Computers in Human Behavior 60 (2016) 534–541

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

A comprehensive study on the effects of using data mining techniques to predict tie strength

Mohammad Karim Sohrabi*, Soodeh Akbari

Department of Computer Engineering, Semnan Branch, Islamic Azad University, Semnan, Iran

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form
23 February 2016
Accepted 24 February 2016
Available online xxx

Keywords:
Data mining
Tie strength
Profile-behavioral based model
Classification techniques

ABSTRACT

The use of social networks has grown noticeably in recent years and this fact has led to the production of numerous volumes of data. Data that are widely used by users on the social media sites are very large, noisy, unstructured and dynamic. Providing a flexible framework and method to apply in all of these networks can be the perfect solution. The uncertainties arising from the complexity of decisions in recognition of the Tie Strength among people have led researchers to seek effective variables of intimacy among people. Since there are several effective variables which their effectiveness rate are not precisely determined and their relations are nonlinear and complex, using data mining techniques can be considered as one of the practical solutions for this problem. Some types of unsupervised mining methods have been conducted in the field of detecting the type of tie. Data mining could be considered as one of the applicable tools for researchers in exploring the relationships among users.

In this paper, the problem of tie strength prediction is modeled as a data mining problem on which different supervised and unsupervised mining methods are applicable. We propose a comprehensive study on the effects of using different classification techniques such as decision trees, Naive Bayes and so on; in addition to some ensemble classification methods such as Bagging and Boosting methods for predicting tie strength of users of a social network. LinkedIn social network is used as a real case study and our experimental results are proposed on its extracted data. Several models, based on basic techniques and ensemble methods are created and their efficiencies are compared based on F-Measure, accuracy, and average executing time. Our experimental results show that, our profile-behavioral based model has much better accuracy in comparison with profile-data based models techniques.

1. Introduction

Social graphs and statistical methods construct the basis of traditional researches in mining and analyzing social information. However, despite their improvements, there are some issues with the above-mentioned methods such as having a large number of variables and the need for making hypothesis (because of their statistical nature). The purpose of this research is to develop a method that can overcome the defects of former methods and make the extracted rules closer and more understandable to human language.

Many researchers have adopted identifying tie strength as an analytical framework for the study of individuals and organizations. For example, Google Scholar claims that more than 29,000 articles have cited research paper of "The Strength of Weak Ties" by Granovetter (1973). Social support provided by strong ties may actually be effective in improving the mental health (Schaefer, Coyne, & Lazarus, 1981). Due to the increasing volume of data in the 21st century, and regarding the benefits of using data mining techniques including: being dynamic, real time analysis, real data analysis, creating real models, rising computing power, having advanced algorithms and powerful software tools, have caused data mining to be in a special place among the various sciences. Thus, data mining techniques can be helpful in diagnosing tie strength (Tan, Steinbach, & Kumar, 2005; Larose, 2005).

The problem of imbalanced class should be considered in using data mining techniques for predicting tie strength. A dataset is imbalanced if the classes are not approximately equally represented. There have been attempts to deal with imbalanced datasets in real issues such as pollution, risk management, fraud detection, and medical diagnosis. In these cases, standard classifier learning

* Corresponding author.
E-mail addresses: Amir_sohraby@yahoo.com (M.K. Sohrabi), Sou.Akbari@gmail.com (S. Akbari).

http://dx.doi.org/10.1016/j.chb.2016.02.092
0747-5632/

algorithms have a bias toward the classes with greater number of instances, since rules that correctly predict those instances are positively weighted in favor of the accuracy metric, whereas specific rules that predict examples from the minority class are usually ignored (treating them as noise), because more general rules are preferred. In such a way, minority class instances are more often misclassified than those from the other classes (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012). In this paper, we focus a class imbalanced data-set, where there is a minority class (strong tie), with the lowest number of instances, and a majority class, with the highest number of instances (weak tie) so we used over sampling and under sampling techniques to solve imbalanced class-variable problem.

In this paper, we propose a comprehensive study of the effect of using different classification techniques such as Decision Trees (DTs), Naive Bayes (NB), Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), K Nearest Neighbors (KNNs), and ensemble classification methods such as Bagging and Boosting methods to predict tie strength of users in a social network. A case study on the database of LinkedIn users is also conducted. It is worth noting, required Information is obtained from LinkedIn via API functions with the users' approval. Then the effective data mining technique for eliciting the tie strength between users with their friends on this social network has been proposed. Finally, decision tree-bagging model with an accuracy of about 86% and F-measure of about 42% and after that artificial neural network-Bagging model with accuracy of about 85% and F-Measure of about 77% performed better. Also Naïve bays-AdaBoost and SVM-AdaBoost had the lowest accuracy in predicting the Tie Strength.

Here are the main contributions of this paper:

1) We propose a new model to use classification techniques and apply related data mining methods on the users' behavior of social networks to predict their tie strength.
2) A new framework has been designed to implementing and applying different classification techniques such as decision trees, and some ensemble classification methods such as Bagging and Boosting, for predicting tie strength of users of a social network.
3) We use LinkedIn social network as case study and compare efficiency of based and ensemble methods on profile-behavioral based model and profile-data based model for its users, based on F-measure, accuracy and average executing time.

The remaining of this paper is organized as follows: in Section 2 related works are introduced. Tie Strength analysis in social network and data mining background reading are two different subsections of the related works. In Section 3, our new model is represented and the flowchart of the mining process is proposed. Experimental results and evaluations are proposed in Section 4, and we conclude our work in Section 5.

2. Related work

In the beginning of this section, a brief history of tie strength in social networks is represented and then some important works in data mining will be explained.

2.1. Tie strength analysis in social networks

The concept of tie strength was introduced in 1973 by Granovetter (1973), who defines it as a function of duration, emotional intensity, intimacy and exchange of services through which, ties are split into ‘strong’ and ‘weak’. Lin, Ensel, and Vaughn (1981) suggested that social distance, education level, socioeconomic status, political affiliation, race, and sex are effective on tie strength. Marsden and Campbell (1984) also did some researches to predict Tie Strength But a key constraint was that the participants were asked to state ten major characteristics of their three closest friends. In 1988, the paper (Krackhardt & Stern, 1988) demonstrated that a strong relationship between employees of different organizational sub-units can help an organization to resist in the crisis. Wellman and Wortley (1990) believe that providing emotional support such as providing advice on family problems indicates strong tie. In 1992, the paper (Krackhardt, 1992) discussed that strong partners create crisis and pressure for institutional changes in the organization. Granovetter (1995) also demonstrated that weak ties, in contrast with strong ones, are more beneficial for job seekers. Wilson, Boe, Sala, Puttaswamy, and Zhao (2009), Viswanath, Mislove, Cha, and Gummadi (2009), and Backstrom, Bakshy, Kleinberg, Thomas, and Itamar (2011) have studied the activities of Facebook users, taking the different interactions into account as a sign. In 2009, these symptoms for Wilson et al. (2009) are links of each user in the social graph, including wall-posts and photo-comments and interactions among friends on Facebook. Viswanath et al. (2009) carried out their study only with the wall-posts in order to study different interaction patterns, influenced by the overall structure of the network over time. Kahanda and Neville (2009) studied the nature and strength of the relationship between Facebook members using the characteristic features (gender, marital status, status, etc.), topological features (User connectivity graph of friendship), transactional characteristics (wall posts, upload photos and group members) and network-transactional features (shared wall-posts) to realize the “special friends”. They concluded that the most prominent features in predicting tie strength are network-transactional ones. Burt (2009) believes that structural factors are effective in shaping tie strength, such as network topology and informal social circle. In 2011, Backstrom et al. (2011) studied how users allocate attention to their Facebook friends, by taking messages, comments, Wall—post's and information on the number of times each user's profile page or photo submissions is viewed by another user into account. In 2011, the classification of ties was performed through a formula in which the number of users and the relation degrees had been used (Cho et al., 2011). Again In 2011 a study was conducted using regression analysis. This study was based on this principle that Tie Strength is a combination of the variables such as: friendship, the intensity of feelings, intimacy, mutual trust and mutual services, considering tie strength as a linear combination of predictive variables and network structures. The proposed model is based on a data set of two thousands of Facebook. In this study with an accuracy of about 85% more than 70 variables were used (Gilbert & Karahalios, 2009). In 2012, a survey based on questions and answers have been proposed (Panovich, Miller, & Karger, 2012). First, users respond to questions that had been put on their Facebook page. The responses were analyzed and the feedbacks from respondents were obtained. Then regarding a psychological research, which says: “weak ties provide more useful information than strong ties”, this research proves the contrary. In (Servia-Rodríguez, Díaz-Redondo, Fernández-Vilas, & Pazos-Arias, 2013), needed information has been extracted through Facebook API's (with users permission). variables used in this study as a sign of interaction are: private messages, wall-posts, uploaded photos and videos, and the number of Likes, comments, the number of tags in the pictures considering the number of people in the photo, membership of the joint groups, last updated posts, conversations and so on. In 2014, evaluation of performance testing, is performed by means of BFF (Fogués, Such, Espinosa, & Garcia-Fornes, 2014). This tool predicts tie strength through relationships in social pages of users. In this study, tie-strength prediction has been based on a linear combination of 14