A Scalable Data Mining Model for Social Media Influencer Identification

Jyoti Sunil More[1,2(✉)] and Chelpa Lingam[3]

[1] Ramrao Adik Institute of Technology, Mumbai, India
[2] Department of Computer Engineering, Lokmanya Tilak College of Engineering (Affiliated to University of Mumbai), Navi Mumbai, India
jyotis8582@gmail.com
[3] Department of Computer Engineering, Pillai's HOC College of Engineering (Affiliated to University of Mumbai), Rasayani, India
chelpa.lingam@gmail.com

Abstract. Social network mining is a growing research area which combines together different fields such as machine learning, graph theory, parallel algorithms, data mining, optimization, etc., with the aim of dealing with issues like behavior analysis, finding interacting groups, finding influencers, information diffusion, etc. in a social network. This paper deals with one of these important issues i.e., Influencer Identification in social networks. This paper presents a data mining modelling approach for a twitter network, to find the most influential user among the given pair of users. This could be scaled over the entire network. We used a data mining model to score the test data and predict the influential user among the given pair of users. This approach of modeling can potentially be used for building many of the marketing and sales strategies wherein the influencer may be motivated for diffusing information or new ideas.

Keywords: Data mining · Influencers · Social network mining · Decision tree · Logistic regression

1 Introduction

Data mining is the process of studying data having different hidden behaviour, analyzing the patterns and deploying it to produce significant information. This information further can be used for carrying out predictive analytics and descriptive analytics. To construct a data mining model, we need to uncover the characteristics of dataset, create a model and deploy it [1].

Social network analysis [SNA] is the process of analyzing the behavior and interactions between different entities in the social networks. SNA has a great potential to evaluate the issues like, the likelihood of a particular community to grow, probability that the node gets influenced by other node, probability that a node acts as an influencer, etc. [1, 2].

The main aim of SNA is to explain the dependencies between the attributes of related nodes and predict the attributes like link probabilities, node behaviour, etc. in a given social network [1].

A. Unal et al. (Eds.): SmartCom 2016, CCIS 628, pp. 625–631, 2016.
DOI: 10.1007/978-981-10-3433-6_75

In this context, social influence can be defined in terms of conformity or the act of manipulating attitudes and behaviors, or dominance on peer group. Three broad varieties of social influence have been identified by Herbert Kelmen [3] as: Compliance, Identification and Internalization. Social Media Influencer Marketing [4] is the strategy of identifying the influencing people and motivating them to form new customer pool for the owners. It is irrespective of the size of pool of the audience of the influencer. The influencer can reach to consumers more efficiently than the direct ability of the brand itself.

2 Related Work

Different measures have been developed to identify the influencers in social networks. There are various ways to measure influence, some of them include number of followers, outreach, degree centrality, etc. Some of the tools used for this are Klout score, Kred, topsy, peerindex, etc.

Duanbing Chen, L. Lu, M. Sheng Shang, Y Cheng Zhang, T. Zhou [5], to observe the pattern in which the influence gets propagated among the nodes, proposed a semi-local centrality measure. For this, they used Susceptible Infected Recovered (SIR) model. In this, different centrality measures were used to rank the nodes. They analyzed the same for several real networks and showed that their approach involving the degree centrality of the nodes was relatively efficient to determine the influential nodes than other network parameters. Christine Kiss and Martin Bichler [6], carried out comparison of different centrality measures with respect to their impact on message distribution i.e., diffusing information in social networks. They examined existing measures and also evaluated the outdegree and SenderRank as centrality measures for the message distribution in social networks. They found that the performance of SenderRank was relatively better than other parameters. Na Li and Denis Gillet [7] investigated the functioning measures of scholars and observed their influence. The experimentation was carried out by aggregating various network centrality metrics.

Zsolt Katona, Peter Pal Zubcsek and Miklos Sarvary [8], in their experiment, analyzed the social network data and attempted to identify the impact of the word of mouth in the network. Their main goal was to discover how the structure of local communication network affects the information diffusion process. Their findings showed that in case of strong communities, beyond the network size, the influence of word of mouth is more effective. Their proposed model has a potential to identify customer pool who in turn will act as influencers for the diffusion activity of the new product or service.

Patrali Chatterji [9] in her research emphasized the significance of recommendation and referral behavior to social media and incorporates the role of underlying covariates. E. Bakshy, J. Hofman, W. Mason, D. Watts [10] investigated two groups of users and found that the users who have a record of being influential in the past and the users who have large number of followers, generate the largest share of information that is passed in successive levels. They concluded that the information diffusion through the word of mouth directly gets affected by the number of potential influencers and hence they are

the potential targets. Isabel Anger, Christian Kittl [11] proposed an approach which presented a quantitative method of determining twitter SNP. They proposed that there are two major aspects of Twitter: content and influence. They deliberately dropped an important factor i.e. number of followers and also proposed that the influence is largely affected by personal relations. Eytan Bakshy, Brian Karrer, Lada Adamic [12] developed a simple model, which exploited the information about the evolving structure of social network. This model gives a greater flexibility and shows the significance of network effects in the adoption of social network content. It is used to analyze the influence identification by exploiting the relationship and the information adoption rate among the nodes.

3 Problem Formulation

Given a social network interpreted as a directed graph, G = (V, E) where, V represents set of social network nodes (users) and E represents the network edges. The objective is influence detection, i.e. finding the set of nodes which can be targeted to diffuse an idea across the network such that the spread will be maximum. This can be viewed as a scalable problem. Our contribution is to define and analyze a model using the data mining stages, for two users at a time. This analysis can be scaled to n nodes in the network.

We consider a dataset comprising of only two users and their influence status. This analysis can be further scaled to other users from the network and overall influencers could be possibly found out. The dataset [13], provided by Peerindex, comprises a standard, pair-wise data records, meant for preference learning task. Each record describes two individuals, A and B with few characteristics. For each person, 11 pre-computed, positive numeric features based on twitter activity (such as number of mentions received, number of followers, etc.) are provided [13].

The binary label in the dataset represents a human judgment mentioning which one of the two individuals is more influential. A label '1' means participant A is more influential than B. 0 means participant B is more influential than A. The objective is to train a machine learning model which predicts the human judgment on who among A and B is more influential, with high accuracy.

The data consisted of 11 various parameters like: followers, retweets mentioned, number of posts, etc.

Exploring the data generally involves basic analysis of a dataset. This can be achieved by observing variable summaries and visual plots. By exploring our data, we can observe few characteristics of the data i.e. its range, its numeric characteristics mean, its spread, its deviation etc. There are numerous inherent problems associated with the data that include missing values, outliers i.e. one exhibiting abnormal characteristic, erroneous data, and skews in the data distributions. This in turn will affect the choice of the most appropriate tools for preparing and transforming our data and for learning the patterns in the data.

The input training data is found to be skewed and is largely messed with outliers. The dataset is first normalized. The linear model is checked for skewness, and skedesticity and was found adequate to proceed. The data is further tested for the data showing

abnormal characteristics i.e. outliers. The outliers are the data points which cause noise in the model and hence they should be detected and treated appropriately. A boxplot can be used for detecting the outliers as shown in Fig. 1. The outliers can be smoothened by substituting them with mean value. Correlation between all the variables is checked for collinearity and multicollinearity.

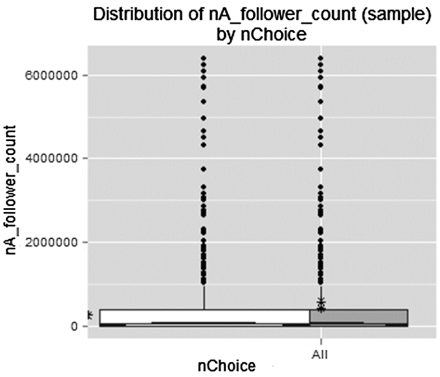

Fig. 1. The box plot showing the outliers

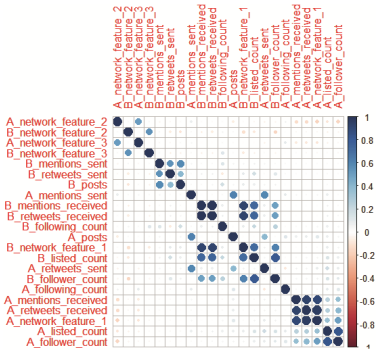

Fig. 2. Correlation matrix

Correlation in a given model can be either positive or negative. It represents the relationship mutually between two variables, and also cardinal i.e. one against all other variables. For Example, a strong correlation is observed between the parameters, A_retweets_received and A_mentions_received as shown in Fig. 2.

4 Model Building and Validation

This stage involves considering various data mining models depending on the characteristics of data. The predictive analytics phase may use mathematical models, graphical models, statistical models, etc. to draw inferences. This stage is considered as core of predictive data mining.

The entire dataset is distributed as 70/15/15 for training, validation and testing respectively. The model building involves two functions- Description and Prediction. For descriptive analytics, the techniques like clustering, summarization, linguistic summary, visualization etc. can be used. For predictive analytics any of the techniques like classification or regression can be used. Depending on the type of the data one of the suitable technique is chosen. In this context, the prediction is to be done by modeling the binary value- Choice of the most influential participant among the given pair of twitter users. The resulting probabilities are restricted to [0, 1] through the logistic distribution function [14]. The logistic regression predicts the event rate, i.e., probability of the event. Hence logistic regression technique is preferred for this case study. The general equation for logistic regression [14] is given as-

$$\text{Log}(p/(1-p)) = \beta 0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_n X_n + e \quad (1)$$

The log of odds function denotes the expected probability that the outcome is present, X_1 to X_n are independent variables and β_0 to β_n are the regression coefficients. Null Hypothesis is defined as: The data coefficients β are null. Our objective is to model and show that the β values are not null. After reviewing the statistical significance of the parameters using the Z score and P values, some of the insignificant parameters were dropped from the regression. The retained parameters were as follows-

A_listed_count, A_mentions_received, A_mentions_sent, A_network_feature_2, B_listed_count, B_mentions_received, B_mentions_sent, B_retweets_sent, B_network_feature_2, B_network_feature_3.

Using the outlier test it is confirmed that no studentized residuals remain in the data. The logistic regression model is shown in Fig. 3. Another modeling technique was used to predict the influential user i.e. Regression tree or classification tree.

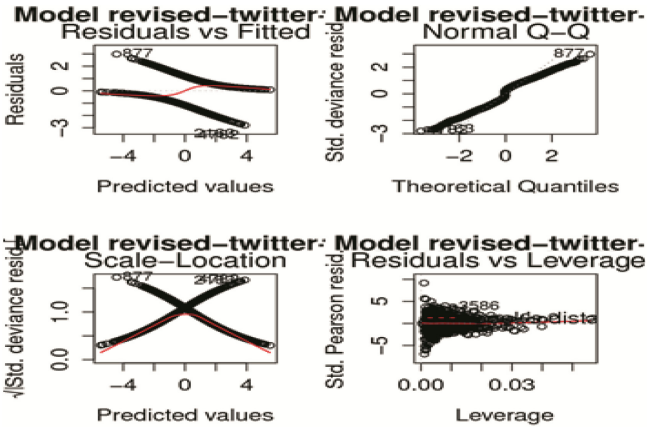

Fig. 3. Logistic regression model

5 Results and Interpretation

The variables actually used in the tree construction are- A_follower_count, A_network_feature_1, A_network_feature_2, B_follower_count, B_listed_count, B_network_feature_1, B_network_feature_3.

Error matrix for the Linear model on revised twitter data and the error matrix for the Decision Tree model on revised twitter data could be compared. The precision and recall parameters could be used to evaluate the models. The graphs when plotted are visualized and are shown in the following Figs. 4, 5, 6 and 7.

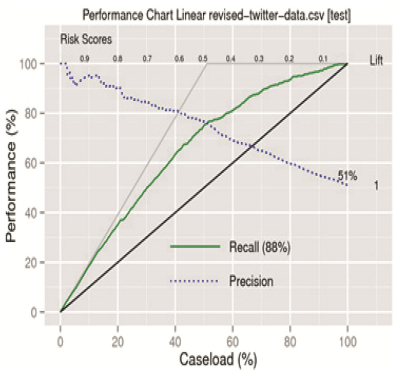

Fig. 4. The precision-recall for logistic regression model

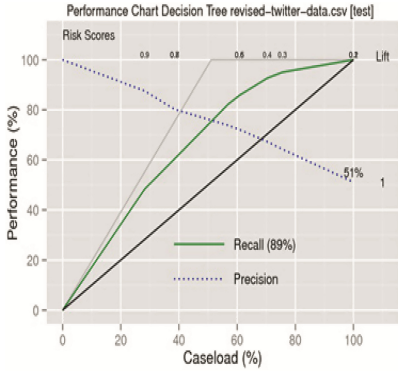

Fig. 5. The precision-recall for decision tree model

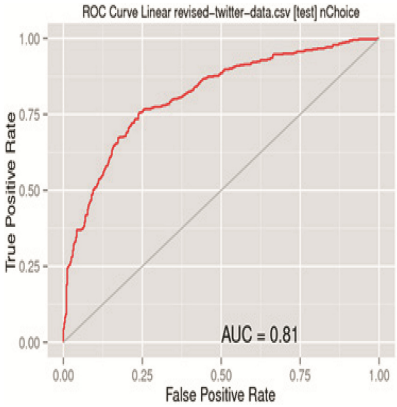

Fig. 6. ROC chart for logistic regression model

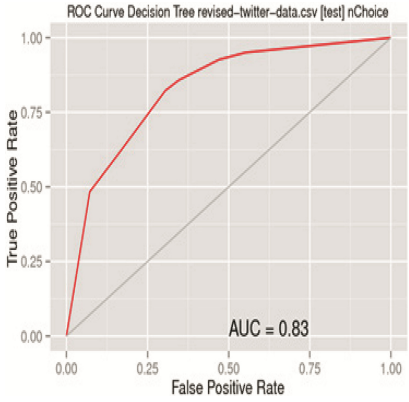

Fig. 7. ROC chart for decision tree model

6 Deployment

This is the final stage of predictive analytics. In this, the elite models build from the training set and chosen parameters, are preserved and are applied to the test data in order to generate predictions of the expected results. It is sometimes termed as score.

The decision tree and Logistic regression models have above average accuracy (Figs. 4, 5, 6 and 7) and hence can be used to find the score of the test data. The test data is scored on their probabilities i.e. what is the probability that either A or B is influencer.

7 Conclusion and Future Scope

Social network mining is a growing, exciting area of research that has a great scope, with the contribution of many research fields. Understanding the user behavior in a social

network has a potential to improve the marketing strategies by targeting precisely the influencers for information diffusion, or propagating ideas and thus optimizing the viral marketing techniques. From the point of view of mining and analysis of social networks, it is also necessary to develop efficient graphical models to exploit the structural properties of network and different statistical methods which are efficient and uniform. Different influence maximization approaches can be proposed. Optimizing the spread could be one of the goals. Considering large size of the social networks and redundant nature of sub graphs in a social network graph, the parallelism in the graphical structure could be exploited.

References

1. More, J.S., Lingam, C.: Reality mining based on social network analysis. In: Proceedings of IEEE International Conference on Communication Information and Computing Technology (ICCIT), pp. 1–6 (2015)
2. Huang, F., Cheng, N.X., Xiao, R.: An approach to mining social networks in chat room. J. Comput. Inf. Syst. **1**, 135–143 (2011)
3. Kelman, H.: Compliance, identification, and internalization: three processes of attitude change. J. Conflict Resolut. **2**, 51–60 (1958)
4. www.grouphigh.com
5. Chen, Duanbing, Lü, L., Shang, M.S., Zhang, Y.C., Zhou, T.: Identifying influential nodes in complex networks. Phys. A Stat. Mechan. Appl. **391**(4), 1777–1787 (2011)
6. Kiss, Christine, Bichler, Martin: Identification of influencers- measuring influence in customer networks. Decis. Support Syst. **46**, 233–253 (2008)
7. Li, N., Gillet, D.: Identifying influential scholars in academic social media platforms. In: ASONAM Proceedings IEEE/ACM International Conference on Advances in Social Network Analysis and Mining, pp. 608–614 (2013)
8. Katona, Z., Zubcsek, P.P., Sarvary, M.: Network effects and personal influences: the diffusion of an online social network. J. Mark. Res. **48**(3), 425–443 (2011)
9. Chatterjee, P.: Drivers of new product recommending and referrel behavior at social network sites. Int. J. Advertising **30**(1), 77–101 (2011)
10. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J: Everyone's an influencer: quantifying influence on twitter. In: WSDM Proceedings of Fourth ACM International Conference on Web Search and Data Mining, pp. 65–74 (2011)
11. Anger, I., Kittl, C.: Measuring influence on twitter. In: International Conference on Knowledge management and Knowledge Technologies. ACM (2011)
12. Bakshy, E., Karrer, B., Adamic, L.A.: Social influence and the diffusion of user-created content. In: Proceedings of the 10th ACM conference on Electronic commerce, pp. 325–334 (2009)
13. www.kaggle.com
14. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, pp. 130–137. Springer, New York (2013)