Friend Recommendation System using Ensemble Voting

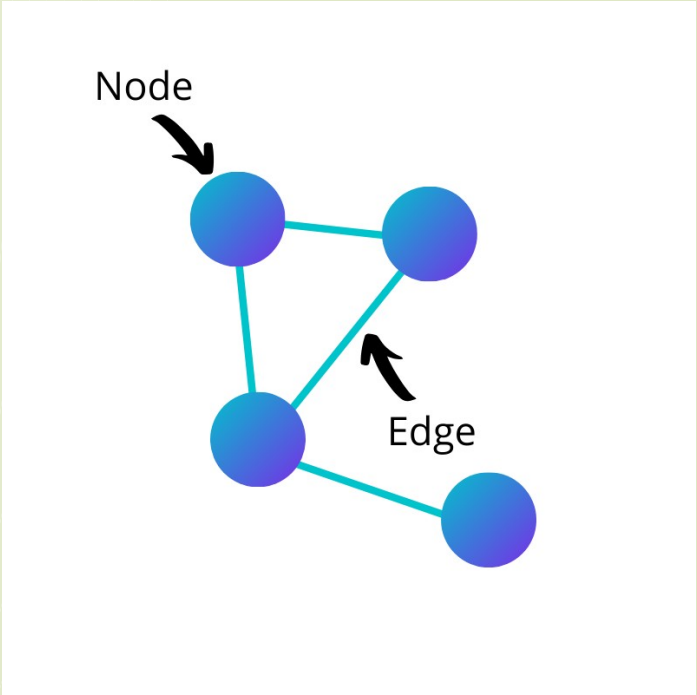
Dhariya Parikh

Yashwanth Varre

Viswanathan Appakkudal Ramani

Objective

- Given two users who are not friends on social media, predict whether or not the two users would become friends in the future.
- Three persons 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.



Source: https://towardsdatascience.com/how-to-get-started-with-social-network-analysis-6d527685d374

Data

- Data Resource: https://snap.stanford.edu/data/
- Dataset: Facebook dataset
- About the data:
  - The Facebook dataset contains 4039 users and 88,234 friend connections distributed among the 4039 users
  - Number of triangles is 1612010
  - File size: 1MB

Steps

- Load the datasheet.
- Extract features from the given data. Jaccardian, Adam index, Common Neighbours, Resource Allocation, and Preferential Attachments are the features extracted from the data.
- Eliminate nodes which don’t have any connection
- Split train/test set
- Perform Machine Learning to get accuracy score
- Tune the model to enhance the accuracy

Features

- Lests assume A,B to be two members, and their friends are represented as sets.
  - A= {1,2,3,4,5} B = {4, 5, 6, 7}
- ***Jaccard similarity index*** compares members for two sets to see which members are shared and which are distinct.
  - J(A, B) = UNION(A,B) = UNION(A,B)/ INTERSECTION(A,B)
  - J(A,B) = 2/7
- ***Common Neighbors***: Number of mutual friends between two individuals
  - CN(A,B) = UNION(A,B)
  - CN(A,B) = 2

Features

- ***Resource Allocation Index*** is measure that calculates amount of resource that a node can send to destination node via its neighbors
  - $RA(A,B) = \psi (x \in CN(A,B)) \sum (1 / CN(x))$
- ***Preferential Attachment*:** More number of mutual friends, more influential the person is
  - $PA(A,B) = CN(A) * CN(B)$
  - $PA(A,B) = 5*4 = 20$
- ***Adamic Adar Index***: Higher the value, if amount of shared links between two nodes are higher
  - $RA(A,B) = \psi (x \in CN(A,B)) \sum (1 / \log|CN(x)|)$

Algorithm

- Ensemble of :
  - ANN
  - Logistic
  - Naïve Bayes
  - SVM, Decision Tree, Random Forest (Considered but not used)

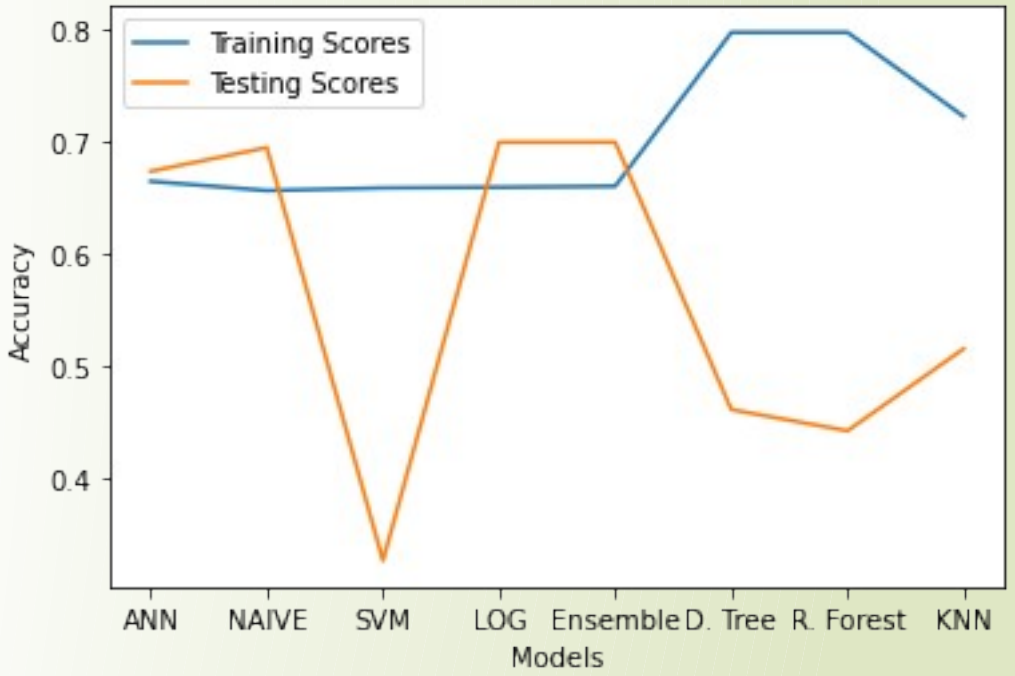
- Weighted Voting Strategy:
  - Find the classifier with maximum accuracy and provide maximum weight for this classifier
  - Eliminate classifiers that overfit/undefit
  - Vote other classifiers relatively based on their accuracy
  - Eliminate classifiers whose accuracy is less than the threshold

Weighted Voting – Approach Illustration

	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
Votes	4	5	0	5	0	0	0

- SVM, KNN, Decision Tree, Random Forest – Overfitting. Hence, Eliminated
- Logistic has the highest Accuracy (69.9%).
- Assumed Variance = 2% . Relative Voting,
  - 68% - 70% = 5 points
  - 66% - 68% = 4 points
  - 64% - 66% = 3 points
  - 62% - 64% = 2 points
  - 60% - 62% = 1 point
  - <60% = Elimination

Results



Ensemble of Weighted Voting Model:

- Training Accuracy: 66%
- Testing Accuracy: 69.9%

References

- J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.
- Ali Choumane, Zein Al Abidin Ibrahim. Friend Recommendation based on Hashtags Analysis
- Qi Yang, Aleksandr Farseev, Andey Filchenkov I Know Where You Are Coming From: On the Impact of Social Media Sources on AI Model Performance
- Marzieh Pourhojjati-Sabet, Azam Rabiee. A Soft Recommender System for Social Networks
- J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, “Make new friends, but keep the old: recommending people on social networking sites,” in Proceedings of the 27th international conference on Human factors in computing systems, ser. CHI ’09. New York, NY, USA: ACM, 2009, pp. 201–210. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518735

References

- 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)
- Noble, W. What is a support vector machine?. *Nat Biotechnol* **24,** 1565–1567 (2006). https://doi.org/10.1038/nbt1206-1565
- Garson, D. (2008). *Logistic regression.* Retrieved December 3, 2008, from http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm.