Expert Systems with Applications 41 (2014) 6575–6584

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Improved churn prediction in telecommunication industry by analyzing a large network

Kyoungok Kim [a,1], Chi-Hyuk Jun [a,1], Jaewook Lee [b,*]

[a] Department of Industrial and Management Engineering, Pohang University of Science and Technology, 790-784 Pohang, Kyungbuk, South Korea
[b] Department of Industrial Engineering, Seoul National University, 151-744 Seoul, South Korea

ARTICLE INFO

Article history:
Available online 17 May 2014

Keywords:
Churn prediction
Network analysis
Community detection
Diffusion process

ABSTRACT

Customer retention in telecommunication companies is one of the most important issues in customer relationship management, and customer churn prediction is a major instrument in customer retention. Churn prediction aims at identifying potential churning customers. Traditional approaches for determining potential churning customers are based only on customer personal information without considering the relationship among customers. However, the subscribers of telecommunication companies are connected with other customers, and network properties among people may affect the churn. For this reason, we proposed a new procedure of the churn prediction by examining the communication patterns among subscribers and considering a propagation process in a network based on call detail records which transfers churning information from churners to non-churners. A fast and effective propagation process is possible through community detection and through setting the initial energy of churners (the amount of information transferred) differently in churn date or centrality. The proposed procedure was evaluated based on the performance of the prediction model trained with a social network feature and traditional personal features.

1. Introduction

Many mobile telecommunication companies face extremely challenging business environments because the market is already saturated. As a result, customer retention is one of the most important issues for producing higher revenue and margin because retaining customers is less costly and more profitable than attracting new subscribers (Euler, 2005). Therefore, building an accurate predictive model for latent churners helps telecommunication companies to survive in a competitive environment.

The main goal of churn prediction is to classify customers into churners and non-churners, providing efficient target marketing for potential churners. This problem can be viewed as a binary classification problem, and thus numerous machine learning techniques such as logistic regression (Kim, 2006; Lima, Mues, & Baesens, 2009; Mozer, Wolniewicz, Grimes, Johnson, & Kaushansky, 2000), neural networks (Hung, Yen, & Wang, 2006; Mozer et al., 2000), decision trees (Hung et al., 2006; Lima et al., 2009) and support vector machines (Archaux, Martin, & Khenchaf, 2004) have been successfully applied for churn prediction in telecommunication companies. However, most of these studies predict potential churners based only on individual customers without considering the relationships among customers.

Some studies have attempted to distinguish potential churners from non-churners using the results of network analysis based on call detail records (CDRs) collected from telecommunication companies to reflect the realistic situation of customer influence on one another. Richter, Yom-Tov, and Slonim (2010) computes the churn probability by initially calculating group churn scores and then obtaining individual churn scores based on social rank within their social group, thereby distinguishing churners by analyzing social groups that customers belong to Dasgupta et al. (2008) used one of the propagation processes and predicted potential churners based only on the results of the propagation process, thereby providing social ties and their relevance to churn in the networks. The main assumption of these studies is that a certain customer's churn is shared with other people linked through their network, which increases the churn probability for non-churners. This has been proven by these two works. However, there is still a room for improvement because they utilized only the social relations in the CDR data.

In the present paper, we investigate the effect of relationships among customers in the mobile telecommunication company on the customer retention by introducing a new variable obtained from a network analysis. This new variable, called the network

* Corresponding author. Tel.: +82 2 880 7176; fax: +82 2 889 8560.
E-mail address: jaewook@snu.ac.kr (J. Lee).
[1] Tel.: +82 54 279 2894.

http://dx.doi.org/10.1016/j.eswa.2014.05.014
0957-4174/

variable, is calculated through the propagation process. In our study, we adopt the spreading activation (SPA) model for the propagation process and use the network variable as well as the traditional customers' personal information. In addition, we consider various characteristics of all initial churners who spread churning information, whereas existing research treated initial churners as the same (Dasgupta et al., 2008). We also adopt community detection to implement a propagation process for a large network constructed using the CDR data. The following two assumptions are implemented in our study to improve the existing model.

Churning information shared in the same community has more effect on customers in that community than that from the other community. Based on this assumption, we apply community detection algorithm to a large network to partition the network into exclusive smaller networks. Churners who unsubscribe from the service in advance cannot exert influence on people in other communities because all edges between the two different communities are removed after community detection; hence, the time for the propagation process can dramatically decrease.

All initial churners for the propagation process are not the same. Some churners may be more effective and influential in leading other customers' churn than other churners. We consider two factors to determine the relative importance of initial churners: churn date and centrality in their community. With regard to churn date, initial churners who quit the service or transfer to competitors at a date close to the specific date for the propagation process may have more influence on other subscribers. In addition, nodes with high centrality in the community represent the central or key customers in the community and may thus be more influential to other customers. Therefore, customers who unsubscribed and have high centrality can lead more customers to quit the service following prior churners.

These two assumptions are verified using the customer personal information data and the CDR data from the mobile telecommunication company in this study.

The remainder of the paper is organized as follows: In Section 2, we provide literature reviews related with our study and in Section 3, give the detailed process of our experiments. In Section 4, we show and analyze results of experiments. Finally, we conclude the paper in Section 5.

2. Literature review

2.1. Community detection

Inhomogeneity that shows a high level of order usually exists in real networks such as a social network because real networks are not random graphs. The aim of community detection in a graph is to divide a network into non-overlapping groups of nodes with dense connections internally and sparse connections between different groups as much as possible. Community detection has been applied in areas such as clustering customers or web pages, providing an efficient way to deal with a large number of clients or web pages.

Identifying communities is a popular research area, but it is a difficult task and intractable. Several types of community detection algorithms have been proposed to simultaneously obtain better-quality communities and faster computation time. The first one is the divisive algorithm, which detects edges that connect nodes of different communities and removes them from the graph (Girvan & Newman, 2002; Newman & Girvan, 2004). This type of algorithm starts with only one community and splits them into several communities as the algorithm proceeds. The key point of this algorithm type is to distinguish the inter-community links that can allow for their identification. The second type is the agglomerative algorithm, which merges similar nodes or communities iteratively (Pons & Latapy, 2006). Unlike divisive algorithms, agglomerative algorithms initially set different communities to all nodes. Other widely developed methods include modularity-based algorithms, which optimize the modality function originally introduced to define termination condition for divisive algorithms (Clauset, Newman, & Moore, 2004). In this type of algorithm, the higher the value of modularity, the better the partition. However, it is impossible to reach the exact maximum value of modularity because there are several ways to divide a graph into smaller subsets. In addition, modularity optimization has been proved to be an NP-complete problem (Brandes et al., 2006). As a result, many algorithms have attempted to reach acceptably good solutions with reasonable approximations or heuristic ways. According to optimization techniques, these algorithms have several variations such as greedy methods (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008; Clauset et al., 2004), simulated annealing methods (Guimerà & Nunes Amaral, 2005; Guimerà, Sales-Pardo, & Amaral, 2004), spectral optimization methods (Newman, 2006a, 2006b) and etc.

In our study, to cluster customers with dense connections, we applied the Louvain algorithm, one of the modularity-based greedy community detection algorithms, to the network constructed by CDR data since it is reported that it can handle a large size of networks and outperforms other methods in terms of computation time (Blondel et al., 2008).

2.2. Centrality measure

Centrality measure is an evaluation criterion, which is a real-valued function that assigns each vertex in a network some value to distinguish more important or more central vertices from others. There are numerous measures for finding the key nodes in a network. Among these measures, degree, closeness, betweenness, and eigenvector centrality are widely used. Degree centrality is the number of edges that links to other nodes (Freeman, 1978). In a directed network, two separate measures of degree centrality, in degree and out degree, are defined. In degree is the number of edges directed to the node while out degree is the number of edges the node directs to other nodes. Closeness centrality is the mean length of all the shortest paths from a node to all other connected nodes in a network. However, the distance between two unconnected nodes is infinite. As a result, another closeness centrality is defined by the inverse of the average length of the shortest paths to or from all the other vertices in a graph (Freeman, 1978). In this setting, closeness is one when a node is maximally close to all other points. Betweenness centrality is the number of shortest paths that pass through a node divided by all the shortest paths in a network (Freeman, 1978). This measure shows which nodes are more likely to be in connected paths between other nodes. Eigenvector centrality corresponds to the values of the first eigenvector of the graph adjacency matrix; one node's eigenvector centrality is proportional to the sum of the centralities of the nodes directly connected to it (Bonacich, 1987). Nodes with high eigenvector centrality are generally those that are connected to other nodes with high eigenvector centrality. One of its variations is Google PageRank algorithm, which prioritizes searching results by considering how many links exist from highly linked pages.

In this paper, eigenvector centrality measure is used to identify the key customers who exert influence on other customers because high value of eigenvector centrality generally indicates a customer who is more central to the main pattern of distances among all customers. Moreover, we only consider the centrality measures of each node within the same community.

2.3. Propagation process

Propagation process is a model to describe the dynamics in a network, such as transferring epidemics, computer virus, and any