Pattern Recognition Project Report
Pace University, December 04, 2020

Friend Recommendation System using Ensemble Voting and Distance Heuristics

Viswanathan Appakkudal Ramani, Dhariya Harsh Parikh, and Yashwanth Varre
Instructor: Professor Sung-Hyuk Cha
Pace University

Abstract

Friend Recommendation System helps to forecast future friends by analyzing social network using an ensemble of classifiers and weighted voting strategy. Social network analysis is a mechanism through which graphs are analyzed to find patterns and deduce relationships between multiple nodes. Three-person P1, P2, P3 form a length-3 following chain if a person P1 follows person P2 and person P2 follows P3. Given a length-3 following chain, it may make sense to recommend to P1 to follow person P3. It may make even more sense doing that if there are several 3-following chains between P1 and P3.

We considered the Facebook dataset provided by Stanford which contains 4039 nodes and 88234 edges and makes a directed network. We cleaned the graph by eliminating isolated nodes, and nodes with fake edges. We extracted features such as Jaccard Similarity, Adamic-Adar Index, Preferential attachment, Resource Allocation Index, and Common Neighbors [1]. Common neighbors find the number of mutual friends between two persons. Jaccard similarity compares members for two sets to see which members are shared and which are distinct. Resource Allocation Index is a measure that calculates the amount of resource that a node can send to the destination node via its neighbors. Preferential attachment measures the influence of a node over the other. Adamic-Adar index measures the number of shared links between two nodes. We modeled the features using an ensemble of ANN, Logistic regression, and Naïve Bayes classifiers using a weighted voting mechanism [2]. Classifiers are weighted based on their accuracy score; the classifier with maximum accuracy gets the highest weights, and weights reduce for a classifier as the accuracy reduces. And, if the accuracy is less than a threshold, or if the classifier overfits/underfits the data, the classifier is eliminated, thereby keeping only significant classifiers.

1. Introduction:

In today's world, social networking tools such as Facebook, Twitter, and Whats-app are quite prominently used among-st the people. Everyone is in search of making new friends, and these tools help people from across the world to connect and learn new things. As these tools are predominantly used by people all over the world, there are lots of connections between one and the other. This gives rise to big data analytics on Social Networks. This is report is based on the project on Social Network analysis which helps to forecast people who could most-likely be friends tomorrow.

2. Input and Output:

The project uses the open-source dataset provided by Stanford University which is called 'Facebook-dataset'. The data contains 4039 nodes and 88234 edges, meaning the dataset contains 4039 users and 88234 connections distributed amongst 4039 users. In graph theory nodes are the entities and edges are the relationship that defines the entity. The project considers a directed graph, which means, two nodes are to each other. If A is undirected to B, means A is connected to B, and B is connected to A. The input dataset contains 2 input columns and 1 result column. Input columns correspond to Source and Recipient nodes, whereas the output column is Boolean. The output states true, if a connection exists between source and recipient, and false otherwise.

3. Terminologies:

- **Graphs:** Graphs are pairwise relations between objects. Two or more nodes connected by an edge(s) forms a graph
- **Isolated nodes:** Nodes which do not have any relationship to another node(s) is/are called an isolated node
- **Fake edges:** Edges that are connected to zero or one node.
- **Ensemble Voting:** A voting ensemble involves averaging the predictions made by multiple classifications or regression models.

4. Features:

As part of the Machine Learning Lifecycle, the first data cleaning was performed. Isolated nodes and fake edges were removed from the graph. The dataset contains only two input columns, the source, and the recipient nodes. This is not a sufficient feature to make a model. The project constructs features based on four metrics, they are common neighbors, Jaccard similarity measure, Adamic-Adar index, Preferential Attachment, and Resource Allocation Index.

- **Common Neighbors:** Common neighbors find the number of common connections of friends of A, B [3].

  - $CN(A,B) = | \Gamma(A) \cap \Gamma(B) |$
    - $\Gamma(A)$ = neighbors of A
    - $\Gamma(B)$ = neighbors of B

- **Jaccard Similarity Measure:** The Jaccard coefficient measures similarity between two sets. It is defined as the CN(A,B) divided by the size of the union of the sets [4]. This gives information about how likely a suggestion can be made between two nodes based on the Jaccard measure value. As the score is closer to 1 higher the possibility a suggestion can be made.

  - The range of Jaccard Similarity Measure is [0,1].
  - $J(A,B) = CN(A,B) / | \Gamma(A) \cap \Gamma(B) |$ [3]

    (or)
  - $J(A,B) = | \Gamma(A) \cap \Gamma(B) | / | \Gamma(A) \cap \Gamma(B) |$
    - $\Gamma(A)$ = neighbors of A
    - $\Gamma(B)$ = neighbors of B

- **Adamic-Adar Index:** It's a measure of the number of shared links between two nodes. Adamic, Adar observed that if the number of links is greater the resource sharing is higher, which in turn affects making new friends. [4]. To calculate AAI at least one shared link must be present between both the nodes otherwise the score would be zero.

  - $AA(A,B) = \psi (x \in | CN(A,B) |) \sum (1 / \log| CN(x) |)$

- **Preferential Attachment:** It is based on the principle "new vertices attach preferentially to sites that are already well connected" [5]. The underlying ideology is that, if a node has more neighbors, a new node that intends to join the node, will tend to connect with the node which has more neighbors.

  - Mathematically represented as a product of the degree of two nodes(number of common neighbors each node contains).
  - $PA(A,B) = |CN(A)| * |CN(B)|$
    - CN(A) = common neighbors of node A.
    - CN(B) = common neighbors of node B.

- **Resource Allocation Index:** It is a measure that calculates the amount of resource that a node can send to the destination node via its neighbors

- $RA(A,B) = \psi (x \in |CN(A,B)|) \sum (1 / |CN(x)|)$

5. Algorithms:

Features are extracted from the Stanford dataset as mentioned in Section 4. The extracted features are modeled using an ensemble of classifiers. The project involves a supervised classification mechanism, and, the given dataset is divided into a train-test set with a ratio of 70:30. The new features are modeled using k-nearest neighbors, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, Artificial Neural Network, Naïve Bayes algorithm, and the accuracy for each classifier is jotted down. Then, the project involves selecting and weighing the classifiers based on their accuracy scores. The algorithm for weighted voting is as follows:

- Find the classifier with maximum accuracy and provide maximum weight for this classifier
- Eliminate classifiers that overfit/underfit
- Vote other classifiers relatively based on their accuracy
- Eliminate classifiers whose accuracy is less than the threshold

The accuracy scores of 7 seven considered classifiers are as follows:

	ANN	**Naïve Bayes**	**SVM**	**Logistic**	**Decision Tree**	**Random Forest**	**KNN**
Test Accuracy	67.3	69.5	32.6	69.9	46.1	44.2	51.5
Train Accuracy	66.4	65.6	65.8	65.9	79.7	79.7	72.2

Table1: Train Test Accuracy scores of 7 classifiers

From Table1, we can notice that SVM, KNN, Decision Tree, and Random Forest classifiers overfit the data, hence the classifiers are eliminated from consideration. As a result, only Naïve Bayes, Logistic, and ANN classifiers are considered for the project. We then provide maximum weight to the Logistic Regression classifier as it has the maximum test accuracy of 69.9%. We set a threshold accuracy of 60%, below which classifiers will be eliminated. We assume an accuracy variance of 2%, and for every decrement of 2%, we reduce the weight for the classifier by 1 point. Hence, the weight for each classifier looks like the one mentioned in Table2.

	ANN	**Naïve Bayes**	**SVM**	**Logistic**	**Decision Tree**	**Random Forest**	**KNN**
Votes	4	5	0	5	0	0	0

Now, the results are modeled based on the considered classifiers. The actual result can be deduced as:

$$\text{Actual Result} = \sum(\text{weight}(i) * \text{result of the classifier}) / \text{total\_weight} \quad (1)$$

i = considered classifier
weight(i) = weight of the considered classifier
total_weight = sum of weights of each classifier

From table2, and (1) we can deduce that:
ActualResult = 4*(result_of_ann) + 5*(result_of_logistic) + 5*(result_of_naive_bayes) / (4+5+5)

Modeling the databased on the above-mentioned strategy, we get a training accuracy of 66% and test accuracy of 70%.

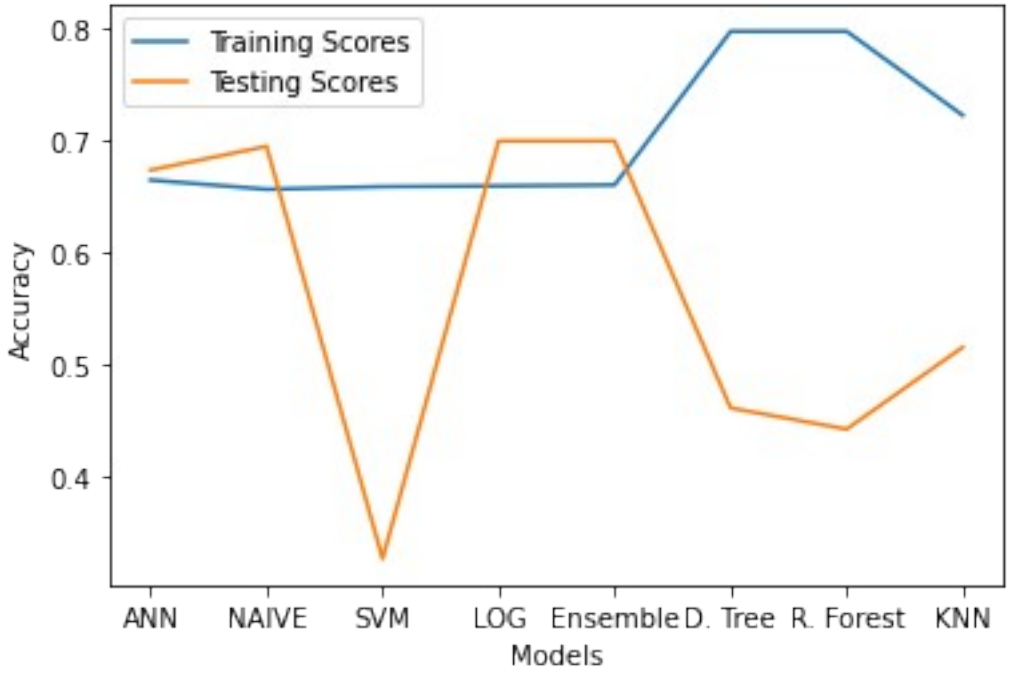

Figure1: Train-Test accuracy scores of various classifiers and Ensemble classifier

From Figure1, we can see that, there is a marginal improvement in training accuracy, although the test accuracy remains the same as that of a logistic classifier.

6. Research Extensions:

To improve the accuracy of the model, the following scopes could be considered:

- Extract more features, and if there are more than 20 features, perform a Principal Component Analysis to reduce the dimension of the data set to the required number of features
- Regularize the overfitting classifiers and add them to the ensemble voting strategy
- Use Deep Learning mechanisms, such as Drop out Neural Network, Neural Network with sigmoid (for hidden layers), and Softmax (for output layer) activations.
- Use the Preferential voting mechanism to perform ensemble learning

- Analyze the data and generate novel features

7. References:

(1) Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L., Elovici, Y., Link Prediction in Social Networks using Computationally Efficient Topological Features. In: Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom), pp. 73–80. IEEE (2011).

(2) S. Chebrolu, A. Abraham and J.P. Thomas, "Feature deduction and ensemble design of intrusion detection system", International journal of computers and security, vol. 24, no. 4, pp. 295-307, 2005

(3) Lin Yaoa , Luning Wanga, Lv Pana , Kai Yaob, "Link Prediction Based on Common-Neighbors for Dynamic Social Network", The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016)

(4) Leydesdorff, 2008, L. Leydesdorff, "On the normalization and visualization of author co□citation data: salton's Cosine versus the Jaccard index". J. Am. Soc. Inf. Sci. Technol., 59 (1) (2008), pp. 77-85

(5) Albert-Laszlo Barabasi and Reka Albert, 'Emergence of scaling in random networks', science, 286(5439), 509–512, (1999)

(6) Tao Zhou, Linyuan Lu, and Yi-Cheng Zhang, 'Predicting missing links via local information', The European Physical Journal B, 71(4), 623– 630, (2009).