Received 21 March 2019
Revised 31 May 2019
Accepted 19 June 2019

Feature intersection for agent-based customer churn prediction

Sandhya N.
Department of Information Technology,
Cochin University of Science and Technology, Kochi, India
Philip Samuel
Department of Computer Science,
Cochin University of Science and Technology, Kochi, India, and
Mariamma Chacko
Department of Ship Technology,
Cochin University of Science and Technology, Kochi, India

Abstract

Purpose – Telecommunication has a decisive role in the development of technology in the current era. The number of mobile users with multiple SIM cards is increasing every second. Hence, telecommunication is a significant area in which big data technologies are needed. Competition among the telecommunication companies is high due to customer churn. Customer retention in telecom companies is one of the major problems. The paper aims to discuss this issue.
Design/methodology/approach – The authors recommend an Intersection-Randomized Algorithm (IRA) using MapReduce functions to avoid data duplication in the mobile user call data of telecommunication service providers. The authors use the agent-based model (ABM) to predict the complex mobile user behaviour to prevent customer churn with a particular telecommunication service provider.
Findings – The agent-based model increases the prediction accuracy due to the dynamic nature of agents. ABM suggests rules based on mobile user variable features using multiple agents.
Research limitations/implications – The authors have not considered the microscopic behaviour of the customer churn based on complex user behaviour.
Practical implications – This paper shows the effectiveness of the IRA along with the agent-based model to predict the mobile user churn behaviour. The advantage of this proposed model is as follows: the user churn prediction system is straightforward, cost-effective, flexible and distributed with good business profit.
Originality/value – This paper shows the customer churn prediction of complex human behaviour in an effective and flexible manner in a distributed environment using Intersection-Randomized MapReduce Algorithm using agent-based model.

Keywords Redundancy, Hadoop, Distributed, MapReduce, Agent-based, Intersection-randomized

Paper type Research paper

1. Introduction

In the modern era, mobile communication has become an integral part of every human being (Almana *et al.*, 2014). In the telecommunication industry, different customer service providers attempt to increase the market share for a profitable business. Mobile customer retention is considered to be a significant factor for investigation, as it is the centre of interest for developing a profitable relationship with customers. In this telecommunication business, the telecom service providers are competing for customer satisfaction by offering different customer services. Mobile customer retention is directly proportional to mobile user satisfaction (Qureshi *et al.*, 2013). It is costly to entice new mobile users into their service providers than to maintain existing mobile users (Burez and Van den Poel, 2009). To sustain in the telecom industry, mobile service providers introduce different service features. By improving the quality of mobile communication service features, we can enhance customer retention.

Data Technologies and Applications
Vol. 53 No. 3, 2019
pp. 318-332

2514-9288
DOI 10.1108/DTA-03-2019-0043

In the mobile communication industry, mobile SIM portability is an on-going process (Wai-Ho Au *et al.*, 2003). This scenario affects the telecom market in a negative manner (Wai-Ho Au *et al.*, 2003). It is important to avoid such a negative situation and we should find a suitable solution to overcome the problem of customer churn, that is by maintaining the customers before customer attrition (Coussement and Van den Poel, 2008). Different artificial intelligence methods and other statistical methods are adapted to handle this problem (Dasgupta *et al.*, 2008). Usually, telecommunication service providers offer different retention policies as trial and error methods to entice new customers (Bandara *et al.*, 2013). Customer satisfaction is an important criterion that avoids this customer churn problem. In the telecom industry, customer churn causes the loss of revenue generation in the business field (Jiang *et al.*, 2013).

Handling massive datasets in the telecommunication user details involves the application of big data, machine learning and artificial intelligence techniques in the telecommunication industry (Ahmed *et al.*, 2011). Much of the valuable information in the mobile communication industry is concealed in data (Apte and Hong, 1996). Hence, it is possible to examine these data from different perspectives for business growth (Verbeke *et al.*, 2012). Research in the mobile communication industry can be broadly divided into three types: fraudulent detection in the mobile communication; customer churn prediction in telecommunication; and telecommunication network fault detection and segregation (Hung *et al.*, 2006). From the input dataset, all the data features are not required for prediction. Duplicate and irrelevant customer call details are removed and relevant features are selected for accurate prediction (Hadden *et al.*, 2007). Survival of the industries such as telecommunication, banking, medical and other marketing environments depends on a large number of customers. The loss occurring due to customer churn is very serious in the telecom industry (Kim *et al.*, 2012).

Another issue in the telecommunication industry is the large volume of data generation, which is difficult to analyse and process. This challenge is handled in our approach, using MapReduce (Li *et al.*, 2014) with machine learning (Fan *et al.*, 2008) techniques. Excessive volume of an unstructured dataset is processed using MapReduce functions in a distributed environment. To rectify this customer churn situation to overcome the loss in the telecommunication industry, service providers look forward to different statistical methods and machine learning algorithms. The previous history of the customer details has to be analysed using machine learning methods to make some customer predictions (Bandara *et al.*, 2013) and (Fortuny *et al.*, 2013).

Customer churn depends on complex human behaviour. Hence, an intellectual agent-based model would be appropriate to analyse customer churn (Bonabeau, 2001; Brantingham *et al.*, 2005). However, it does not handle redundancy issues. In our approach, big data computation with machine learning techniques (Amin *et al.*, 2014; Han *et al.*, 2012; Farquad *et al.*, 2014) was found to be an effective method for identifying customer churn prediction. Redundant data from the customer details dataset are removed, thereby reducing the map partitions. In the map phase, the chunks of the data partition occur. Duplicate removal reduces the number of maps as well as Reduce functions.

This paper is organised as follows. Section 2 discusses related work; Section 3 discusses the proposed methodology; Section 4 describes result analysis; and Section 5 presents the conclusion and future work of the paper.

2. Related works

Qureshi *et al.* (2013) describe certain data mining techniques for churn prediction. Customer DNA dataset is used for the prediction of 1,06,000 users traffic data. Customer usage has been analysed for three months. The re-sampling technique is adopted to solve different class imbalance problems. The statistical methods implemented are regression analysis

method, artificial neural networks, K-means, decision tree (Bin *et al.*, 2007), CHAID, exhaustive CHAID, CART and QUEST. During the result analysis and discussion, all these methods are compared on the basis of different measures such as precision, recall and *F*-measure. The potential churners have been identified using the above methods, and exhaustive CHAID was found to be the most accurate. The prediction accuracy of this method was found to be 75.4 per cent.

Jiang *et al.* proposed an implementation using WEKA tool that handles data mining applications for churn prediction analysis. Customer classification was done not as a churner or likely, but each customer was classified as a potential churner or non-churner. The framework discussed was based on the knowledge discovery data process. In total, three different datasets, small, medium and large, with varying attributes were considered. The efficiency and performance of the decision tree and logistic regression techniques were compared. The accuracy achieved with decision tree was much greater than logistic regression. Using logistic regression analysis and decision tree analysis, the prediction accuracies were 86 and 91.4 per cent, respectively.

Idris *et al.* (2012) proposed an approach based on genetic programming together with Adaboost to solve the telecommunication churn problem. The Adaboost way of boosting refers to the process of developing more than one programmes in a class and computing the total summation of outputs of the genetic programming problems for the ultimate prediction. Orange Telecom and Cell2cell telecommunication datasets were used for testing. Finally, 89 per cent of accuracy was obtained for Cell2Cell and 63 per cent accuracy was obtained for Orange mobile communication. Zhang *et al.* suggested the accuracy of the mobile customer churn prediction on the reaction of different attributes of the network. Depending on the behaviour of user phone calls, the attributes of the network considered are communication among the customers and social networking topologies. Comparisons of different networking attribute models are discussed. Differentiation of traditional attribute, network attribute and various combinations of attributes together is done. It states that the customer churn prediction accuracy has greater improvement while connecting different networks attributes.

Michael Prez *et al.* in the paper proposed telecommunication users with call drops or other customer service failure, continuous telecommunication service collapse and the success rate of restoration. Telecommunication multi-system operator dataset over ten months duration is used for the case study and the implantation purpose. Depending on the plan duration of the customer with service failure, it has been considered for further statistical analysis and research. Another approach is based on 30-day period customer churn due to service failure. Using telecommunication billing information, dataset is evaluated. They concluded that customers using real-time streaming voice, video and data probably left the telecommunication service provider due to service failure.

Ahmed *et al.* (2011) proposed a method using data mining and agent-based method. For implementation and testing, the dataset is used other than random values. The agent-based method considered each parameter as a real-world entity. The data mining method used for analysis is a decision tree. This method analyses the patterns used by drivers, which describe the driver behaviour. During the car journey, procating of the child in the car can be predicted in a good manner. Dataset is taken from various road surveys. The simulation and the result analysis have been done using data mining and the agent-based model.

Idris *et al.* (2012) addressed Cell2Cell telecommunication companies dataset with 40,000 entities. Feature classifiers are adopted in the pre-processing phase (Yang and Moody, 2000). Classifiers such as minimum redundancy and maximum relevance (mRMR), Fisher's ratio and *F*-score methods are used. Linear search technique is adopted for the improved performance.

Accuracy among the classifiers is 76.2 per cent for mRMR, 65.25 per cent for *F*-score ratio and 69.1 per cent for Fisher's ratio. Arroyo *et al.* (2010) suggested exogenous designing of the system with the agent-based model. In this system, data mining methods and statistical methods are described. Different phases of the model, such as initialising, designing, validating, are involved in the system study. Model simulation is done using data mining methods. The mentat dataset model is used for the analysis purpose.

Petkovski *et al.* (2016) proposed churn prediction analysis with Macedonia telecommunication service with 22,461 customers dataset as a case study. The methods adopted for classification are C4.5, logistic regression and K-nearest neighbour. Compared to these three methods, logistic regression has the highest accuracy with 94.3 per cent. The disadvantage of this paper is the larger execution time and the usage of more memory resources. Karanovic *et al.* (2018) focussed on churn prediction of telecom customers of a company called Orange. The pre-processing methods are performed to handle missing data from the dataset. Convolution neural network is adopted for classification. Limitation is that feature selection and network complexity are not found out. The average execution time is also longer. Existing systems can be improved by adding better customer churn prediction. Further, the existing systems do not handle input data cleaning or pre-processing redundancy techniques.

3. Proposed methodology

In this paper, during data pre-processing phase, Intersection-Randomized Algorithm (IRA) is applied to clean the unwanted data from the input dataset. Telecommunication customer churn prediction is one of the real challenges because of customer behaviour. Hence, it is considered for our case study. Input dataset is telecommunication customer details of an anonymous service provider from 2015 to 2017 with 18 months customer call details [DATASET]. The customer call dataset is the crucial input dataset that is used for customer retention. The dataset we obtained is large and unstructured. Hence, we implement the proposed Intersection-Randomized Algorithm on MapReduce platform. Telecommunication customer retention is predicted using the agent-based model (ABM) (Sandhya *et al.*, 2018).

The unstructured user dataset should be cleaned by removing unwanted data to obtain better results in the application execution. All the entities in the dataset are not required for prediction analysis. Some entities are not complete and some of the entities are having "null" values. Incomplete or null value entity fields are removed. So data duplication has to be removed and data integration is required for the better performance of the system. Unstructured redundant input data are given as input. This dataset is cleaned by removing redundant unwanted data using the Intersection-Randomised Algorithm. The irredundant dataset, thus obtained, is used to train the agent-based customer retention prediction model (Figure 1).

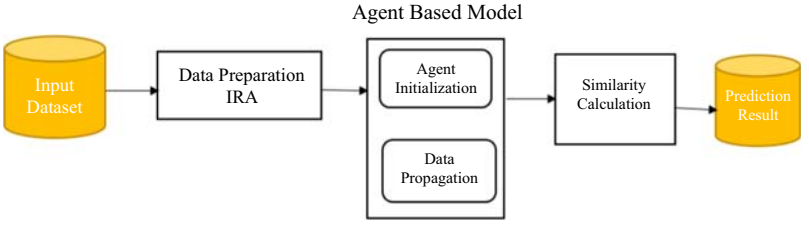

Figure 1.
System design for IRA with ABM

3.1 MapReduce technology

MapReduce technology has made large complex data processing easier and more efficient. This technology is applied for batch processing of large volumes of data. MapReduce is a parallel data processing approach for execution on a computer cluster. The performance of the MapReduce jobs depends on the input split sizes of the file chunk size. We choose *InputFileFormat* class in which the entire file would be processed by a single map task and a parameter would be set, indicating that the input files are splittable, which means that we can run multiple map tasks against different parts of a single input file. For example, we need to process a 10-GB file in a single map task. We can set the split size to 1-GB file chunks to each node in the cluster and then we can process with ten map tasks against the single 10-GB file, with each map task processing a unique 1-GB range of the file data.

In general, depending on the volume of the dataset and the processing ability, the input data can be divided into n chunks. These chunks are passed to the Map function. All the chunks are processed simultaneously at the same time, which contains the parallel processing of data. Then, sorting and shuffling occur, which leads to a collection of similar patterns.

Finally, reducers combine them to a consolidated output. The input dataset volume depends on the scalability of the algorithm, and we can increase the number of processing units. MapReduce programming model defines two user-defined functions Map and Reduce. Input data have been divided into chunks. Each chunk generates a Map function that performs sorting and filtering of the input data. The input data to Map function is in the form of a key/value pair. The Map function in the MapReduce job is fed with data stored on the distributed file system, which are split across nodes. The Map tasks are started on the compute nodes and Map function is applied to each key/value pair, thus having intermediate key/value pairs as output. These intermediate key/value pairs are stored in the local file system and sorted by the keys.

The Reduce function performs the addition operation. The Reduce function takes the intermediate key/value pair as the input. A Reduce function is applied to all values grouped for one key and, in turn, generates key/value pairs. In the distributed file, the reducer function's key/value is written.

3.2 Intersection-randomized algorithm

The proposed model implements Intersection-Randomized Algorithm with MapReduce functions. The Clean dataset with reliable data and without problematic error is the data pre-processing requirement. Redundancy removal technique helps to get a well-defined dataset for customer retention feature prediction. During data processing, 24 features of the customer call dataset [3] are taken for analysis. The algorithm first compares each "feature_variable" of customer dataset between all other customers in a parallel manner.

In each iteration, n feature_variables are compared for each user from 1 to $n+1$. If the same feature_variable exists for one "userid", then the second entry is considered to be replicated and value "0" is assigned for the replicated entry. If "null" value is identified with any of the feature fields, then it will assign value "0". Each customer with all these possibilities is taken into consideration and included in the training dataset for final analysis. The key–value pairs with value "1" will be taken in the training dataset for customer retention feature description. As shown in Figure 2. Feature_variable1 through feature_variable24 of user call details dataset has been compared and all the probabilities for redundant, unclean, or error data between the users are tested. The final collection of a dataset with the clean and reliable dataset is obtained for predicting customer retention features.

Intersection-Randomized MapReduce Algorithm.
Algorithm 1: Intersection-Randomized MapReduce Algorithm.
// value:- var_f1: Variable feature1;
// key:- UID1: first UserID
Output: key value pair: key UserID and value Boolean number representing removal entity status
map (String UID1, String var_f1)
FUNCTION cust_retention (UID1, var_f1)
STATIC FUNCTION cust_retention
(MISSING, INVALID, UID, var_f)};
for each UserID UID1in value Variable
feature
EmitIntermediate (UID1, "1")
if (var_f1.MISSING || 0)
if (var_f1 || (UID1 = = INVALID))
EmitIntermediate UID1
Algorithm: Reduce function
// key: - UserID
// value: - totalvar_f: list of features
reduce (String UID, Iterator values)
for each totalvar_f in values
Emit(AsString(output)

Algorithm 1 shows Map function and Reduce function of Intersection-Randomization Algorithm. Each feature is compared with every other feature of all customers. This algorithm is implemented using Map and Reduce functions. The partitioning of input dataset into chunks to perform Map function occurs in the Mapping side in parallel. In the Reduce function, each "userid" and each variable feature are gathered (Cust_Retention(UID, [var_f1,var_f2..var_fn])) on the basis of their keys. The Reduce function collects all the sorted Map functions, finds the matching feature variables and returns key–value pairs. Each of the customers is compared in the same manner in parallel. This algorithm returns value "0" for the duplicate feature field and returns value "1" for the field for further processing. Customers with feature value "1" are taken for customer churn behaviour prediction.

3.3 Agent-based model

Behaviour of human species is complex. Agents are used to represent the behavioural model. Multi-agents are useful in understanding different perspectives of human behaviour. The ABM acts as actions and communications between agents for collective entities such as organisations. The ABM is used to predict some complex phenomena, such as human behaviour that depends on some situations, or it affects the system or group as a whole to deal with it. Analysing the customer call details for a period of time and depending on their

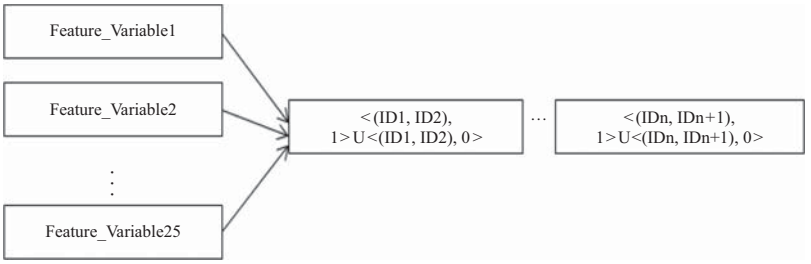

Figure 2.
Intersection randomization

behaviour customer retention features can be suggested. It will be easy and accurate to communicate customer perspectives through agent-based activities.

In today's technology changes, artificial intelligence has a crucial role. Softwares designed on the basis of agents are widely used nowadays. Complexity increases during the analysis phase in traditional systems. In traditional systems, it is difficult to find out the interdependencies between the feature, whereas it is easier in the ABM. ABM is useful for complicated real-time system modelling. In traditional systems, analysis will be difficult when multiple entities communicate with one another. Simulations of microsystem design models are possible due to large-scale data handling power of big data frameworks with ABM.

As shown in Figure 3, agent types, objects and attributes are identified, and they are represented as entities (E1, E2, etc.). Agents' active environments are defined with some rules and conditions. Functions are defined for agent–agent interactions and entity–agent interactions while implementing ABM.

ABM can be computationally connected with real-world objects. Each real entity can be represented mathematically or analysed statistically. These can be converted to computer programming languages. Depending on the real entity, the behaviour of the agents varies. Agents should be in active mode and to make decisions, the properties of agents are describes as:

- agents are capable of distinguishing the working environment, and they can react to internal changes;
- we can define agents on the basis of the objectives of the application;
- the capacity of the agent to communicate with each other at sudden requirements;
- the agents capacity for failure recovery and to achieve different methods to satisfy their objectives; and
- potential of the agents without direct involvement by human beings, that is controlling internal state space and behaviour by artificial intelligence.

The customer variable features depend on customer behaviour. This customer behaviour varies in different situations and it can be complex. To reduce this complexity, we design an ABM. Agents are the functional units of this model. The positive side of this model is that ABM is flexible. Agents can be increased or decreased depending on the features of input dataset. The addition of the agents to the agent-based model is simple. It is also easy to change the aggregation of the agents and to subgroup them depending on the description and behaviour. The ABM improves the performance of the system.

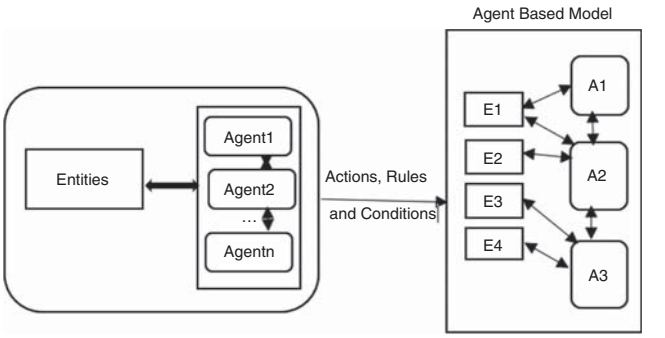

Figure 3.
The architecture of agent-based model

The variable features are agents in the agent-based model. We have designed a model to suggest the customer retention variable feature on the basis of the customer dataset shown in Table I.

Nowadays, it is easier to migrate to different networks. This behaviour of customers affects user retention. Agents are designed as churn agents and retention agents. Both agents are again subdivided into some subgroups, as shown in Table II.

Agent interactions occur dynamically. The coding rules of the agents are as follows: churn agents are considered as CH_A and retention agents as R_A. Thus, state space of the agents is defined below (State Space of Agents Features). State Space of Agents Features:

(neither CH_A or R_A) as
nCoR
(only CH_A) as oC
(only R_A) as oR.

As shown above (State Space of Agents Features), three state spaces of the agent interactions have been defined. The first state space "nCoR" cannot be defined between the churn agent or retention agent. The state "oc" is described as the only churn agent. The explicit behavioural feature interactions describe the churn agent. The "Or" state space is only retention agents concerning the interaction between the variable features.

No.	Variable feature	Description	No.	Variable feature	Description
1	A_Len	Account Length	14	Day _Charg	Day time call charges
2	MO_Churn	Monthly Customer Churn	15	Eve_ Calls	Number of calls in the evening
3	D_Mins	Day time call duration of the customer	16	Eve_ Charge	Evening call charges
4	E _Mins	Evening call duration of the customer	17	Night _Calls	Number of calls at night
5	N_ Mins	Night call duration of the customer	18	Night_ Charge	Call charges at night
6	Int_ Mins	International call duration of the customer	19	Intl _Calls	Number of international calls
7	CS_Calls	Customer service calls	20	Int_ Charge	Number of customer service calls
8	Total_Churn	Customer churn during last one year	21	State	Customers state code
9	Int _Plan	Customers have taken International plan	22	A _Code	The area code of the customer
10	VMail _Plan	Calling plans opted	23	Mon_Chg	Monthly charges of the customer
11	Day_ Calls	Number of calls at day time	24	Contract	Plan duration of the customer
12	St_Vd	Online data streaming videos	25	V_Chat	Online voice chatting
13	Network_Q	Network quality of telephonic calls.	26	D_Plans	Online data plans

Table I. Variable features of customer call dataset

Agents	Subgroups
CALL_DURATION (CD)	NIGHT_CALL_DURATION(NCD), DAY_CALL_DURATION (DCD), EVENING_CALL_DURATION (ECD)
CALL_CHARGES (CC)	MONTHLY_CALL_CHARGES(MCC), INTERNATIONAL_CALL (IC)
CUSTOMER_SERVICE_CALL (CSC)	PLAN_BASED(PB), INTERNET_BASED (IB), NETWORK_BASED (NB)
SIM_EXPIRY (SE)	PLAN_DURATION (PD)
NETWORK_PLANS (NP)	POSTPAID (PT), PREPAID(PP), INTERNATIONAL_PLANS (IP), DATA_PLAN(DP)
NETWORK_FALIURE (NF)	VIDEO_STREAMING (VDS), VOICE_CHATS (VCT)

Table II. Agents and subgroups

These three state spaces are described for this system. The system can change the status at any point of time, which makes the system show complexity depending on customer variable features. The relationships between the features are defined for the proposed model on the basis of the state-space features. The rules of the agent-based model prediction for customer retention are defined as follows:

- Rule 1: If call time duration, that is day call duration, increases (CD_DCD) for a user, then call charge agent with monthly call charge (CC_MCC) increases or CD_NCD increases then CC_MCC increases, the probability of only churn increases.
- Rule 2: If NP_PT to NP_PP or vice versa change occurs, then CSC enquiry should accompany the customer; otherwise, only churn occurs.
- Rule 3: If SIM_EXPIRY (SE) then intent to CSC_PB applicable, then chances of *neither churn or retention* increases.
- Rule 4: If IC rises, then reduced NP_IP will increase the chance of the occurrence of only retention.
- Rule 6: If CSC continuously enquired the customer based on CSC_PB, CSC_IB and CSC_NB with customer satisfaction during SE, then there is a high chance of the occurrence of only retention.
- Rule 7: If NF_VDS or NF_VCT, then there is a high chance of the occurrence of only churn.
- Rule 8: If NF_DCD or NF_ECD or NF_NCD, then there is a high chance of the occurrence of only churn.
- Rule 9: If NP_DP rate is reduced and NP_DP options are high, then there is a high chance of the occurrence of only retention.
- Rule 10: If NP_DP rate is high and NP_DP options are low, then there is a high chance of the occurrence of only churn.

In this system, the agent interaction is more, that is agents communicate with every other agent. This is called global communication, which increases the system performance to a high level and also the interactions occur in a certain order. The communications between the agents happen in a parallel manner. Hence, this model achieves parallelism.

According to the state-space features and rules of the ABM, affinity propagation is measured. Similarity measure of potential input data points is taken for the calculation of clusters in the affinity propagation method. If the similarity measure is nill, then that feature does not belong to that telecommunication feature cluster set. The responsive agents are represented as responsive matrix $R(i, t)$, that is x_t agents interact with x_i and available agents as $A(i, t)$, that is x_i available at an object x_t. Responsive agents of each dataset are found in (2) and distance measure is obtained as follows:

$$d(i,t) = -\left\|x_i - x_t\right\|^2. \tag{1}$$

$$R(i,t) \leftarrow d(i,t) - \max_{k' \neq k}\left\{A(i,t') + d(i,t')\right\}. \tag{2}$$

The proposed prediction agent-based model is implemented using MapReduce function with these defined agent-based rules. The data set [DATASET] refers to 18-month customer call details that have been used for training customer retention agent-based model, and this system predicts customer retention features best suited for the specific customers based on the variable features.

4. Result analysis and discussion

Customer retention prediction using Intersection-Randomized Algorithm with agent-based model illustrates the customer retention requirements. Depending on the customer behaviour, churn out reasons are characterised as variable features that can be modelled as agents in the system. As shown in Figure 4, telecommunication characteristics such as Network_quality (88 per cent) are the highest for customer retention requirement. Network_Plans has 79 per cent effect on the customers. International_plans has 35 per cent effect. Monthly_cal_rate has 74 per cent effect. Customer_service_calls is required for 22 per cent of customers. Day_call_charge is needed for 40 per cent of customers. Night_call_charge is required for 22 per cent customers, International_call_rate is needed for 18 per cent customers and Data_plan is required for 63 per cent of customers.

Table III shows percentage of customers on the basis of their agent-based behaviour. According to the agent-based rule prediction, 50,000 customers were analysed on the basis of the ABM: 70 per cent have oR behaviour, 20 per cent have oC behaviour and 10 per cent have nCoR behaviour. Analysing the data of 75,000 customers, we found that 63 per cent have oR behaviour, 26 per cent have oC behaviour and 11 per cent have nCoR behaviour. Data set with 1,00,000 customers is analysed, and it is seen that 60 per cent have oR behaviour, 34 per cent have oC behaviour and 6 per cent of the customers have nCoR behaviour. Hence, it is seen that customer behaviour depends on the variable features of telecommunication. Customer retention behaviour can be maintained on the basis of the variable feature rate.

The average similarity score by using affinity propagation is calculated by different customer data sets. Different data set observations are evaluated and are shown in Figure 5. The average similarity score is 0.808 for data set1, with 6 months of customer telecommunication details. The average similarity score is 0.87 for data set2 with 12-month customer telecommunication details, and an average similarity score is 0.89 for data set3 with 18-month customer telecommunication details. We can find the average similarity score value for each of the different data sets. From our results, it can be observed that the average similarity score is increasing for every six months. This short-term prediction with

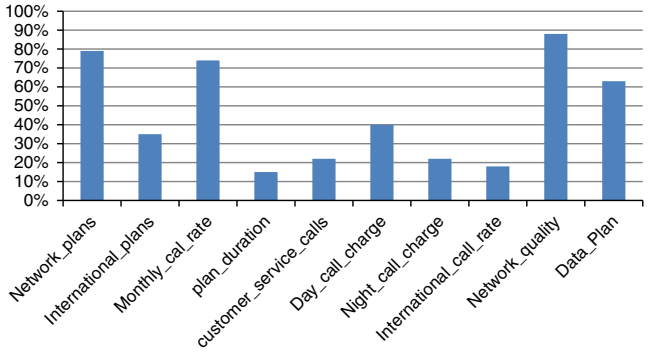

Figure 4.
Customer retention requirements

No. of customers	oR behaviour (%)	oC behaviour (%)	nCoR behaviour (%)
50,000	70	20	10
75,000	63	26	11
100,000	60	34	6

Table III.
Customers versus agent-based behaviour

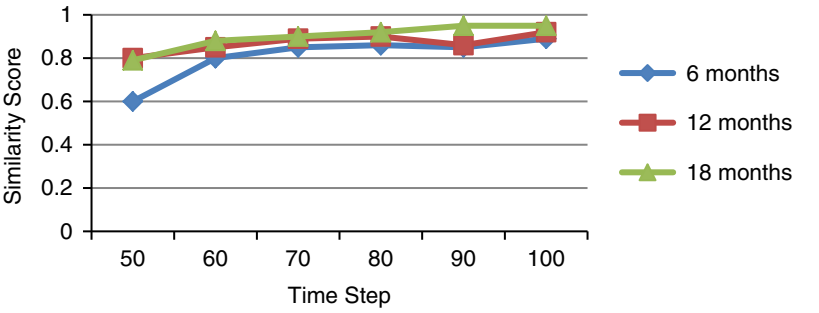

Figure 5.
Similarity score versus time step

6-month difference helps to improve customer retention to some extent. Hence, our method helps the company to take managerial decisions to avoid customer churn.

Prediction accuracy of the model is defined with the following terms: number of positive churn users, number of positive non-churn users, number of negative churn user, number of negative non-churn users. In our ABM, we denote ABM_P for precision, ABM_R for recall and ABM_A for accuracy:

$$ABM\_P = \frac{\mathrm{NPC}}{\mathrm{NPC}+\mathrm{NNC}}, \tag{3}$$

$$ABM\_R = \frac{\mathrm{NPC}}{\mathrm{NPC}+\mathrm{NNNC}}, \tag{4}$$

$$\mathrm{Pr}ediction\_Accuracy = \frac{\mathrm{NPC}+\mathrm{NPNC}}{\mathrm{NPC}+\mathrm{NPNC}+\mathrm{NNC}+\mathrm{NNNC}}. \tag{5}$$

Table IV shows precision, recall and accuracy obtained from Equations (3)–(5). As shown in Table IV, ABM_A has an average accuracy of 95.8 per cent. Thus, our proposed ABM shows good predictive power. The result indicates that adding good features and selecting better classifier may enhance predictive modelling more significantly for telecommunication service providers faced with competition.

As shown in Table V, the time taken for programme execution with redundant data from the data set will be more when compared with IRA model implementation. This is tested

Table IV.
Precision, recall and accuracy for ABM

Users	ABM_P	ABM_R	ABM_A
50,000	0.73	0.75	0.943
100,000	0.72	0.69	0.96
150,000	0.788	0.743	0.97
200,000	0.786	0.846	0.973

Table V.
Average time taken to execute the application

Users	Average time taken without IRA in seconds	The average time taken with IRA
50,000	90	50
75,000	160	80
100,000	220	135

using different customer data sets and shown above. The performance is high for the churn prediction using IRA model. As shown in Table V, the average time taken to execute the application with the IRA is lesser than without IRA. In total, three different data sets with 50,000, 75,000 and 100,000 customers are considered. With using the IRA, the time taken is 50 sec, 80 sec and 135 sec, respectively. Without removing the data duplication, the average time taken is 90 sec, 160 sec and 220 sec, respectively. The time taken for execution without IRA increases almost 50 per cent than with the IRA model. The performance of the application increases with IRA model implementation and the time taken to execute the application decreases.

5. Conclusion and future work

This paper focusses on telecommunication customer retention prediction based on different human behaviour. One of the challenging situations in the telecommunication industry is user churn. Customer satisfaction of the service provider depends on the product service quality, product price, and different personal aspects. Our proposed prediction model helps in the customer retention prediction using IRA using the MapReduce method with the agent-based model. Duplication of data in a large volume of input data set is removed by Intersection-Randomized Algorithm using Map function and Reduce function. Customer retention of this data set is analysed in a flexible fast agent-based variable feature modelling. The reliability of IRA with ABM is evaluated in terms of affinity propagation and prediction accuracy with respect to flexible behavioural changes. We make rules for agents to make them competent for learning from previous experience. Depending on the state space and the defined rules, agents can adapt to the behavioural changes.

Using IRA with MapReduce processing and using ABM enhance the prediction accuracy by removing duplicates from the data set. Hence, this prediction system is simple, flexible and distributed. The proposed system is flexible because of the dynamic properties of agents, which is designed using simple global rules and state-space features. The performance accuracy of the system increases because of the enhanced macroscopic behaviour of the agents. Our prediction is fast because of the distributed environment of MapReduce functions.

References

Ahmed, S., Kobti, Z. and Kent, R.D. (2011), "Predictive data mining driven architecture to guide car seat model parameter initialization", in Watada J., Phillips-Wren G., Jain, L.C. and Howlett, R.J. (Eds), *Intelligent Decision Technologies. Smart Innovation, Systems and Technologies*, Vol. 10, Springer, Berlin, Heidelberg.

Almana, A.M., Aksoy, M.S. and Alzahrani, R. (2014), "A survey on data mining techniques in customer churn analysis for telecom industry", *Journal of Engineering Research and Applications*, Vol. 4 No. 5, pp. 165-171.

Amin, A., Khan, C., Ali, I. and Anwar, S. (2014), "Customer churn prediction in telecommunication industry: with and without counter-example", *Nature-Inspired Computation and Machine Learning*, Springer, pp. 206-218, available at: http://dx.doi.org/10.1007/978-3-319-13650-9_19

Apte, C. and Hong, S.J. (1996), "Predicting equity returns from securities data with minimal rule generation", *In Advances in Knowledge Discovery and Data Mining*, pp. 541-560.

Arroyo, J., Hassan, S., Gutiérrez, C. and Pavón, J. (2010), "Re-thinking simulation: a methodological approach for the application of data mining in agent-based modelling", *Computational & Mathematical Organization Theory*, Vol. 16, pp. 416-435, doi: 10.1007/s10588-010-9078-y.

Bandara, W.M.C.A., Perera, S. and Alahakoon, D. (2013), "Churn prediction methodologies in the telecommunications sector: a survey", *2013 International Conference on Advances in ICT for Emerging Regions (ICTer), 2013*, pp. 172-176.

Bin, L., Peiji, S. and Juan, L. (2007), "Customer churn prediction based on the decision tree in personal handyphone system service", *2007 International Conference on Service Systems and Service Management, IEEE*, pp. 1-5, available at: http://dx.doi.org/10.1109/ICSSSM.2007.4280145

Bonabeau, E. (2001), "Agent-based modeling: methods and techniques for simulating human systems", *Proceedings of National Academy of Sciences*, Vol. 99 No. 3, pp. 7280-7287.

Brantingham, P.L., Glässer, U., Kinney, B., Singh, K. and Vajihollahi, M. (2005), "Modeling urban crime patterns. Viewing multi-agent systems as abstract state machines", *Proceedings ASM'05, Vol. 3*, pp. 101-117.

Burez, J. and Van den Poel, D. (2009), "Handling class imbalance in customer churn prediction", *Expert Systems with Applications*, Vol. 36 No. 3, pp. 4626-4636.

Coussement, K. and Van den Poel, D. (2008), "Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques", *Expert Systems with Applications*, Vol. 34 No. 1, pp. 313-327.

[DATASET]. available at: www.kaggle.com/datasets

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. and Joshi, A. (2008), "Social ties and their relevance to churn in mobile telecom networks", *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, pp. 668-677, doi: 10.1145/1353343.1353424.

Fan, R.E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J. (2008), "LIBLINEAR: a library for large linear classification", *Journal of Machine Learning Research*, Vol. 9, pp. 1871-1874.

Farquad, M., Ravi, V. and Raju, S.B. (2014), "Churn prediction using comprehensible support vector machine: an analytical CRM application", *Applied Soft Computing*, Vol. 19 No. 1, pp. 31-40.

Fortuny, E.J., de, Martens, D. and Provost, F. (2013), "Predictive modeling with big data: is bigger really better?", *Big Data*, Vol. 1 No. 4, pp. 215-226.

Hadden, J., Tiwari, A., Roy, R. and Ruta, D. (2007), "Computer assisted customer churn management: State-of-the-art and future trends", *Computers & Operations Research*, Vol. 34 No. 10, pp. 2902-2917.

Han, S.H., Lu, S.X. and Leung, S.C. (2012), "Segmentation of telecom customers based on customer value by decision tree model", *Expert Systems with Applications*, Vol. 39 No. 4, pp. 3964-3973, available at: http://dx.doi.org/10.1016/j.eswa.2011.09.034

Hung, S.Y., Yen, D.C. and Wang, H.Y. (2006), "Applying data mining to telecom churn management", *Expert Systems with Applications*, Vol. 31 No. 3, pp. 515-524.

Idris, A., Khan, A. and Lee, Y.S. (2012), "Genetic programming and adaboosting based churn prediction for telecom", *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1328-1332.

Jiang, S., Fiore, G.A., Yang, Y., Ferreira, J. Jr, Frazzoli, E. and González, M.C. (2013), "A review of urban computing for mobile phone traces: current methods, challenges and opportunities", *In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, UrbComp '13, ACM Press, Chicago, IL, 11 -14 August*, p. 1-9.

Karanovic, M., Popovac, M., Sladojevic, S., Arsenovic, M. and Stefanović, D. (2018), "Telecommunication services churn prediction – deep learning approach", *2018 26th Telecommunications Forum (TELFOR)*, Belgrade, pp. 420-425, doi: 10.1109/TELFOR.2018.8612067.

Kim, N., Jung, K.-H., Kim, Y.S. and Lee, J. (2012), "Uniformly subsampled ensemble (use) for churn management: theory and implementation", *Expert Systems with Applications*, Vol. 39 No. 15, pp. 11839-11845.

Li, F., Ooi, B.C., Tamer Ozsu, M. and Wu, S. (2014), "Distributed data management using MapReduce", *ACM Computing Surveys (CSUR)*, Vol. 46 No. 3, 42pp.

Petkovski, J.A., Risteska Stojkoska, B., Trivodaliev, K. and Kalajdziski, S. (2016), "Analysis of churn prediction: a case study on telecommunication services in Macedonia", pp. 1-4, 10.1109/TELFOR.2016.7818903.

Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A. and Rehman, A. (2013), "Telecommunication subscribers' churn prediction model using machine learning", *Eighth International Conference on Digital Information Management (ICDIM 2013)*, pp. 131-136.

Sandhya, N., Philip, S. and Mariamma, C. (2018), *Randomized Agent Based Model for Mobile Customer Retention Behaviour Prediction*, Springer Innovations in Communications and Computing, Cochin University of Science & Technology, Kerala.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and Baesens, B. (2012), "New insights into churn prediction in the telecommunication sector: a profit driven data mining approach", *European Journal of Operational Research*, Vol. 218 No. 1, pp. 211-229.

Wai-Ho Au, K.C.C., Chan and Xin, Y. (2003), "A novel evolutionary data mining algorithm with applications to churn prediction", *IEEE Transactions on Evolutionary Computation*, Vol. 7 No. 6, pp. 532-545.

Yang, H.H. and Moody, J. (2000), "Data visualization and feature selection: new algorithms for non-Gaussian data", *Advances in Neural Information Processing Systems*, Vol. 12, pp. 687-693.

Further reading

Baqueiro, O., Wang, Y.J., Mcburney, P. and Coenen, F. (2009), "Integrating data mining and agent-based modeling and simulation", in Perner, P. (Ed.), *Advances in Data Mining: Applications and Theoretical Aspects*, ICDM 2009, Lecture Notes in Computer Science, Vol. 5633, Springer, Berlin, Heidelberg.

Bhatotia, P., Wieder, A., Rodrigues, R., Acar, U.A. and Pasquin, I.R. (2011), "Incoop: MapReduce for incremental computations", *In Proceedings of the 2nd ACM Symposium on Cloud Computing (SOCC '11), ACM, New York, NY, Article No. 7*, 14pp, available at: https://doi.org/10.1145/2038916.2038923

Chu, C.T., Kim, S.K., Lin, Y.A., Yu, Y., Bradski, G.R., Ng, A.Y. and Olukotun, K. (2006), "Map- Reduce for machine learning on multicore", *NIPS '06*, MIT Press, pp. 281-288.

Dahiya, K. and Bhatia, S. (2015), "Customer churn analysis in telecom industry", *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, pp. 1-6.

Folcik, V. and Orosz, C. (2006), *An Agent-Based Model Demonstrates That The Immune System Behaves Like A Complex System And A Scale-Free Network*, SwarmFest, University of Notre Dame, South Bend, IN.

Frey, B.J. and Dueck, D. (2007), "Clustering by passing messages between data points", *Science*, Vol. 315 No. 5814, pp. 972-976, doi: 10.1126/science.1136800.

Guan, R., Shi, X., Marchese, M., Yang, C. and Liang, Y. (2011), "Text clustering with seeds affinity propagation", *IEEE Transactions on Knowledge & Data Engineering*, Vol. 23 No. 4, pp. 627-637, doi: 10.1109/tkde.2010.144.

Hassouna, M. and Arzoky, M. (2011), "Agent-Based modelling and simulation: toward a new model of customer retention in the mobile market", *Proceedings of the Summer Computer Simulation Conference*, pp. 30-35.

Idris, A. and Khan, A. (2012), "Customer churn prediction for telecommunication: employing various features selection techniques and tree based ensemble classifiers", *2012 15th International on Multitopic Conference (INMIC)*, pp. 23-27.

Liu, D.S. and Fan, S.J. (2014), "A modified decision tree algorithm based on genetic algorithm for mobile user classification problem", *The Scientific World Journal*, No. 1, 11pp, available at: http://dx.doi.org/10.1155/2014/468324

Macal, C. and North, M.J. (2010), "Tutorial on agent-based modelling and simulation", *Journal of Simultion*, Vol. 4, pp. 151-162, available at: https://doi.org/10.1057/jos.2010.3

North, M.J. and Macal, C.M. (2007), *Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation*, Oxford University Press, Oxford.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016), "Why should I trust you?: Explaining the predictions of any classifier", *Proceedings of ACMSIGKDD International Conference on Knowledge Discovery, Data Mining*, pp. 1135-1144.

Wooldridge, M. (2002), *An Introduction to MultiAgent Systems*, John Wiley & Sons, Chichester.

Zhang, X., Liu, Z., Yang, X., Shi, W. and Wang, Q. (2010), "Predicting customer churn by integrating the effect of the customer contact network", *2010 IEEE International Conference on Service Operations and Logistics and Informatics (SOLI)*, pp. 392-397.

About the authors

Sandhya N. is Researcher in Information Technology Division, School of Engineering, Cochin University of Science and Technology (CUSAT). She holds a Master Degree (MTech) in Computational Engineering and Networking from Amrita Viswa Vidya Peetham (2009–2011). She has three years of experience in teaching and four years of experience in Research and Development in IT Company. She has published a paper on 2012 7th International Conference on Computer Science and Education (ICCSE). Her research interest includes Big Data Analytics, Machine Learning and Artificial Intelligence. Sandhya N. is the corresponding author and can be contacted at: nairsands@gmail.com

Dr Philip Samuel is Professor in Department of Computer Science, Cochin University of Science and Technology (CUSAT). He holds Master's Degree (MTech) in Computer and Information Science from Cochin University of Science and Technology and PhD Degree in Computer Science and Engineering from Indian Institute of Technology (IIT Kharagpur). He has more than 17 years of experience in teaching and research as a faculty at Cochin University of Science and Technology. He was the former Head, Information Technology Division, at Cochin University of Science and Technology. He has published more than 35 research papers in International Conferences and Journals. His research interest includes big data analytics, distributed computing and automated software engineering.

Mariamma Chacko was born in 1961 at Changanacherry, India. She received her Bachelor's Degree in Electrical Engineering from University of Kerala in 1985, Master's Degree in Electronics from Cochin University of Science and Technology in 1987 and PhD Degree in Computer Engineering from Cochin University of Science and Technology in 2012. She has been working as Associate Professor in the Department of Ship Technology at Cochin University of Science and Technology since 2006. She joined the Cochin University as Member of the Faculty in 1990. From 1987 to 1990, she was associated with the Department of Electronics, Cochin University of Science and Technology as Research Associate. She has 12 research publications to her credit, and her research interests include validation and optimization of embedded software, sensorless control of BLDC motors and power quality.