Churn Prediction on Huge Telecommunication Industry Data using Optimization Data Classification Model

Karthikeyan A
Assistant Professor Mr. K. Nirmal.
Krishnasamy College of Engineering and Technology,
Cuddalore.

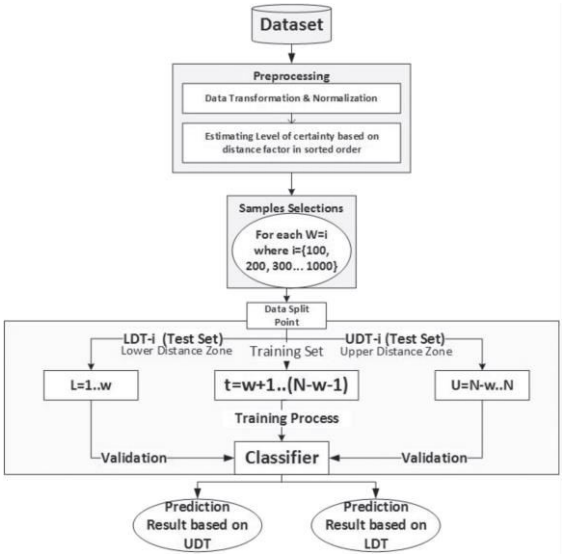
ABSTRACT

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 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 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 project presents a metaheuristic based churn prediction technique that performs churn prediction on huge telecom data. A Neural Network form of Deep Learning algorithm is used as the classifier. It was observed that proposed algorithm works best on churn data and the algorithm provides effective and faster results.

Introduction

The customers are considered one of the most important asset for a business in numerous dynamic and competitive companies within a marketplace In competitive market, companies in which the customers have numerous choice of service providers they can easily switch a service or even the provider.

Such customers are referred to as churned customer The causes of customer churn can be due to dissatisfaction, higher cost, low quality, lack of features, and privacy concerns Many organizations e.g., financial service are ever more focusing on establishingand maintaining the long-term relationships with their existingcustomers Loyal customers can be considered long-term customers that are not only profitable for the company but also are great ambassadors in the market One of the industry wherein this phenomenon is observed is the Telecommunication Industry (TCI). CCP in TCI is an increasingly well-known domain and popular research problem in the literature in recent It is reported that TCI is suffering from the substantial problem of customer churn due to fierce competition, saturated markets, dynamic condition, and launching new attractive offers It is observed that acquiring new customer can be more expensive for companies as compared to retention of the existing customer (Also, the researchers have confirmed that CCP approaches can improve a company's revenue and good reputation in market.



This is partly because most of the time, customer churn and non-churn have resembling features and behavior which increases the classification error rate. In other words, we can say that the classifier is uncertain about the decision and the level of certainty varies from case to case in the TCI. In this paper, we introduce a novel CCP approach using distance factor focusing on different distance zones (e.g., upper zone (greater distance factor's value) and lower zone (small distance factor's value)) pertaining to estimate the certainty of the classifier. Furthermore, the proposed CCP approach will not only predict the customer churns but can also calculate the level of the certainty of the prediction by evaluating the classifier's decision into the following categories, (i) customer churn and non-churn with high certainty, (ii) customer churn and non-churn with low certainty. The low certainty can be considered as uncertain classification for predicting the customer churns. The distance factors in term of upper and lower zones has not been considered for CCP in TCI yet. The proposed approach towards the target industry, exploring the discussed unexplored factors, can play a pivotal role in CCP models.

Related work

The review in this section is primarily related to exploring the stateof- the-art techniques for CCP that have been adopted for CCP. hybrid neural networks approach for CCP in a CRM dataset of the American telecommunication company. They used an approach in which they have combined artificial neural network (ANN) and self-organized map (SOM) for CCP model. The ANN is used for data reduction in which unrepresentative data was filtered out from the training set. Then, the output of the first step is put into the SOM to build prediction model. The results indicate that combination of ANN+SOM outperform the single neural network with respect to accuracy. However, it can be observed that data reduction and filtering in first method (i.e., ANN) leads to loss of samples from the training set.

Apart from the above mentioned discussion and to the best of our knowledge, there is no study which has focused and considered the role of the distance factors in developing the CCP model for TCI. Therefore,this paper presents a novel CCP model based on distance factors to efficiently predict customer churns and also estimate the level of certainty of the classifier's decision in a given TCI dataset. The next section introduces the propose methodology and empirical setup of this study.

METHODOLOGY

In this section, the detailed descriptions of the proposed empirical study. explains the problem statement and provide details of the empirical setup, and evaluation setup, respectively.

The problem statement

The CCP is binary classification problem where all the customers are divided into two possible behaviors: (i) Churn, and (ii) Non-Churn. Further, the churn behavior can be classified into the following subcategories: (a) voluntary customer churn, in which a customer decides to leave the service or even company, and (b) involuntary customer churn, in which the company or service provider decides to terminate a contract with the customer This study addresses the voluntary customer churns due to difficulty in predicting this type of customer churn while it is easier to filter out the involuntary customer churn by simple queries. On the other hand, the literature revealed that existing studies have been published but still there is no agreement on choosing the best approach to handle CCP
problem.

Empirical setup

We designed an empirical study to evaluate the proposed CCP model where we have focused on distance factor using different distance zones (i.e., Upper and Lower zones) in the given TCI datasets. Fig. 1 visualize the overall process of the framework. For this study, we have selected arbitrary four publicly available datasets. The dataset-1 consists of 3333 samples and each sample represent individual customer; whereas, the ratio of churn and non-churn customers is 85.5% and 14.49%, respectively. Similarly, datasets-2, 3 and 4 contain 7043, 18,000 and 100,000 samples, respectively. Further, detail about these datasets is provided in Table 1.

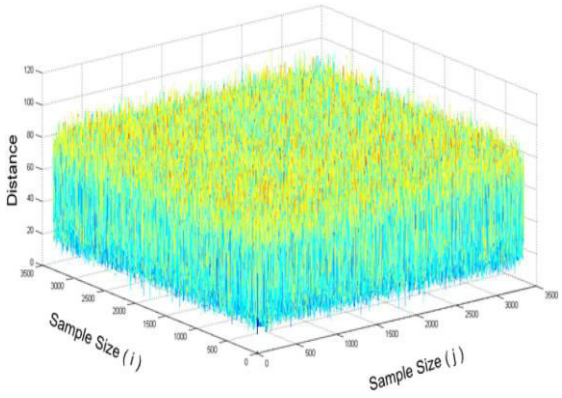


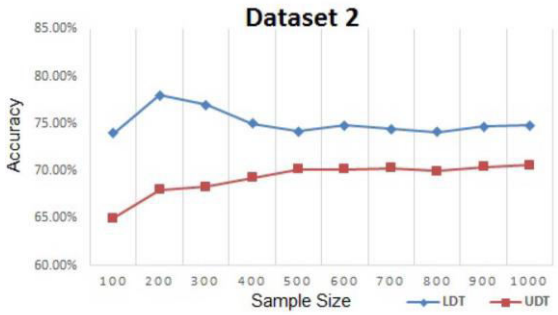
Fig. 2. Shows the distances between one sample against the rest of instances.

Data preprocessing

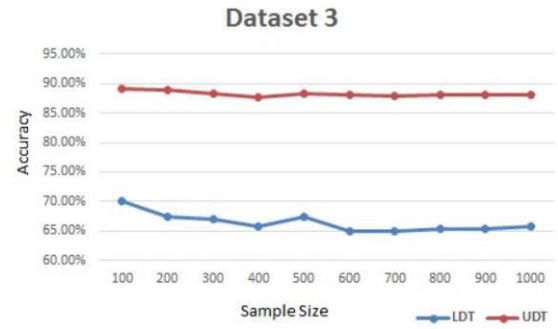
In the preparation step, we have discretize by size, the values that exists in each attribute of the dataset, and then assigned certain labels e.g., Zero to Nine (0–9) possible values, to each discretized group. The discretizing by size leads to selecting the numerical attributes to nominal attributes and grouped them into specific size of bins. We then divide the total number of values in an attribute by size of bin. Ultimately, it produced specific list of values in different number of groups of an attribute. The step by step procedure for data preprocessing and discretization is elaborated as following:

1. Ignore the attributes consisting unique values which represents identity of the sample or descriptive text that serves for informational purposes and does not effect on the models training process.

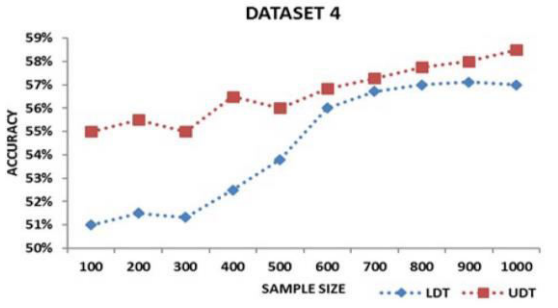
(a)



(b)



(c)



(d)

2. Normalize the categorical values (such as 'yes' or 'no') into 0s and 1s where each value represents the corresponding category, then transformed the 0 and 1 into the same range and assign the same labels which applied for the rest of the attributes.

3. Find the distinct count of each value in every attribute, and also calculate the frequencies of these values in corresponding attributes.

4. Divide the range of values into 10 possible groups and assigned 0–9 label to each group in all the attributes.

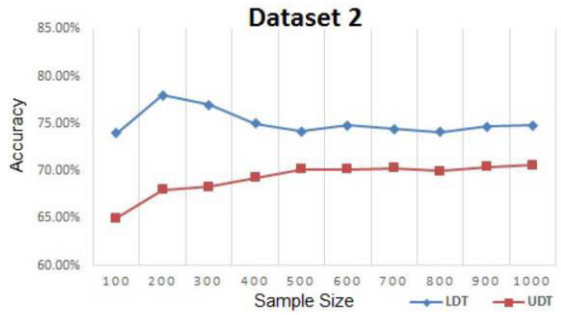
Evaluation setup

In this section, a benchmarking framework is setup to present and evaluate the performance of the proposed study. (i) samples at lower distance side (lower zone), (ii) samples at upper distance side (upper zone), and (iii) samples that are in the middle (dark color) and can be seen as the major part of the samples.

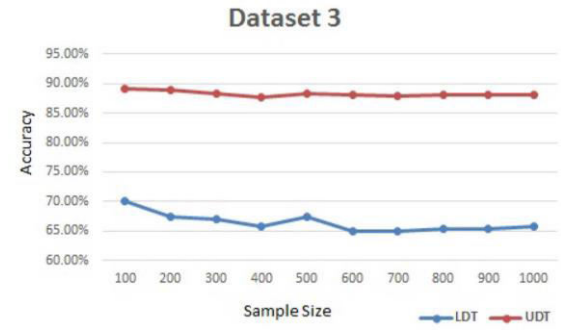
Selection of samples for building the CCP model

The training set is usually used to train the model while test set is used in order to estimate how well the proposed model has been trained (performance evaluation of the model). In this study, we have introduced a novel procedure for training and validation process for finding the expected certainty (i.e., high and low) of the classifier's decision based on the distance factors as well as its impact on the classification performance. We assume the test set is new data where the value of the

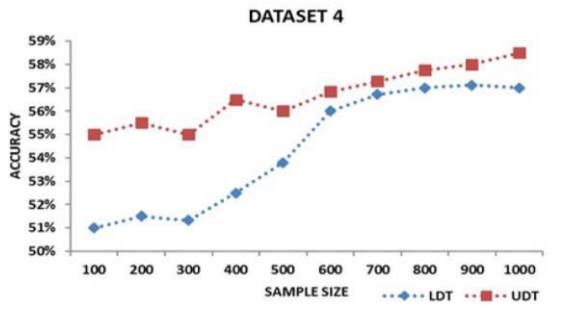
(a)



(b)



(c)



(d)

class label is obtained from the proposed predictive model. We then collected the predictions result from the trained classifier on the inputs from the test sets of both upper and lower zone samples and then compared them to obtain the empirical results of these test sets. This process allows us to evaluate the performance of the proposed model on the given test sets for both UDT and LDT highlighting the level of certainty of the classifier for both. For this purpose, initially, both zones sizes are set to 100 samples where first 100 and last 100 samples are selected for lower and upper zones of the distance between the samples, and then with each iteration the zone's size is increased by adding next 100 samples.

Results and discussion

In this section, we have explored the results of the proposed empirical study and evaluated these results through the state-of-the-art evaluations measures (i.e., precision, recall, f-measure, and accuracy).

The proposed study further demonstrated the benefit of incorporating the distance factor and creating different data zones. Often, the researchers focused on a model which achieved higher accuracy rather than focusing on the level of certainty of the classifier predicting the Churn behavior. This study revealed that the decision maker should equally focus to propose level of certainty to effectively predict the customer churn and non-churn with high certainty as well as with low certainty. Customers identified by classifier with low certainty creates uncertain situation in the TCI because such customers may change their mind and may become churn from non-churn customer and vice versa. We can say that the performance of CCP model under uncertain situation in TCI is inversely proportional to accuracy.

CONCLUSION

The terrific growth of digital data and associated technologies, there is an emerging trend, where industries become rapidly digitized. These technologies are providing great opportunities to identify and resolve diffuse problem of customer churn, particularly in TCI. Through a novel CCP approach, we have extracted insightful level of certainty of classifier decision based on the distance factor and also

categories the customers into different customer groups based on lower zone and the upper zone of distance. Further, we empirically evaluated the impact of the level of certainty of classifier before the classification customers churn and non-churn. . Overall, the proposed study offers two main contributions to the existing literature such as: (i) introduced a novel approach for CCP in TCI based on distance factor, and (ii) revealed the effects of the distance factor in different distance zones (upper and lower zones) to estimate the expected certainty of the classifier decision. It is also investigated that distance factor is strongly co-related with the certainty of the classifier because the customers in lower zone shown uncertain behavior as compared to upper zone (certain behavior). Additionally, the performance of the resulting models was evaluated with four stateof-the-art evaluation measures which gave consistent and robust results.

REFERENCES

Abbasimehr, H. (2011). A Neuro-Fuzzy Classifier for Customer Churn Prediction. International Journal of Computer Applications, 19(8), 35–41.

Ahmed, A. A., & Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. Egyptian Informatics Journal, 18(3), 215–220. http://www.sciencedirect.com/science/article/pii/S1110866517300403http://dx doi.org/10.1016/j.eij.2017.02.002.

Amin, A., Al- Obeidat, F., Shah, B., Tae, M., Khan, C., Durrani, H., & Anwar, S. (2017). Just-in-time customer churn prediction in the telecommunication sector. Journal of Supercomputing. http://dx.doi.org/10.1007/s11227-017-2149-9.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, 237. http://dx.doi.org/10.1016/j.neucom.2016.12.009.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., ... Hussain, A. (2016).