Analysis of Probabilistic Models for Influence Ranking in Social Networks

Pranav Nerurkar, Aruna Pavate, Mansi Shah and Samuel Jacob

Abstract Influence is a phenomenon occurring in every social network. Network science literature on Influence ranking focuses on investigation and design of computational models for ranking of nodes by their influence and mapping the spread of their influence in the network. In addition to this contemporary literature seeks efficient and scalable influence ranking techniques that could be suitable for application on massive social networks. For this purpose joint and conditional probabilistic models could be a way forward as these models can be trained on data rapidly making them ideal for deployment on massive social networks. However identification of suitable predictors that may have a correlation with influence plays a major role in deciding the successful outcome for these models. The present investigation proceeds with the intuition that interaction is positively correlated with influence. Furthermore, through extensive experimentation it identifies a joint probabilistic model and trains it on interaction characteristics on nodes of a social network for influence ranking. A qualitative analysis of these models is presented to highlight its suitability.

Keywords Social network analysis · Social influence analysis
Network centrality · Influence Ranking

P. Nerurkar (✉)
Department of CE & IT, VJTI, Mumbai, India
e-mail: pranavn91@gmail.com

A. Pavate
Department of CE & IT, Atharva CoE, Mumbai, India
e-mail: gavkare@gmail.com

M. Shah
Department of CE & IT, Rizvi CoE, Mumbai, India
e-mail: mansishah928@gmail.com

S. Jacob
Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan, India
e-mail: samueljacob@atharvacoe.ac.in

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_23

1 Introduction

Social influence is referred to as the behavioural change or alternation in performance of actions of an individual brought about due to interchanges conducted with other individuals. It has known to be the causal link for other well documented phenomena seen in social networks such as social competition, peer pressure, homophily, information spread, network evolution. Influence is also important in constraining the flow of dynamics within a network [1].

In online social networks, social influence may depend on factors such as strength of relationships between nodes, distance between the nodes, number of paths for traversal from a node to its neighbors and characteristics of the individuals in the network [2]. However, for the purpose of developing a computational model that shall measure influence quantitatively and qualitatively, a statistic based measure is required. Statistic based measures proposed in the literature for measuring influence are centrality, between-ness, closeness, decay etc. These methods focused only on the structural characteristics of a social network to calculate the influence of nodes in it. Influence Maximization technique goes beyond such simple statistical measures and provides an alternate method to measure influence. However it is a subset selection problem and hence is fundamentally different from Influence ranking which is a measurement problem. Hence it is not the focus of the present inquiry.

Joint and conditional probabilistic frameworks could be the suitable methods to measure influence. This is because such techniques have been experimentally verified on wide range of scenarios. The advantage of these methods is that they can be trained rapidly on data. This makes them suitable for deployment on massive social networks such as Twitter, Facebook etc. However, a key aspect that determines the success of these methods is the selection of suitable predictors for training them. This inquiry selects statistics about user interactions for training four probabilistic models. Extensive experiments are then performed using these frameworks on Twitter data. Standard evaluation metrics are used to select the most optimal model out of these. The aim of this work is to extend the literature on use of probabilistic models to rank nodes on influence.

2 Review of Literature

The edge and node measures for calculating centrality are techniques that rely on structural features which ignoring attribute level data of a node or interaction characteristics of a node in the network. To overcome these drawbacks Influence Maximization technique and Probabilistic Generative models were proposed.

2.1 Quantifying Influence

A formal computational definition for influence is proposed in [3]. The nodes are x_i and x_j and t and $t-1$ are the time instants, a_{ij}^t denotes value of the adjacency matrix at time instant t for nodes i and j.

$$\frac{p((x_i^t, x_j^t) > (x_i^{t-1}, x_j^{t-1}) | a_{ij}^{t-1} = 0, a_{ij}^t = 1)}{p((x_i^t, x_j^t) > (x_i^{t-1}, x_j^{t-1}) | a_{ij}^{t-1} = 0)} \tag{1}$$

The numerator is the conditional probability that two nodes that were not linked at instant $t-1$ are linked at t have seen increase in their similarity. The denominator calculates the conditional probability that two nodes that were not linked at $t-1$ see an increase in their similarity at t as compared to $t-1$. This method however does not differentiate the influence from various angles [2].

2.2 Influence Maximization (IM)

The formal definition of the problem is: Given a social influence graph $G(V, E)$ with V, E representing vertices (individuals) and edges (social relationships) respectively. $P(u, v)$ is the probability that v is activated by an already active node u sharing a directed edge (u, v). Independent cascade model allows a small sets of seed nodes to be activated. Then a node u can activate its neighbor v with probability $p(u, v)$. Influence spread is denoted as the maximum number of nodes activated. The problem is to maximize the influence spread [4]. IM is an NP-hard problem and hence a greedy approximation algorithm is used that can theoretically guarantee influence spread is within 63% of the optimal influence spread. However, the greedy algorithm requires the evaluation of the influence spread given a seed set. This step is time consuming.

In contrast to this line of work, statistics of interaction by nodes were also explored for measuring their influence in the network. Interaction occurs in the network through activity performed by the node, response received from its neighbors on such an activity, propagation of activity further throughout the network, activating reaction from nodes not directly connected to each other etc. Klout rank is one such measure that ranks nodes in a network as per their influence based on their statistics of interaction [5]. These statistics are collected from multiple social networks of which the node is a participant. Klout rank uses a feature vector of 3600 attributes which are statistics of user interaction on the social platform and analyzes interactions between the user and other participants on the social network to generate a score. Alexy et al. have investigated use of machine learning for calculating influence ranks based on data generated from social network activity by the users [6]. Behnam et al. modelled influence of a node in a network based on metrics such as number of followers and ratio of affection [6]. The purpose of the current inquiry is to advance the existing knowledge in the field of influence analysis by providing experimental evaluation of

conditional and joint probabilistic learning techniques to model influence of nodes on social networks.

3 Mathematical Model

Learning techniques build a function $g \in G$, where G represents the hypothesis space. Function $g : X \rightarrow Y$ maps the input space X to the output space Y. If F denotes the space of scoring functions f such that $f : X * Y \rightarrow R$ then g is as shown in Eq. 2.

$$g(x) = \arg\max_{y} f(x, y) \tag{2}$$

Then a suitable conditional probability model or joint probability model is chosen for mapping f to g. For a conditional probability model, g is calculated as $P(y|x)$ and for a joint probability model g is $P(x, y)$. A method to choose the appropriate model is structural risk minimization in which a regularization penalty called L_2 norm which is $\sum_j \beta_j^2$ is incorporated in to the cost function $J(\theta)$ to minimize over-fitting.

3.1 Learning Technique for Ranking Influential Nodes

Conditional probability model or joint probability models are then used in the below technique to identify the most suitable model amongst them for comparing influences and ranking nodes as per their influence.

Algorithm 1: Influence Ranking model

```
Set initial seed for random numbers;
Set the training control values;
Set the tuning grid for parameter search;
for each parameter set do
    for each re-sampling iteration set do
        hold out specific samples;
        Pre process the data (Center and Scale);
        Fit the model on the remaining samples;
        Predict the held out samples;
    Calculate the average performance across held out predictions;
Determine the optimal parameter set;
Fit the final model to all the training data using optimal parameter set ;
```

Table 1 Description of the data-set

Sr. No.	Name	Description
1	Training set size	5500
2	Test set size	5952
3	Feature vector	8
4	Classification	Binary

Table 2 Attributes in the feature vector

Sr. No	Feature list	Sr. No	Feature list
1	Follower count	5	Following count
2	Listed count	6	Mentions received
3	Retweets received	7	Mentions sent
4	Retweets sent	8	Posts

4 Experimental Study

4.1 Data-Set

Data-set used for the experiment consists of features extracted about user interaction characteristics (Tables 1 and 2) from Twitter—an online social network [6]. Twitter is a graph $G = (V, E)$ with n nodes having k attributes which have been formed using its interaction characteristics. These have been collected from the nodes activity observed over an online social network. Any two nodes, A, B of G are picked and a feature vector x_i is built by combining their individual interaction characteristics i.e. $x_i = a_i, b_i$. This is used to build a dataset X that contains training samples of the form $(x_1, y_1)...(x_n, y_n)$ such that x_i is the feature vector of the ith example. The corresponding class label of x_i is y_i which represents the human judgment about which one of the two individuals in x_i is more influential. Thus $y_i \in 0, 1$ such that $y_i = 0$ means first user is more influential and $y_i = 1$ means that the second user is more influential.

Learning models described in Sect. 4.2 are used in the Influence learning strategy proposed in Sect. 3.1. Performance metrics used for selection of the appropriate model are Accuracy, Kappa, Area under the ROC curve, specificity, sensitivity, Log-loss, Precision and Recall.

Table 3 Parameters of the Optimal Ex-LM model

Model	nhid	actfun
Ex-LM-1	14	Radial basis

Table 4 Parameters of the Optimal Rf model

Model	mTry	Metric
Rf-1	2	Accuracy
Rf-2	1	ROC
Rf-3	3	Logloss
Rf-4	7	AUC

4.2 Result

4.3 Artificial Neural Network: [7, 8]

This model is a single hidden layer feed-forward neural network whose tunable parameters are number of hidden units ($nhid$) and activation function ($actfun$) chosen using the random hyper-parameter optimization [9]. The optimal model was chosen by fivefold cross validation repeated 5 times on the basis of Accuracy metric (other metrics aren't calculated for this model as it doesn't give class probabilities) (Table 3).

4.4 Random Forests (Rf): [10]

Random forests were the next model trained on the interaction characteristics. Tunable parameter is the number of trees ($mTry$) for this model. Four different optimal models were selected one each for the metrics Accuracy, Area under ROC curve (ROC), Area under Precision-recall curve (AUC) and logloss. fivefold cross validation was used with 5 times repeat (Table 4).

4.5 Stochastic Gradient Boosted Trees (GBM): [11]

Stochastic Gradient Boosted Trees were the next model trained on the interaction characteristics. Tunable parameter is the boosting iterations ($n.trees$), maximum tree depth ($interaction.depth$), ($shrinkage$) and minimum terminal node size ($n.minobsinnode$) for this model. Four different optimal models were selected one each for the metrics Accuracy, Area under ROC curve (ROC), Area under Precision-recall curve (AUC) and logloss. fivefold cross validation was used with 5 times repeat (Table 5).

Table 5 Parameters of the Optimal GBM model

Model	n.trees	interaction.depth	shrinkage	n.minobsinnode	Metric
GBM-1	2695	2	0.06	12	Accuracy
GBM-2	190	1	0.42	13	ROC
GBM-3	32	2	0.56	12	Logloss
GBM-4	196	10	0.32	20	AUC

Table 6 Parameters of the Optimal xgbTree model

Model	nrounds	max-depth	eta	gamma	col-bt	min-cw	subsample	Metric
xgbTree1	573	10	0.09	5.01	0.64	19	0.91	Accuracy
xgbTree2	816	1	0.4	3.02	0.55	11	0.49	ROC
xgbTree3	901	1	0.06	7.65	0.64	8	0.72	logloss
xgbTree4	175	2	0.05	2.7	0.59	13	0.81	AUC

4.6 Extreme Gradient Boosted Trees (xgbTree): [11]

Extreme Gradient Boosted Trees model was trained on the interaction characteristics. Tunable parameter are the Boosting Iterations ($nrounds$), Max Tree Depth ($max - depth$), Shrinkage (eta), Minimum Loss Reduction ($gamma$), Sub-sample Ratio of Columns ($col - bt$), Minimum Sum of Instance Weight ($min - cw$), and Sub-sample Percentage ($subsample$) for this model. Four different optimal models were selected one each for the metrics Accuracy, Area under ROC curve (ROC), Area under Precision-recall curve (AUC) and log-loss. fivefold cross validation was used with 5 times repeat (Table 6).

Optimal models selected from the above procedure were selected evaluated on the test data. Based on the evaluation metrics it can be inferred the xgbTree models fit the test data better than other ML techniques applied on the training dataset. xgbTree models are ensembled predictors of balanced decision trees and also have the advantage of not overfitting the data.

Table 7 demonstrates that Extreme Gradient Boosted Trees (xgbTree) were found to be the most appropriate probabilistic models amongst those reviewed in this study for influence ranking on the Twitter dataset. xgbTree model supports parallelization of tree construction, cache optimization of data structures and algorithm to make best use of hardware, distributed computing and out-of-core computing for very large datasets that do not fit into memory for training very large models using a cluster of machines [12].

Table 7 Results of Optimal ML based model on test data

Model	Accuracy	Kappa	ROC	Sensitivity	Specificity
Ex-LM-1	0.66	0.32	–	–	–
Rf-1	0.77	0.54	0.77	0.77	0.78
Rf-2	0.77	0.54	0.77	0.76	0.77
Rf-3	0.77	0.54	0.77	0.76	0.78
Rf-4	0.77	0.54	0.77	0.76	0.78
GBM-1	0.78	**0.57**	0.78	0.76	0.79
GBM-2	0.78	0.55	0.78	0.77	0.78
GBM-3	0.78	0.56	0.78	0.76	0.79
GBM-4	0.76	0.52	0.76	0.74	0.77
xgbTree-1	**0.79**	**0.57**	0.79	**0.77**	**0.8**
xgbTree-2	0.77	0.53	0.77	0.75	0.78
xgbTree-3	0.78	**0.57**	0.78	**0.77**	0.79
xgbTree-4	**0.79**	**0.57**	**0.79**	**0.77**	**0.8**

5 Conclusion

This inquiry was conducted on the intuition that interaction occurs between entities of any social network and this could play an important role in modelling of influence. Hence statistics related to interactions made by entities in a social network were utilized to model influence and rank the entities as per their influence. Joint and conditional probability based techniques were used such as Neural networks, decision trees, ensembled predictors to conduct the inquiry. Based on the evaluation of the performance of the learning techniques on the data it was concluded that Extreme gradient boosted trees were the most suitable amongst the techniques used. The key advantage of this technique over other methods would be that it could scale well over massive online networks compared to other influence ranking models.

References

1. Aggarwal, C.C.: An introduction to social network data analytics. In: Social Network Data Analytics, pp. 1–15 (2011)
2. Sun, J., Tang, J.: A survey of models and algorithms for social influence analysis. In: Social Network Data Analytics, pp. 177–214 (2011)
3. Scripps, J., Tan, P.-N., Esfahanian, A.-H.: Measuring the effects of preprocessing decisions and network forces in dynamic network analysis. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 747–756. ACM (2009)
4. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146. ACM (2003)

5. Rao, A., Spasojevic, N., Li, Z., DSouza, T.: Klout score: measuring influence across multiple social networks. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 2282–2289. IEEE (2015)
6. Nerurkar, P., Bhirud, S.: Modeling influence on a social network using interaction characteristics. Int. J. Comput. Math. Sci. 152–160 (2017)
7. Zurada, J.M.: Introduction to Artificial Neural Systems, vol. 8. West St. Paul (1992)
8. Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 IEEE International Joint Conference on Neural Networks, 2004. Proceedings, vol. 2, pp. 985–990. IEEE (2004)
9. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**, 281–305 (2012)
10. Breiman, L.: Random forests. Mach. Learn. 5–32 (2001)
11. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232 (2001)
12. A gentle introduction to XGBoost for applied machine learning, Sep 2016