Egyptian Informatics Journal xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Churn prediction on huge telecom data using hybrid firefly based classification

Ammar A. Q. Ahmed *, Maheswari D.

Rathnavel Subramainam College of Arts & Science, Coimbatore, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 25 September 2016
Accepted 10 February 2017
Available online xxxx

Keywords:
Firefly algorithm
Simulated annealing
Telecom churn prediction
Data imbalance
Data sparsity
Huge data

ABSTRACT

Churn prediction in telecom has become a major requirement due to the increase in the number of telecom providers. However due to the hugeness, sparsity and imbalanced nature of the data, churn prediction in telecom has always been a complex task. This paper presents a metaheuristic based churn prediction technique that performs churn prediction on huge telecom data. A hybridized form of Firefly algorithm is used as the classifier. It has been identified that the compute intensive component of the Firefly algorithm is the comparison block, where every firefly is compared with every other firefly to identify the one with the highest light intensity. This component is replaced by Simulated Annealing and the classification process is carried out. Experiments were conducted on the Orange dataset. It was observed that Firefly algorithm works best on churn data and the hybridized Firefly algorithm provides effective and faster results.

1. Introduction

Increase in the number of telecom providers has led to a huge rise in competition and hence customer churn. Currently organizations have their major focus on reducing the churn by focusing on customers independently. Churn [1] can be defined as the propensity of a customer to cease business transactions with an organization. The major requirement now is identification of customers who have high probabilities of moving out. The ability of an organization to intervene at the right time could effectively reduce churn.

Churn occurs mainly due to customer dissatisfaction. Identifying customer dissatisfaction requires several parameters. A customer usually does not churn due to a single dissatisfaction scenario [2]. There usually exist several dissatisfaction cases before a customer completely ceases to do transactions with an organization. Several properties associated with the customer and their mode of operations with the organization are recorded by the organizations. This represents the customer's behavior data. Analyzing this data would present a clear view of the customer's current status [3]. Hence this can be used as the base data for churn prediction. The major difficulty arising from this mode of operation is that the data under discussion tends to be very huge. The hugeness can be attributed to the behavioral nature of the data, depicting all the product lines dealt with by the organization. Further, due to the requirement of structural representation of the data, all the instances are bound to contain all the properties corresponding to a generic customer in the organization [4,5]. This leads to data sparseness, since customers will be associated with only a few properties and not all the properties pertaining to the organization. The hugeness of data and sparsity acts as the major difficulties in the process of churn prediction.

Large companies interact with their customers to provide a variety of services to them [6]. Customer service is one of the key differentiators for companies. The ability to predict if a customer will leave in order to intervene at the right time can be essential for pre-empting problems and providing high level of customer service. The problem becomes more complex as customer behavior data is sequential and can be very diverse.

Churn is an unavoidable process in any industry. However, though difficult, it is possible to identify the causes of churn using several approaches.

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

* Corresponding author.
E-mail addresses: ammar.aqahmed@gmail.com (A. A. Q. Ahmed), mahelenin@gmail.com (D. Maheswari).

http://dx.doi.org/10.1016/j.eij.2017.02.002
1110-8665/

Please cite this article in press as: Ahmed AAQ, Maheswari D. Churn prediction on huge telecom data using hybrid firefly based classification. Egyptian Informatics J (2017), http://dx.doi.org/10.1016/j.eij.2017.02.002

2. Related work

This section discusses the recent approaches for churn prediction. A risk prediction technique that identifies probable customers for churn was presented by Coussement et al. in [7]. This technique utilizes Generalized Additive Models (GAM). These models relaxe the linearity constraints, hence allowing complex non-linear fits to the data. This technique is exhibited to improve marketing decisions by identifying the risky customers and also providing visualizations of non-linear relationships.

A neural network based customer profiling technique that can be used for churn prediction was presented by Tiwari et al. in [8]. This technique differs from the other proposed techniques by the fact that most of the techniques are only able to identify the customers who will instantaneously churn. However the neural network based churn prediction model proposes to predict customer's future churn behavior, providing the much required buffer for the organizations to perform prevention activities. A similar neural network based model includes [22,24]. The approach in [22] is based on the 80-20 rule to identify the key attributes affecting churn, while that of [24] involves identifying the major features of the data to determine churn.

A regression based churn prediction model was presented by Awnag et al. in [9]. This method identifies churn by using multiple regressions analysis. This technique utilizes the customer's feature data for analysis and proposes to provide good performance.

Class imbalance plays a major role in affecting the reliability of a classifier. The major issue existing due to class imbalance is that the minority class is not well represented and hence the classifier is undertrained on the minority classes. The technique proposed by Zhu et al. in [10] proposes to eliminate this issue by using transfer learning techniques. The approach presented in [10] operates by training the classifier using customer related behavioral data obtained from related domains. This approach has its major focus on the banking industry and the results are proposed to exhibit enhanced performance. Another technique that considers the imbalance nature of data to perform churn prediction was presented by Xiao et al. in [15]. A comparison of sampling techniques for effectively operating on churn data was presented by Amin et al. in [16]. Game theory based churn prediction techniques [17] are also on the raise.

The complex nature of churn behavior has also enabled several publications on churn prediction using multiple models. A churn prediction model based on cluster analysis and decision tree algorithm was presented by Li et al. in [11]. This technique operates on China's Telecom data. Another technique utilizing multiple prediction techniques was proposed by Le et al. in [12]. This technique utilized a combination of k-Nearest Neighbor algorithm and sequence alignment. This technique has its major focus on the temporal categorical features of the data to predict churn.

Utilizing heuristics for predictions are on the raise due to the complex nature of data. A rule generation techniques that employs heuristics for customer churn prediction in telecom services was presented by Huang et al. in [13]. A combination of Self Organizing Maps (SOM) and Genetic Programming (GP) to identify and predict churn was presented by Faris et al. in [14]. SOM is utilized to cluster the customers and then outliers are eliminated to obtain clusters depicting customer behaviors. An enhanced classification tree is built using GP.

A boosting algorithm that proposes to improve the prediction accuracy of classifier models was proposed by Lu et al. in [18]. This method boosts the learning process by using a combination of clustering and logistic regression. A similar prediction boosting technique using Genetic Algorithm was proposed by Idris et al. in [19]. This is also an ensemble model utilizing multiple techniques for the prediction process. Other ensemble based prediction techniques include [20,21,1,23].

3. Churn prediction on huge data using hybrid firefly based classification

Churn prediction on huge data utilizes Hybrid Firefly algorithm to effectively identify churn. This technique modifies the comparison component of the actual firefly algorithm with Simulated Annealing to provide faster and effective results.

A. Firefly algorithm: WorkingFirefly algorithm [25] is a nature inspired metaheuristic algorithm that was inspired by the behavior of fireflies attracting other fireflies by flashing lights. The intensity of the light plays a major role in determining the attractiveness of a firefly. It works on the following assumptions:

- All fireflies are unisexual, hence any firefly can be attracted to any other firefly.
- Attractiveness is proportional to the brightness of a firefly.
- For any two fireflies, the brighter one will attract the other.
- Brightness decreases as the distance between the fireflies increase.
- If no firefly is brighter than a given firefly, then it moves randomly.

For an optimization problem, the brightness of a firefly is associated with the objective function. The objective function contains all the parameters dependent on applications, hence expresses the degree of importance that the current solution holds.

B. Firefly algorithm: pros and cons

Firefly algorithm, due to its metaheuristic nature, can effectively identify optimal solutions when compared to other statistics based classification algorithms. Movement of the fireflies are directed by the intensity of the fireflies, provided by the firefly intensity parameter. The usage of a single dependent parameter leads to lesser memory requirements, hence this algorithm is capable of operating on huge data.

The major drawbacks of this algorithm is that for every iteration, a firefly is compared with every other firefly in the system [26], hence increasing the number of computations. Hence as the number of fireflies in the search space increases, the level of computations also increases to a large extent.

C. Hybrid Firefly: architecture

The hybrid firefly architecture is proposed to eliminate the problem of huge computational requirements due to comparisons. The working of hybrid firefly algorithm is presented in Fig. 1.

Building the search space marks the beginning of the classification process. The initial population of fireflies is generated and are distributed across the search space. The distribution of fireflies is carried out in random. Position of each firefly is recorded and the initial intensity of the fireflies (Intensity) are identified on the basis of their distance from the test data.

$$Intensity_i = 1/\sqrt{\sum_{j=1}^{attr}(X_{test,j} - X_{i,j})^2} \quad (1)$$

where $X_{test,j}$ refers to the j^{th} attribute of the test data and $X_{i,j}$ refers to the j^{th} attribute of the firefly i.

Please cite this article in press as: Ahmed AAQ, Maheswari D. Churn prediction on huge telecom data using hybrid firefly based classification. Egyptian Informatics J (2017), http://dx.doi.org/10.1016/j.eij.2017.02.002