Feature Importance Network (NSCI 295 Final)

Sage Hahn

April 2019

1 Introduction

There exists within computational neuroscience, among a number of other fields, a significant ideological gap between descriptive statistics and machine learning or classification based approaches. The general trade off in approaching problems of interest within a machine learning framework, i.e., complicated classifiers, is a boost to predictive performance at the cost of interpretability. On the other hand, the more traditional descriptive statistic approaches concern themselves for the most part with attempting to explain what is going on, and therefore by design yield only results under a certain threshold of complexity. Within the scope of this work I attempt to further bridge the gap between complexity and explain-ability via the construction of a feature importance network.

In order to properly introduce the the construction and analysis of the proposed feature importance network, it is important to first introduce the problem of interest. This work is concerned with the use of machine learning methods in order to distinguish between alcohol dependent and control human subjects from structural MRI acquired from the Enhancing Neuro-Imaging Genetics Through Meta-Analysis (ENIGMA) Addiction Working Group (Mackey *et al.*, 2016). A machine learning based evolutionary search in then employed on regions of interest extracted from each subject. While these techniques will be briefly explained below, the scope of the paper is primarily concerned with a network based analysis approach which receives as input results from the evolutionary search. I further attempt to show the merit of modeling feature importance within a network in comparison to a more naive approach. There are two broader questions of which modelling feature importance could potentially help with, one of simply better understanding the dynamics behind feature importance, and second in choosing a set of features for optimal classifier performance.

2 Methods

The provided dataset from the ENIGMA Addiction group contains a complicated mix of data from over 20 unique sites whose distribution of alcoholics vary drastically i.e, some sites contain only alcoholics and some only controls, only five contain both. In total there are 1652 subjects with full brain data, of which 692 have been diagnosed alcohol dependent. I make use of 150 Freesurfer derived Desikan ROI measurements from each subject (bilateral cortical thickness and surface area and sub-cortical volume) as input to the evolutionary search based classifier (Desikan *et al.*, 2006).

From a high level the evolutionary search is designed to identify and test different subsets of features (brain regions) as input to a machine learning binary classifier. While any back-end classifier can be used, within the scope of this project I made use of a cross validated logistic regression with l2 normalization implemented in the python library scikit-learn (Pedregosa *et al.*, 2011). Specifically, and due to the complex multi-site nature of the data, at each individual round of the evolutionary search a score is assigned to a given set of features as the average receiver under the characteristic operator curve (ROC AUC) score from five nested site left out evaluation schemes. The nested left out evaluation scheme works as follows: The logistic regression classifier is trained and evaluated (using 3-fold CV for regularization selection) on all available subjects, with the exception of subjects from the left out site, and with access to only the subset of features being evaluated. The trained classifier is then tested on the remaining subjects from the left out site yielding an

ROC AUC for that set of features and that site. This scheme is repeated five times across all available with sites with both alcoholics and controls available, producing an average ROC AUC.

The previous paragraph describes the behavior of the evolutionary search at the level of an individual set of features being evaluated one time. More broadly, a population of initially randomly generated subsets of 3-5 features are created and each individual then evaluated. Next, a Pareto tournament is run where two individuals are randomly compared. If one individual dominates the other then the dominated individual is removed from the population. The criteria for one individual to dominate another requires the dominating individual to have a greater score as well as less than or equal to the same number of features, therefore optimizing the global set of features for both performance and sparsity. The tournament continues to compare random individuals until only half of initial population remains, thus generating a two dimensional Pareto front of ‘optimal’ feature sets. The missing spots in the population are then filled in with a mix of new random individuals and mutated copies of existing individuals. A mutation defined as a feature within the set being randomly changed, added or removed, and therefore producing a similar but different set of features then the original. This process as described represents the initializing of a population and one complete generation. Within this work I ran 50 unique populations of 100 individuals each, all for 500 generations. One last Pareto tournament is run after the last generation yielding for each population 50 final subsets of features (2500 total) all with an assigned score.

The large number of subsets provides both the motivation and data behind my choice of modelling the problem within a network. A naive approach towards making sense of the feature sets simply says: assign each feature a weighted score based on how many subsets it appears in, the weighting relative to each feature sets score i.e., a feature that appears in a set with a high score should contribute more than a feature that shows up only in a poorly preforming feature set. Two additional constraints can then be optionally applied, namely an initial score threshold where if under a certain score a feature set will not be considered i.e., you would not want to weight a feature highly if it appeared in none of the best preforming feature sets but almost all of the poorly preforming ones. Secondly, a size constraint which divides the importance of a given feature based on the number of other features in the set i.e., each feature in a set of two should be weighted more highly then features in an equally preforming set of fifty. The most obvious flaw in this naive approach can be seen in its inability to capture two features which might ’stand in’ for the other, regardless of their potentially shared importance. It is with this problem of co-variance, among other concerns, that I introduce a network based feature importance model. Furthermore, the concepts of weighting by score, initial thresholding and size constraint all continue to apply in my network definition.

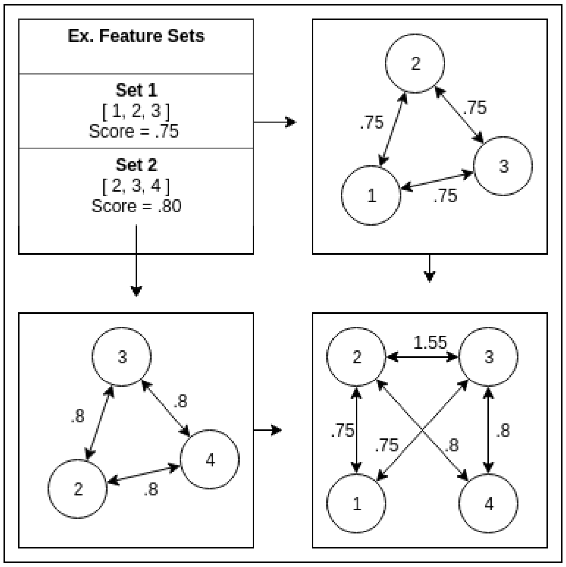

Figure 1: Simple example showing the weighted by score network construction of two arbitrary sets of features, shown separately (top right and bottom left) and merged (bottom left).

Formally, I define the undirected and weighted feature importance network to consist of a set of nodes where each unique feature corresponds to one node. An edge then exists between any two nodes if they appear together in a valid feature set (optionally as determined by passing an initial score threshold). This is a weighted network where all edges have a relative weight in relation to all other edges. In the simple case where I construct a network only weighting by score (no size constraint), an edge’s weight between any two features is defined as the sum of all valid feature set scores in which both keys appear. An example with only two sets of three features is shown in Figure 1. Within the size constraint variant an edges weights are instead calculated from the feature sets score divided by the number of features in that set. Lastly, the edge weights in the network can optionally be normalized by dividing each weight by the sum (or sum divided by number of features) of all valid feature sets scores i.e., an edge between some feature i and j with a weight of 1 would correspond to the case where features i and j appeared together in every valid feature set.

I further define a projection onto the resulting network, similar to a bipartite projection, but primarily concerned with the previously introduced problem of identifying features that ‘stand in’ for other features. Within the projected network a weighted edge is defined linking every node with the neighbors of its neighbors. Formally, for any three nodes X, Y and Z in the original network where edges (X,Y) and (Y,Z) exist I define a weighted edge in the projected network, (X,Z), as

$$w(X,Z)_{proj} = \frac{w(X,Y) + w(Y,Z)}{2} - w(X,Z)$$

The function w() simply referring to the weight of that edge in the original network. Figure 2 shows two simple applied examples, where in both it can be seen that in the resulting projection the relationship (edge weight) between node 1 and 3 is highlighted. The second example further shows a case where the results edge weights between (1,2) and (2,3) are negative after the projection, which under the interpretation of the projected edge weights as signifying an analog of co-variance seems reasonable. This definition notably fails to deal with the case where there are two or more reasonable paths, for example a 4-cycle. In these cases there exists two reasonable possibilities, namely taking an average for each valid projected weight or calculating a sum. The average case though suffers from the possibility that a strong connection could be lost due to the presence of a number of other weaker connections, whereas in the summing case any strong two nodes will contribute. Likewise, while the summing case will seemingly inflate the edge weight between two nodes with numerous valid paths, I would argue the presence of these paths, even if individually weak, are meaningful. Therefore I define the final projected edge weights to represent the sum of all valid projections as defined above.

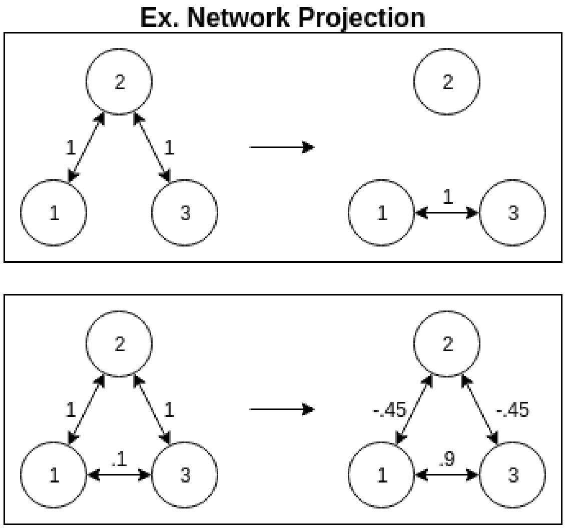

Figure 2: Two basic examples of projected network weights, where the original weights are shown on the left and the right side version shows the projected version.

An important and fairly simple trait of interest of our weighted network is the concept of weighted degree. Weighted degree is defined for all nodes as the sum of the weights of all edges containing that node. If one were to then order nodes by weighted degree, a parallel between our initial naive ranking scheme and this introduced ordering arises. Specifically, the ranking produced by the weighted and size constraint variants of the naive ordering and weighted degree ordering will be exactly the same, whereas without the size constraint they will be only similar. As a simple metric of measuring variability between two different rank spaces I can use the Spearman’s rank-order correlation which for the case of distinct integer ranking between lists of ranks X and Y is calculated as,

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

Where, $\forall i \in X \cap Y, d_i = X_i - Y_i$ and n is the number of observations (Zar, 1972). While proposing some sort of ordering might be useful towards the goal of selecting an optimal final feature set, any proposed ordering will be less useful towards understanding the complexities of the feature space.

One perhaps more fruitful method of extracting an optimal feature set for use in a classifier relates back to my definition in constructing the feature network itself. That is, features are essentially added as cliques (fully connected sub-graphs) and therefore it might be useful to identify the largest clique within the feature network. Maximum cliques can further be

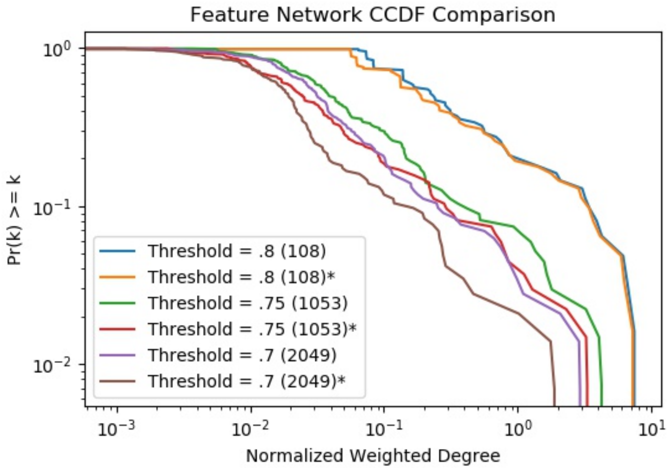

Figure 3: Complementary cumulative distribution function (CCDF) comparison between different network constructions with different initial thresholds and optional size constraint. Where * signifies that the size constraint was applied and the number in parentheses represents the number of valid keys after the threshold.

found under different edge weight thresholds, where by removing edges the resulting maximum clique will be smaller and smaller. The assumption is that these features present within the various sized maximum cliques might represent high preforming subsets of features. Another potentially useful tool in understanding the dynamics of different feature importance graphs is to introduce a measure of clustering. I will make use of an average clustering coefficient across the whole graph, where I use the geometric average of the sub graph edge weights as my clustering coefficient defined as,

$$c_u = \frac{1}{deg(u)deg(u)-1} \sum_{vw} (w(u,v)w(u,w)w(v,w))^{1/3}$$

(Onnela *et al.*, 2005).

3 Results

I introduced a number of different parameters for constructing different feature importance networks. In Figure 3, I explore how different choices of initial threshold and choice of optionally enforcing a size constraint effect the resulting degree distribution. When a higher threshold is enforced (.8 in the example) far less sets of features are considered and perhaps as a consequence the distribution remains more stable regardless of size constraint. When more initial sets are considered, as in the case of .7, then the size constraint has a more noticeable effect. The general trend of the CCDF remains, i.e., you would expect similar power law exponents, but when applying the size constraint the resulting distribution is shifted to the left and the exact shape certainly differ.

Threshold	# Key Sets	Clustering Coef.
.50	2437	.643
.60	2375	.689
.70	2049	.738
.75	1053	.775
.80	108	.816

Table 1: Average clustering coef. across different initial thresholds.

Another way to quantify the different effects of the initial network construction parameters is to compare node rankings. I can easily rank all of the nodes/features by weighted degree as introduced earlier. As I do not have a ground truth top ranking of features to consider, I will instead have to make comparisons with the Spearman's rank-order correlation arbitrarily between different configurations (note: to compare ranks with a different number of features we can just consider the a truncated version of the longer list). For example, a comparison between a threshold of .7 and .8 yields a Spearman's coef. of .28 without size constraint and .09 with. A comparison just changing size constraint with a threshold of .8 yields a coef. of .587, and with a threshold of .7 a coef. of .27. I can also consider a larger comparison, say between a threshold of .6 and .8, where with a size constraint the coef. is .17 and without .01. Lastly, note that if we instead constrain our rank comparison to the top 10 features and repeat the .6 versus .8 analysis we find a coef. of -.32 with a size constraint and -.42 without. These comparisons in general seem to suggest a high volatility in node ranking that is quite dependant on network construction parameters. Likewise, another potentially misleading confound is that when the initial threshold is set lower, say .6, when compared to .8, there are more features available and therefore features that are ranked which might lead to less correlation, despite list truncation.

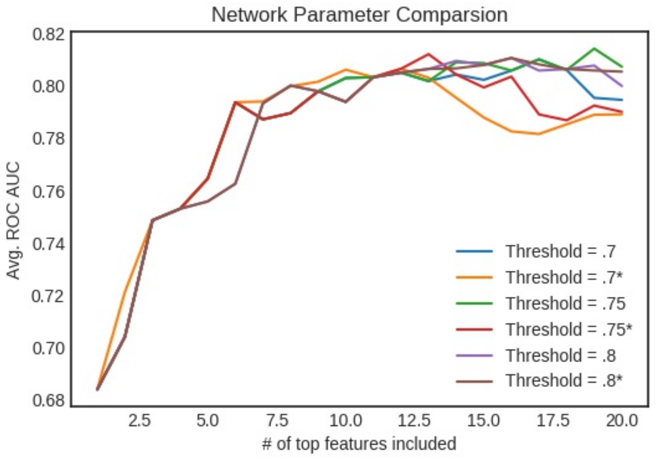

Figure 4: Model evaluation using the top features as extracted from the weighted degree using explicitly the top 1-20 features. * signifies that a size constraint was applied.

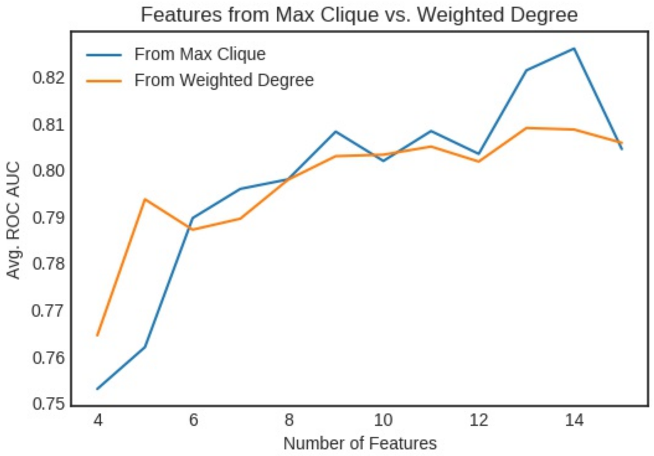

Figure 5: Model evaluation using the top features as extracted from the weighted degree versus features extracted from thresholded maximum cliques, all for an initial threshold of .75 with no size constraint.

I can also examine how the average clustering changes with different initial thresholds (the size constraint will not effect clustering) in Table 1. In general one might expect higher clustering to be a sign that the features in the network translate to more directly useful features. This can be seen in a comparison between a threshold of .8 and .5, where when all key sets are included the resulting network is less clustered. Though, with an initial threshold of .8, only 108 sets of features are preserved which certainly makes obtaining a higher clustering coef. more likely. It unclear if this measurement is at all meaningful in this context.

Based on the above exploration of different feature network parameters it is still difficult to point towards a 'best' representation for understanding feature importance dynamics. One more quantitative way, though still potentially naive, involves testing the performance of the algorithm on the top 1-20 features as extracted by weighted degree as seen in Figure 4. This method of validation importantly only confirms the utility of the parameters in selecting subsets of high ranking features and may very well not represent the best network settings for understanding feature dynamics. While no clear 'best' network parameters emerge just by looking at the graph a few thing stand out, that a threshold of .75 with size constraint yields the second best score with 14 features and that a threshold of .75 without size constraint yields the best score with 19 features. A threshold of .75 also has the highest area under the curve. Therefore, for the remainder of the analysis we will consider a feature comparison network constructed with an initial threshold of .75 with no size constraint.

Table 2 displays the top 10 features as determined by weighted degree as well as the top 10 edges as identified by weight and the top edges as identified by the previously defined network projection. notably, each pair of features in the projection seems to follow the trend of having one feature previously identified as a top feature and one previously lower ranked feature. This trend makes sense as the way I defined the network projection seems to overweight already strong edges e.g., the average between 1 and .1 is still .55.

When running a maximum clique algorithm on this feature network a clique of size 15 is initially found with no threshold. By progressively introducing a threshold maximum cliques of every smaller size can then be found. I then evaluated performance using the subset of features found from these progressively smaller cliques. Interestingly, the features found from a a max clique of size 14 produces an average ROC AUC of .826 versus the highest ROC AUC found from the previous degree ranking, .814. A full comparison is shown in Figure 5.

4 Discussion

Given the original two goals of modelling feature importance as a network, namely improved

Features by degree	Edges by weight		Projected Edges by weight	
L_superiorfrontal_thickavg	L_superiorfrontal_thickavg	R_lateralorbitofrontal_thickavg	L_medialorbitofrontal_thickavg	L_superiorfrontal_thickavg
R_lateralorbitofrontal_thickavg	L_superiorfrontal_thickavg	R_transversetemporal_surfavg	L_transversetemporal_surfavg	R_transversetemporal_surfavg
R_transversetemporal_surfavg	R_lateralorbitofrontal_thickavg	R_transversetemporal_surfavg	L_caudalanteriorcingulate_thickavg	L_superiorfrontal_thickavg
L_put	L_superiorfrontal$_t$*hickavg*	L_put	R_caud	L_superiorfrontal_thickavg
L_precuneus_surfavg	L_put	R_transversetemporal_surfavg	L_supramarginal_surfavg	L_superiorfrontal_thickavg
L_rostralmiddlefrontal_thickavg	L_put	R_lateralorbitofrontal_thickavg	L_superiorfrontal$_t$*hickavg*	R_parsopercularis_surfavg
L_parsopercularis_thickavg	L_precuneus_surfavg	L_superiorfrontal_thickavg	L_medialorbitofrontal_surfavg	L_superiorfrontal_thickavg
L_lateraloccipital_surfavg	L_superiorfrontal_thickavg	L_rostralmiddlefrontal_thickavg	R_inferiorparietal_surfavg	L_superiorfrontal_thickavg
ICV	L_precuneus_surfavg	R_transversetemporal_surfavg	R_pericalcarine_thickavg	L_superiorfrontal_thickavg
R_cuneus_thickavg	R_lateralorbitofrontal_thickavg	L_rostralmiddlefrontal_thickavg	L_bankssts_surfavg	L_superiorfrontal_thickavg

Table 2: Top ranked features and edges under an initial threshold of .75 and no size constraint

understanding of dynamics and improved selection of optimal features, this work provides support for both. In particular, I was able to exploit the network structure towards choosing more optimal features then those found via the weighted degree ranking via identifying progressively smaller maximum cliques. While it is very well possible that there exist better methods i.e., perhaps making use of measures of centrality or clustering might help to identify a more optimal set, it can not be inferred as a result from this work.

Towards the initial stated goal of improving understanding, I believe my approach was mostly successful. The idea around the network projection as mention I believe has the potential to be interesting, but might require tweak to how I calculate edge weight exactly. Future work, or rather work that does not lend itself to figures within a paper, would include generating interactive visualizations of the network and the network projection. This would allow for a more intuitive exploration of feature importance and especially towards interpreting the network projection.

References

Desikan, Rahul S, Ségonne, Florent, Fischl, Bruce, Quinn, Brian T, Dickerson, Bradford C, Blacker, Deborah, Buckner, Randy L, Dale, Anders M, Maguire, R Paul, Hyman, Bradley T, *et al.* 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, **31**(3), 968–980.

Mackey, Scott, Kan, Kees-Jan, Chaarani, Bader, Alia-Klein, Nelly, Batalla, Albert, Brooks, Samantha, Cousijn, Janna, Dagher, Alain, De Ruiter, Michiel, Desrivieres, Sylvane, *et al.* 2016. Genetic imaging consortium for addiction medicine: From neuroimaging to genes. *Pages 203–223 of: Progress in brain research*, vol. 224. Elsevier.

Onnela, Jukka-Pekka, Saramäki, Jari, Kertész, János, & Kaski, Kimmo. 2005. Intensity and coherence of motifs in weighted complex networks. *Physical review e*, **71**(6), 065103.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, **12**, 2825–2830.

Zar, Jerrold H. 1972. Significance testing of the spearman rank correlation coefficient. *Journal of the american statistical association*, **67**(339), 578–580.