence diagrams correspond to features such as connected components (dimension 0), holes (dimension 1), along with their birth and death, as simplicial complex grows during filtration. Landscapes is the piecewise linear function obtained by, tilting the diagonal in persistence diagram, so that it lies on x axis. Points are sampled from these piecewise linear functions which act as landscape features and can be used as input to machine learning algorithm. We are using the landscape kernel described in [2], but we have limited landscape orders till 4. This is because the higher

order landscapes do not contain much relevant information. Order one is most relevant among all. This was visible through experiments too and there was not much observable difference while adding other orders to order 1 during landscape kernel computation.

2. We also tried doing a quick test on the dataset, using multidimensional scaling in order to see which brain networks are similar and fall into same group (cluster).

5 Experiments

- **Persistent Landscape**

  We computed persistence landscapes for each subject using corresponding persistence diagram. There is a function called landscape (Diag, dimension, KK, tseq) in TDA-R package which is used to compute persistence landscapes from persistence diagrams. Here Diag = Persistence Diagram, dimension = dimension of features (In our case its dim 1 features only), KK is order of landscapes we are using (Here we are using upto order 4 landscapes for our experiments), tseq = number of discretized points that are sampled from the landscapes function (For each order we are sampling 2000 points). By landscape order we mean:

  1. 1st order landscape = we sample the function of 1st order at 2000 equidistant points in the filtration range (0, $\sqrt{(}2)$).
  2. 1st to 4th order landscapes = we sample the function of each order at 2000 equidistant points in the filtration range (0, $\sqrt{(}2)$) and append them.

  We performed our experiments with: landscapes (2000 - 8000 features values) only and (baseline + landscapes) (34716 - 42716 feature values) .Kernel that we are using over these landscapes is linear kernel which is defined as dot product between feature vector of pair of subjects. Since we are running SVM with linear kernel we have only one parameter C to tune for landscapes only and two parameters (C, w) to tune for ((w)baseline +(1-w) landscapes).Values for

  C = [0.0000001,0.000001,0.00001,0.0001,0.001,0.01,0.1,1,10,100,1000,10000,100000,1000000]. The values for w range in between (0,1) with 0.1 steps.

- **Classifier: Support Vector Machine**

  I have used standard SVM implementation in Scikit Learn for classification for both datasets. More details about SVM can be found in [4].

- **Cross Validation**

  1. **Leave One Out** As we had total of 175 subjects, we used leave one out (LOO). Here we leave one subject out as test data and trained the model on the remaining 174 subjects and use this model to classify the test subject. Prediction accuracy of the model is either 0 or 1. We repeated this process, leaving out each subject in turn so that in the end we have 175 trained models and each is used to predict label for one test subject. The final accuracy we obtain is the average accuracy of the 175 models.
  2. **Training model** We trained the model (on 174 subject training data) using LOO to tune the parameters as follows: For each parameter combination, we iteratively leave one subject out as test, train the model on remaining 173 subjects and predict label on test subject, taking each of the 174 subjects into consideration. We average the accuracy of 174 models - this is the cross validation accuracy for the given parameter combination. This procedure is repeated for each parameter combination, and at last the combination with highest cross validation accuracy is picked. This parameter combination is used to train the model on entire training data (174 subjects). This forms the final training model.
  3. **Kernel**
     (a) **For Baseline**
     For baseline we use vectorized correlation matrices as feature vectors and trained linear SVM classifier.

(b) **For TDA**
We used Landscapes as features and trained Linear SVM over them.

Since we are using linear kernel here and only parameter required is C which is the SVM parameter.

- **Multi Dimensional Scaling Implementation**

  For MDS results for HCP Dataset, We used Scikit Learn's MDS implementation [3] which takes input as distance matrix and number of components and gives output as embedding of the data points in lower dimensional space having the given number of components while keeping the distances between points preserved. I used three components and thus, MDS embedded the points in 3 dimensional space. This is the quick test to see how points lie in space.

6 Results

Accuracy for different Experiments:

1. **Baseline:** 0.70285714
2. **Using only Persistent Landscapes:**
   - Order 1 : 0.44571428
   - Order 2 (order1 + order2) : 0.60571428
   - Order 3 (order1 + order2 + order3): 0.56571428
   - Order 4 (order1 + order2 + order3 + order4) : 0.52571428

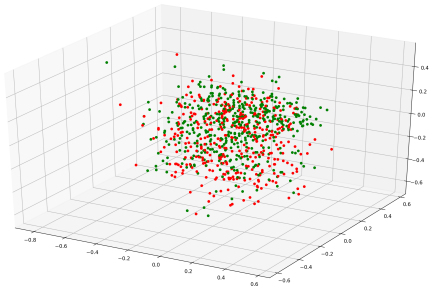
   Linear kernel over these landscapes simply computes the dot product between corresponding order landscapes and adds them for example for any two landscapes uptill order 4 computes (1st*1st' + 2nd *2nd' + 3rd *3rd' + 4th * 4th') Order 1 + 2 seems to give significantly better classification than just order 1. But including order 3 and order 4 landscapes somehow makes it worse. In any case, these results are nowhere near baseline.
3. **Using Baseline + Landscapes:**

   These are the results for combined (w * baseline + (1 - w) * landscape) kernels. The weights go from 0 to 1 in steps of 0.1. In most cases, the trained models have greater weight on baseline kernel (> 0.7).

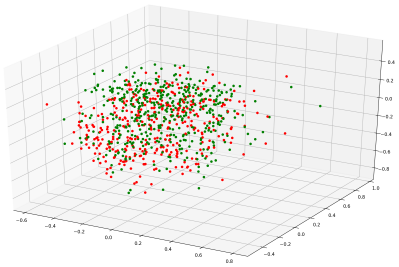
   But the landscape kernel has non zero weight in majority of the cases.
   - Baseline + Order 1 (order1) : 0.71428571
   - Baseline + Order 2 (order1 + order2) : 0.70857142
   - Baseline + Order 3 (order1 + order2 + order3) : 0.71428571
   - Baseline + Order 4 (order1 + order2 + order3 + order4) : 0.71428571

   There is a slight improvement over baseline. The improvement here corresponds to correct classification of 2 more subjects.
4. **Using MDS on Human Connectome Dataset**

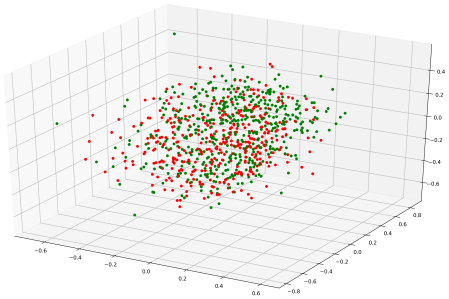
   The results are shown in Figure 2. In all the four cases, the points seem pretty mixed, with no observable difference between the male and female group.



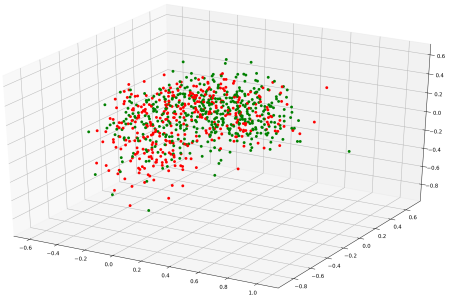
(a) MDS results for Order 1 Landscapes



(b) MDS results for Order 1 + 2 Landscapes



(c) MDS results for Order 1 + 2 + 3 LandscapesMDS results for Order 1 Landscapes



(d) MDS results for Order 1 + 2 + 3 + 4 Landscapes

Figure 2: MDS results

7 Conclusion and Future Work

From the experiements, it is clear that there was not much improvement over baseline. It might be because features are not discriminant enough. We need to extract some discriminative features from the dataset in order to achieve good classification accuracy. We are trying to find the new approaches and trying some other behaviour metrics which can help us for proper classification of dataset.

Future Work: We are still Working on Classification using Landscapes on Human Connectome Dataset. We think that using TDA features along with machine learning will let us discriminate between subjects based on their gender. But still we need to run experiments in order be assured of it. We might try using other behaviour metrics such as age, IQ, etc. in order to get greater insights from the data about subjects.

References

[1] Abide Classification Paper, $https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4309950/$

[2] Xiaojin Zhu, Ara Vartanian, Manish Bansal, Duy Nguyen, Luke Brandl, "Stochastic Multiresolution Persistent Homology Kernel", University of Wisconsin and Madison

[3] Thomas Finley, Thorsten Joachims, "Supervised k-Means Clustering", This work was supported under NSF Award IIS-0713483 "Learning Structure to Structure Mapping," and through a gift from Yahoo! Inc.

[4] "Scikit Learn Multi Dimensional Scaling", http://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html

[5] Cristianini, Nello, Shawe-Taylor, John; An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000. ISBN 0-521-78019-5 (SVM Book)

[6] Wickelmaier, Florian, "An introduction to MDS." Sound Quality Research Unit, Aalborg University, Denmark (2003): 46