2019-01

Software Capstone Project

CSI 4102 (Sector 3)

Recommending Candidate Association between Brain Activation and Behavior

: based on Collaborative Filtering Approach

Advisor : 박상현 교수님

Assistant : 하지환 조교님

Data Engineering Laboratory

Team BRAIN

Junsol Kim, Hyeongjoon Kim, and Jiyun Kim

Table of Contents

1 Research Topic

We aim for predicting the candidate associations between brain activation and human behavior which are unknown. To do so, we apply collaborative filtering approach on the Neurosynth dataset which integrated findings on brain activation and human behavior by integrating neuropsychological findings.

2 Prior Research and Research Goal

2.1 Bioinformatics

Bioinformatics is a field of study which collects, specifies, stores, and analyzes biochemical and biological information using computers and informatic skills. The brightest area in the field is regarding molecular genetics and genomics. The field is highly specialized in genetics but the scope has got larger to other areas such as brain analysis. Meta-analysis in a large scale to define correlations between brain and behavior has been a research topic since near past. (Yakoni, 2012)

There has been studies about meta-analyzing correlations between each brain region and behaviors. ALE, MKDA are the ones which get the most public confidence. (Caspers, 2010, Chang, 2012) However, these methods have a critical limitation that correlations which are unknown or not yet studied are under the shade of ignorance. In other words, they only focus on analyzing the dataset of published topics and suggest a macroscopic view of them. Prior studies regard that it is impossible to estimate correlations unknown or unobserved.

However, collaborative filtering has already been used to find out new gene-disease correlation in the field of genetics. Despite its early introduction, it is notable that brain science rarely made an observable progress soaking the new skills from computer science and using recommendation algorithm. The techniques have used within a limited sectors such as estimating brain tumor. (Ortega-Martorell et al., 2012)

2.2 Keyword-voxel Cluster Matrix

By applying matrix factorization method, we intended to find unknown relationship between a voxel cluster and a keyword. The basic concept of matrix factorization is split original matrix dataset into latent factor matrix for user and item. In our case, user is a keyword and an item is a voxel cluster id. Latent factor matrix means feature matrix which can describe the user of item preference.

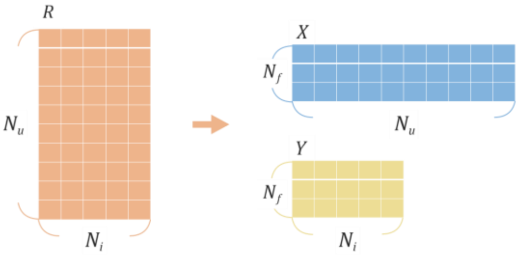

Figure 1. Matrix factorization - split latent factor matrix

By multiplying the transpose of user latent factor matrix and item latent factor matrix, we can predict relationship score matrix. Comparing predict relationship score with original matrix, we can get loss function.

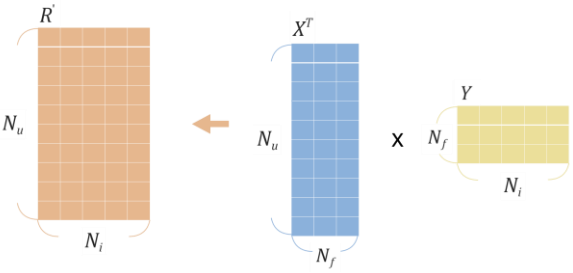

Figure 2. Matrix factorization Concept

We can subtract predicted value with original data to calculate accuracy of our prediction. Then latent factor matrix can be trained in direction which can minimize the difference between prediction and original data. Adding L2-normalization for regularization, loss function can be get.

$$\min_{x^*,\ y^*} \sum_{u,i} (r_{ui} - x_u^T y_i)^2 + \lambda \left(\sum_u ||x_u||^2 + \sum_i ||y_i||^2 \right)$$

Formula 1. Loss function of matrix factorization model

2.3 Research Goal and Contribution

To our knowledge, this project is the first trial which utilizes recommendation algorithm at estimating and specifying new candidate correlations between brain voxel and behavior.

The final goal of the project is to make an educated guess based on Neurosynth dataset and draw a candidate correlation set. Later, we are planning to organize a network map or visualized 3-dimensional graph based on the correlations we bring up with and release it on the web.

3 Research Specification

3.1 Approach

Basically, we use the dataset provided by “Neurosynth” to identify the association between behavioral keywords and brain voxel activation. After identifying association matrix, we preprocess data and apply collaborative filtering approach designed for implicit recommendation system to predict unidentified association.

3.1.1 *Neurosynth*

We use the dataset “Neurosynth” as a foundation to predict unknown candidate associations between behavior and brain activation. It provides matrix-shaped dataset that identifies association between 1335 behavioral keywords and brain voxel by analyzing 14371 neuropsychological studies. 1335 behavioral keywords were chosen by manually identifying keywords in the neuropsychological papers which not only frequently appeared in papers but also have theoretical importance. “Brain voxel” refers to the “pixel with brain volume”, which is an basic unit of the brain functional neuroimage. Amongst various ways to define voxel, Neurosynth followed “MNI152 space” which defines a voxel as 2mm x 2mm x 2mm cube. Whole brain can be divided into 90 x 109 x 90 voxels using this method. Therefore, “neurosynth matrix” can be defined as 1335 x 90 x 109 x 90 matrix which identifies brain-behavior association.

3.1.2. Association Metrics

Neurosynth matrix represents association between behavioral keywords and brain voxel activation by using multiple metrics based on one-way chi-square test (Yarkoni, 2011). Specifically, Neurosynth first identified the number of papers “A” containing specific keyword which reported specific brain voxel to be activated. Then Neurosynth identified the number of papers “B” containing specific keyword that did not report activation of that brain voxel. Then chi-square test was conducted to reject the hypothesis that “A and B are uniformly distributed”. Using the A value, B value and p-value of the chi-square test, Neurosynth created various metrics that represents association between each behavioral keyword and brain voxel activation as shown in the table below.

P(V|K)	Probability that the activation of specific voxel “V” is reported in the paper in which behavioral keyword “K” appeared.
P(K|V)	Probability that behavioral keyword “K” appeared in the paper which reported activation of specific voxel “V”
Z-score (Uniformity test)	Z-scores from a one-way chi-square test that identified if A and B are uniformly distributed
Z-score (Association test)	Z-scores from a two-way chi-square test that identified the presence of a non-zero association between term use and voxel activation.

Table 1. Association metrics calculated in Neurosynth datasets

3.1.3. Modeling for Implicit Data

Neurosynth matrix is an implicit dataset which is different from an explicit dataset in many ways. Firstly, there is no negative feedback (Hu, 2008). In other words, Neurosynth only collected positive association between keyword and voxel activation reported in neuropsychological paper (Yarkoni, 2011). Secondly, it is inherently noisy in that Neurosynth guessed the association reported in the paper would be a true positive association despite possibility that some papers reported false positive association (Yarkoni, 2011).

3.2 Challenges

3.2.1. Choosing Metric

We need to choose a metric among various association metrics provided by Neurosynth datasets. When choosing a metric in our research, two possible issues should have been considered. First, we should choose metric that would be meaningful for neuroscientists so that predicted values of “unidentified association” between keyword and brain voxel can be fruitful. Second, we should make sure that each metric is appropriate to be normalized.

3.2.2. Data Sparsity

There were 1335 keywords and 202135 brain voxels which are 90 x 110 x 90 in the Neurosynth dataset. However, like other bioinformatics dataset, it is very sparse (Zeng et al., 2017). Each keyword is associated with very limited number of brain voxel activation. (avg: 5%) Therefore, we should cluster voxels into categories but in the manner that would convince neuroscientists.

3.3 Estimated Solution

3.3.1. Identifying Binarized Association

Based on false discovery rate approach, we identified binary association between behavioral keywords and brain voxel activations by applying FDR Q<0.01 criteria (Yarkoni, 2011).

3.3.2. Voxel Cluster

To classify too many voxels into voxel clusters, we used whole-brain parcellation atlas as shown below. Each atlas classified voxels according to theoretical or connectivity approach. Theoretical approach classified voxels manually by reviewing previous anatomical and functional research. Connectivity approach classified voxels by analyzing anatomical or functional connections between brain regions. By trying every parcellation scheme, we tried to justify generalizability of our research.

Atlas Name	**Reference**	**Classifying criteria**
Shen-268	Finn et al., 2015	Connectivity approach (Resting functional connectivity-based parcellation)
Brainnetome-274	Fan et al., 2016	Connectivity approach (the anatomical connectivity-based parcellation results, including the MPM maps, probabilistic maps and both the anatomical and functional connectivity patterns)
Neuromorphometrics-167	http://www.neuromorphometrics.com/?page_id=310	Theoretical approach (Manual parcellation)

Table 2. Whole-brain parcellation atlas used for clustering voxel

4 Current Stage

4.1 Data Processing

4.1.1. Raw Data

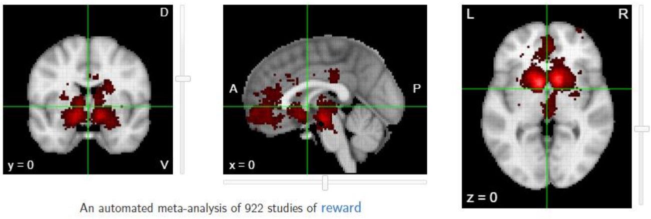

Figure 3. Brain voxels associated with "reward" in the raw data

We collected 1335 original NifTi-1 format data (nii) from 'neurosynth.org'. Each file contains 3-Dimension array which represents strength of relationship between a keyword and a brain voxel coordinate. Since our study focuses on applying recommendation algorithms to brain voxel dataset, we filtered out voxels which don't have any relationship score with any keyword. (e.i. an item which contains no rating from user.) From 836K brain voxels, only 201K are survived. Then assuming keyword as a user and brain voxel as an item, we formulated 1335 x 201235 matrix. This is our original matrix data.

4.1.2. Voxel Clustering

Our original matrix data is too sparse and imbalance among keyword. There were 1335 keywords and 202135 brain voxels. Each keyword has only relationship score to very limited number of voxels. (Average: 5%)

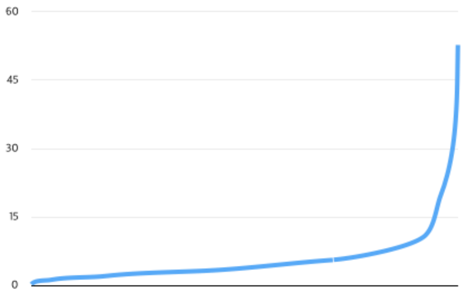

Figure 4. Voxel activation ratio per voxel

So we intended to cluster voxel which are known related from previous neural science studies. Currently, coordinate table clustering is applied. A voxel cluster contains 300 to 1600 voxels. If certain positive keyword relationship score is found more than 30 times, we set relationship score between the voxel cluster and the keyword as 1. In that manner, we formulated 1335 x 268 binary relationship score matrix. (Reduced each 201235 voxels to 268 voxel cluster.) In this manner, we solved sparse and imbalanced data problem.

4.2 BRAIN Model Design

4.2.1. Model Concept

We applied matrix factorization algorithm to our relationship score matrix. Instead of applying naive matrix factorization model, we added some factors to the loss function since we considered our dataset as an implicit dataset. Then optimize our loss function using alternating least squares algorithm, we could train the latent factor matrixes.

4.2.2. Implicit Dataset

Implicit dataset means dataset which contain only positive feedback. So 0 value in implicit dataset can be negative or positive. So during the training process, this 0 values should be considered.

Our relationship scores are based on word count on papers. Most of research in neural science focused on positive relationship between brain voxel and keyword, not negative. In this situation, it is more proper to consider our dataset as an implicit dataset.

So we add confidence level concept from [5] in order to contain 0 rating values to train our latent matrix.

$$c_{ui} = 1 + \alpha r_{ui}$$

Formula 2. Confidence level

According to above formula, 0 rating value will have 1 confidence level which is low. Non-zero rating will have relatively high confidence level. This reflects the facts that 0 rating can have both positive and negative score. Alpha means how non-zero rating value affects latent matrix. This value is dependent on dataset, so better modified from experiment. Currently we set this value as 40.

4.2.3. Loss Function

Adding confidence level to the naive matrix factorization algorithms loss function, we could get our model’s loss function.

$$\min_{x^*,\ y^*} \sum_{u,i} c_{ui}(r_{ui} - x_u^T y_i)^2 + \lambda(\sum_u ||x_u||^2 + \sum_i ||y_i||^2)$$

Formula 3. Loss function

However our loss function contain two latent factor matrix x and y. Optimizing both matrix at once is a non-convex function. So by applying alternating least squares method, optimize one of user and item latent matrix at once.

4.2.4. Alternating Least Square (ALS)

In this method, fix one of latent factor matrix and optimize other latent factor matrix from differentiation of loss function. Next, fix previously trained matrix and optimize fixed one. Repeating these steps until the loss function will converge. The differentiation of loss function for user and item latent factor matrix is like below.

$$x_u = (Y^T C^u Y + \lambda I)^{-1} Y^T C^u p(u)$$

$$y_i = (X^T C^i X + \lambda I)^{-1} X^T C^i p(i)$$

Formula 4. Differentiation of loss function

4.2.5. Current Result

We implemented our loss function and trained using ALS algorithm. Loss function converged after 4 steps. We succeeded to get recommendation result but cross validation is currently under develop. After developing validation, we are going to regulate parameters like confidence level alpha in order to get best prediction result.

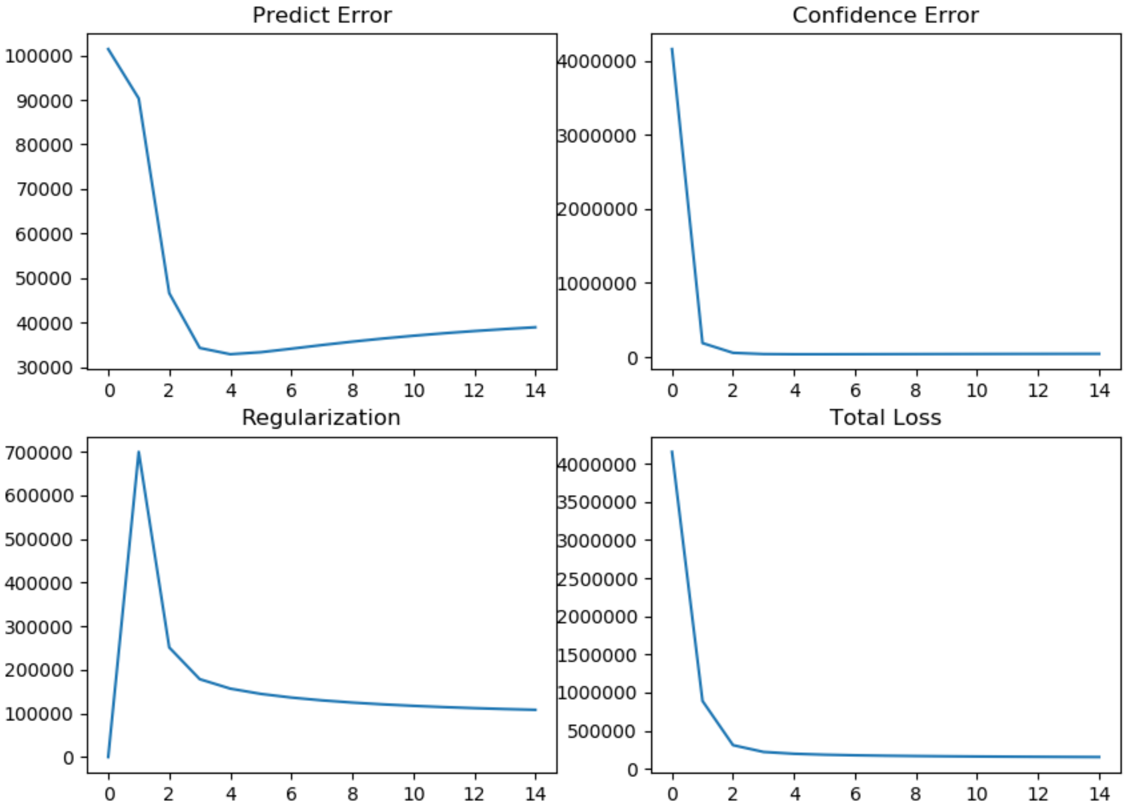

Figure 5. Train Result

5 Schedule and Roles

5.1. Schedule

Date	**Task**
3/15	Submitting project proposal, Preprocessing *Neurosynth* dataset and prior research review
3/22	Building matrix factorization model Having first meeting with the assistant Preparing for the project proposal
3/31	Modifying and optimizing model parameters
4/7	Developing cross-validation tool, Preparing for mid presentation
4/15	Testing cross-validation, Executing additional optimization process
4/30	Reflecting feedbacks
5/15	Analyzing the results, Implementing web dashboard for visualization
5/22	Preparation for showcase in May and final presentation
5/30	Having final presentation / capstone showcase

5.2. Roles

Member	**March 2019**	**April 2019**	**May 2019**
Junsol Kim	Preprocessing data	Conducting experiment	Developing web dashboard
Hyeongjoon Kim	Designing algorithm	Building experiment framework	Optimizing algorithm
Jiyun Kim	Reviewing prior research	Executing experiment, Documenting works	Building and managing database system

6. References

[1] Caspers, S., Zilles, K., Laird, A. R., & Eickhoff, S. B. (2010). ALE meta-analysis of action observation and imitation in the human brain. Neuroimage, 50(3), 1148-1167

[2] Chang, L. J., Yarkoni, T., Khaw, M. W., & Sanfey, A. G. (2012). Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. Cerebral cortex, 23(3), 739-749.

[3] Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A.R., Fox, P.T., Eickhoff, S.B., Yu, C. & Jiang, T. The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. Cerebral Cortex, 26 (8): 3508-3526,(2016).

[4] Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., ... & Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature neuroscience, 18(11), 1664.

[5] Hu, Y., Koren, Y., & Volinsky, C. (2008, December). Collaborative filtering for implicit feedback datasets. In 2008 Eighth IEEE International Conference on Data Mining (pp. 263-272). Ieee.

[6] Lan, W., Wang, J., Li, M., Liu, J., Wu, F. X., & Pan, Y. (2018). Predicting microRNA-disease associations based on improved microRNA and disease similarities. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 15(6), 1774-1782.

[7] Ortega-Martorell, S., Lisboa, P. J., Vellido, A., Simões, R. V., Pumarola, M., Julià-Sapé, M., & Arús, C. (2012). Convex non-negative matrix factorization for brain tumor delimitation from MRSI data. PLoS One, 7(10), e47824.

[8] Yarkoni, T. (2012). Psychoinformatics: New horizons at the interface of the psychological and computing sciences. Current Directions in Psychological Science, 21(6), 391-397.

[9] Zeng, X., Ding, N., Rodríguez-Patón, A., & Zou, Q. (2017). Probability-based collaborative filtering model for predicting gene–disease associations. BMC medical genomics, 10(5), 76.