Fuzzy Clustering with Ensemble Classification Techniques to Improve the Customer Churn Prediction in Telecommunication Sector

J. Vijaya, E. Sivasankar and S. Gayathri

Abstract Retention of customer improves the profit growth of a marketing firm. Customer retention is the process of identifying and retaining the customers who are about to slide from one brand to another on any dissatisfaction. Customer Correlation Management (CCM) process helps the organization in this retention of customer. A company which performs the retention of customer efficiently is achieving a high market value. Nowadays, customer retention is highly required in the telecommunication systems as the sliding of customers from one provider to the other is increasing. Earlier this customer retention was made using single classification techniques which were less efficient than the hybrid models. Clustering the customers who all belong to similar characteristics into groups and then classifying each group is called the hybrid model. Further to improve the classification in our hybrid model, ensemble techniques are used, which are more efficient. In the first phase, fuzzy-based clustering methods such as Fuzzy C-Means (FCM), Possibility C-Means (PCM), and Possibility Fuzzy C-Means (PFCM) are used for clustering the customers into groups. In the second phase, those clustered groups are partitioned into training and testing data using holdout. In the third phase, the training data are given to ensemble models like bagging, boosting, and Random Subspace (RS) algorithms for building the model. The test data are predicted based on the majority voting, provided by the ensemble techniques. Through this analysis, it is found that the proposed fuzzy clustering with ensemble classification techniques provides more accuracy than single classifier and clustering with base classifier.

Keywords CCM · Churn · Clustering · FCM · PCM · FPCM
Ensemble classification · Bagging · Boosting · Random subspace

J. Vijaya (✉) · E. Sivasankar
NIT Trichy, Trichy, India
e-mail: 406114003@nitt.edu; vijayacsedept@gmail.com

E. Sivasankar
e-mail: sivasankar@nitt.edu

S. Gayathri
Temporary Faculty, NIT Trichy, Trichy, India
e-mail: sgayathri@nitt.edu

J. Kalita et al. (eds.), *Recent Developments in Machine Learning and Data Analytics*,
Advances in Intelligent Systems and Computing 740,
https://doi.org/10.1007/978-981-13-1280-9_25

1 Introduction

Due to privatizations throughout the world, telecommunication industry is getting multiplied in the marketing environment. Nowadays, along with voice services telecommunication industry provide data services, online gaming, e-tickets booking, online purchasing, online banking, entertainments, educational services, and many more. Many customers are highly using these provisions that are provided by the telecommunication industries, which help the customer in many ways [1]. Along with the various services provided there are evolving of technology like a shift from 4G to 5G technology also. Because of this, the customers are searching for better service providers and technologies day by day. After identifying a better service provider the customers could easily shift to another service provider even without changing their existing mobile numbers. For example, in India due to the evolving of Jio service provider, churn rate from other service providers was increased enormously [2]. The cost of adding new customers into an organization is six times more than preserving an existing customer. Hence, in order to retain their customer churn prediction becomes a highly wanted technique for every telecommunication service provider [3]. Through the study of many literatures, it is found that machine learning algorithms such as linear regression, logistic regression, Support Vector Machine (SVM) cart, Naive Bayes (NB), K-Nearest Neighbor (KNN), Apriori, K-means, Principle Component Analysis (PCA), bagging, random forest, boosting with Ada-Boost are highly contributing in telecommunication customer churn prediction [4–7]. These machine learning techniques helps not only in the churn prediction of the telecommunication industries but also in various other industries like banking, gaming, social media, online marketing, insurances, restaurants, and so on [8–10]. In this paper, a combination of fuzzy clustering with an ensemble classification techniques-based hybrid churn prediction model is proposed. The data set provided by KDD cup 2009 a French-based telecommunication company on customer information is used for our analysis.

2 Literature Survey

Earlier the customer retention was made using single classification techniques [1] which were less efficient than the hybrid models [4, 5, 8, 11–14]. To cluster the customers who are having the similar characteristics into segments, Hudaib et al. [4] used the clustering methods like K-means, Self-Organizing Map (SOM), and hierarchical clustering algorithms. Then, the clustered segments are fed into the Artificial Neural Networks (MLP-ANN) classifier for hybrid model building and the accuracy is evaluated [4]. Out of 271 attributes in the data set Bose and Chen [5] selected 14 important attribute in which 7 is based on revenue and 7 is based on minute of usage. Then, the selected 14 attributes are fed into five known clustering techniques for segmentation. Further, the segmented clusters are input into the boosted Decision

Tree (DT) classifier for hybrid model building and top-decile lift is evaluated [5]. Rajamohamed and Manokaran [8] used the data set of credit card churn prediction from UCI repository for their hybrid model building which uses improved rough K-means algorithm for clustering and KNN, DT, SVM, NB, and ANN for classification. Hence, they proved their proposed hybrid model to be efficient when compared with a single classifier [8]. Huang and Kechadi [11] built a hybrid model on telecommunication data set which uses proposed weighted K-means algorithm for clustering and first-order inductive learning for classification [11]. Tsai and Lu [12] tried with two hybrid models in which one was the combination of SOM and back-propagation ANN and another by combining a neural with the same neural network and found that the earlier was better than the forth [12]. Two hybrid models were proposed by Hung et al. [13] in which one was a combination of K-mean with DT and another model combining DT with NN and evaluated the result using hit ratio and LIFT in both the models [13]. Vijaya and Sivasankar [14] built a hybrid model on telecommunication data set which uses proposed K-means and K-medoids algorithms for clustering and DT, SVM, NB, KNN, and LDA algorithm for classification.

3 Churn Prediction Models and Methods

Figure 1 depicts the flow diagram of the proposed hybrid model. The data set provided by KDD cup 2009 a French-based telecommunication company on customer information is preprocessed for further processing. In the first phase, fuzzy-based clustering methods such as Fuzzy C-Means (FCM), Possibility C-Means (PCM), and Possibility Fuzzy C-Means (PFCM) are used for clustering the customers into groups. In the second phase, those clustered groups are partitioned into training and testing data using holdout method. In the third phase, the training data are given to ensemble models like bagging, boosting, and random subspace algorithms for building the model. The test data are predicted based on the majority voting, provided by the ensemble techniques. The efficiency is measured using the performance metrics such as accuracy, TPR, and FPR.

3.1 Data Set

The French-based company, orange provides two different sets of data set, namely, small data set and large data set [15]. Here, we have considered the training data of the small data set for our experimentation. The attribute names of the data set are not revealed for maintaining the privacy of the customer. There are 50,000 samples and 230 features in the data set. Among the 50,000 samples, 46328 samples are non-churn samples and 3672 samples are churn samples. Among 230 attributes, 190 are numerical and 40 are string attributes.

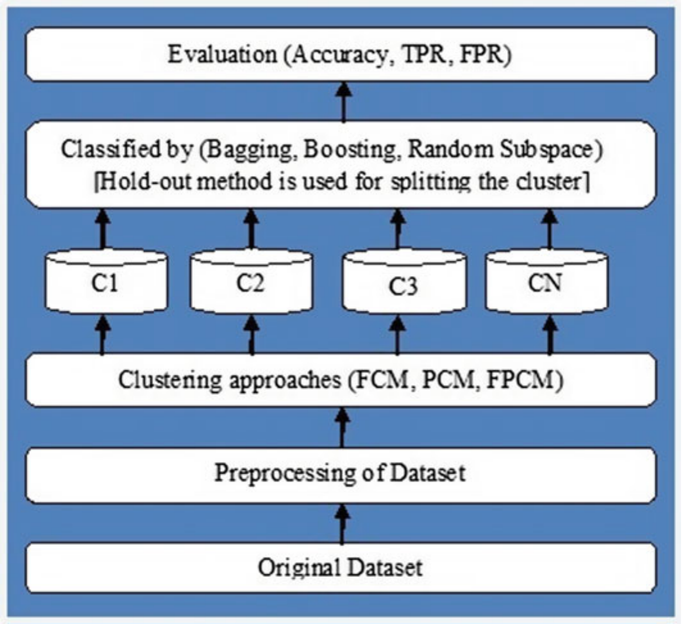
Evaluation (Accuracy, TPR, FPR)
Classified by (Bagging, Boosting, Random Subspace)
[Hold-out method is used for splitting the cluster]
C1
C2
C3
CN
Clustering approaches (FCM, PCM, FPCM)
Preprocessing of Dataset
Original Dataset

Fig. 1 Fuzzy clustering with ensemble classification customer churn prediction model

3.2 Data Set Preprocessing

In the above considered small data set there are 60% of missing values, which needs effective preprocessing is needed to obtain an effective data set for the experimentation. Preprocessing is carried out using the following procedure. If there is 25% of missing value in an attribute they are totally removed from the data set. In doing this, the data set gets reduced to 67 attributes. If there are further missing values in these 67 attributes they are filled in two different principles. That is if it a numeric value it is filled by means and if it's a string value it is filled by mode. Thus, filled string value had a higher deviation, and hence, it is removed from the data set which reduces the count to 49. Further, the string values are converted into numeric value because the clustering algorithm supports only a numerical value. The data set thus preprocessed are normalized using min-max normalization.

3.3 Clustering Techniques

Clustering is basically an unsupervised learning methodology in the wide area of data mining. Considering any data set, clustering can be used to group data with similar features into a common cluster, in which an efficient clustering method will generate clusters such that the interdependency between the clusters is very less. There are many clustering methods that exist. We have chosen three different fuzzy-based clustering techniques which are discussed below.

3.3.1 FCM

J. C. Dunn in 1973, based on fuzzy technique, propounds a clustering algorithm called fuzzy C-means algorithm [16]. There is a drawback in partition-based clustering algorithms where if an object is closer to more than one cluster formed, it has to be clustered only in one cluster which has minimum distance. To overcome this drawback in fuzzy-based clustering algorithm, even if an object is closer to more than one cluster it can be clustered into all the clusters formed to whichever it is closer. In FCM, entire data set D has N tuples is clustered into K number of clusters, where $K \leq N$ is a user-defined variable. Based on the value of K, K cluster centers are formed randomly. For every tuple in the data set, the fuzzy membership value is calculated using Eq. 1 for every cluster center, where the summation of all the fuzzy membership function should be equal to 1. Now, the clusters are formed by grouping the tuples which are having the maximum fuzzy membership function with the cluster center. Next, new cluster centers are identified for every cluster formed using Eq. 2. These steps are repeated again and again till we obtain a minimum objective function which is defined using Eq. 3.

$$M_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}} \tag{1}$$

$$V_j = \frac{\sum_{i=1}^{n} (M_{ij})^m . x_i}{\sum_{i=1}^{n} (M_{ij})^m} \tag{2}$$

$$J(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} \left(M_{ij}\right)^m \left\| x_i - V_j \right\|^2 \tag{3}$$

3.3.2 PCM

To overcome the outliers that are obtained during FCM, possibility C-means algorithm was propounded by R. Krishnapuram in 1996 [16]. Outliers in FCM is found because it depends highly on membership function alone, and hence, another function called typicality matrix is used. FCM also faces the problem of handling large feature data set. In PCM, entire data set D has N tuples is clustered into K number of clusters, where $K \leq N$ is a user-defined variable. Based on the value of K, K cluster centers are formed randomly. For every tuple in the data set, typicality matrix is calculated using Eq. 4 for every cluster center. Now, the clusters are formed by grouping the tuples based on the typicality matrix. Next, new cluster centers are identified for every cluster formed using Eq. 5. These steps are repeated again and again till we obtain a minimum objective function which is defined using Eq. 6.

$$T_{ij} = \frac{1}{(1 + \frac{d_{ji}}{\eta_j})^{\frac{1}{m-1}}} \tag{4}$$

$$V_j = \frac{\sum_{i=1}^{n} (T_{ij})^m . x_i}{\sum_{i=1}^{n} (T_{ij})^m} \tag{5}$$

$$J(U, V) = \sum_{j=1}^{n} \sum_{i=1}^{c} (T_{ij})^m (d_{ji})^2 + \sum_{j=1}^{n} \eta_i \sum_{i=1}^{c} (1 - T_{ij})^m \tag{6}$$

3.3.3 FPCM

Combining the membership function of FCM and typicality matrix function of PCM, Pal and Bezdek in 1997 propounded fuzzy possibility C-means algorithm [16]. For every tuple in the data set, membership matrix is calculated using Eq. 1 and typicality matrix is calculated using Eq. 4 for every cluster center. Now, the clusters are formed by grouping the tuples based on the membership matrix and typicality matrix. Next, new cluster centers are identified for every cluster formed using Eq. 7. These steps are repeated again and again till we obtain a minimum objective function which is defined using Eq. 8.

$$V_j = \frac{\sum_{k=1}^{n} (a.M_{jk}^m + b.T_{jk}^n).x_k}{\sum_{k=1}^{n} (a.M_{jk}^m + b.T_{jk}^n)} \tag{7}$$

$$J(U, V) = \sum_{j=1}^{n} \sum_{i=1}^{c} a * M_{jk}^m + b.T_{jk}^n * \left\| x_i - V_j \right\|^2 + \sum_{j=1}^{n} \eta_i \sum_{i=1}^{c} (1 - T_{ij})^m \tag{8}$$

3.4 Ensemble Classification Techniques

To enhance the weak classifiers formed using a single classification technique, ensemble classification technique is introduced. Here, the size of the ensemble and the classification technique used are the important factors. The following are the most important ensemble classification techniques.

3.4.1 Bagging

Bagging is a bootstrap aggregation algorithm propound by Breiman in 1996 [17]. Considering the training data set, *N* number of bags is formed using sampling with replacement mechanism. Further, each bag is fed into any classifier to build the model where each model produces separate prediction results for every test instance. The majority of the similar output produced by each model will be the final result of each test instance.

3.4.2 Boosting

Boosting is an algorithm that uses the weighted training set [17]. That is it assigned a weight commonly to all the tuples and are given to the classifier to build the model and the performance is evaluated in the model. The weight of the miss predicted tuples is enhanced, and these boosted tuples are again fed into the next classifier for building the model. The previous step is repeated continuously till the miss prediction is reduced.

3.4.3 Random Subspace

Random Subspace (RS) is a feature bagging algorithm propound by Ho in 1998 [17]. Considering the training data set, N number of subsets are formed based on the various combinations of attributes. Further, each subset is fed into any classifier to build the model where each model produces separate prediction results for every test instance. The majority of the similar output produced by each model will be the final result of each test instance.

4 Experiments and Result Analysis

4.1 Performance Measures

The performance of single classification, ensemble classification, and the proposed methods are compared using the performance metrics like accuracy, TPR, and FPR. A good churn prediction model should show high accuracy, high true positive rate, and low false positive rate. Accuracy is the measure of the ratio between accurately predicted churn and non-churn customer with entire customers in the data set. True positive rate is the accurate prediction of churn customers over the given churn customers. False positive rate is the miss prediction of non-churn customer as a churn customer.

Table 1 Performance of single classifiers [14]

Classifiers	DT	KNN	SVM	NB	LDA
Accuracy	87.61	90.91	**92.55**	91.39	89.52
TPR	93.68	98.00	100.0	99.00	97.96
FPR	12.23	02.84	00.00	01.00	00.16

Table 2 Performance of ensemble classifiers

Ensemble classifiers	Bagging	Boosting	Random subspace
Accuracy	92.33	**94.19**	93.32
TPR	97.07	97.73	97.62
FPR	04.29	07.69	06.62

4.2 Setup 1: Performance of Single Classifiers

The preprocessed data set of the experiment consists of 50,000 samples and 49 attributes with one churn attribute. This data set is input into the single classifier model and the performance is evaluated using the accuracy, TPR, and FPR. Holdout method is used to divide the preprocessed data into training part and the test part. The obtained results are tabulated in Table 1 which is already discussed in our earlier paper [14].

4.3 Setup 2: Performance of Ensemble Classifiers

The preprocessed data set of the experiment consists of 50,000 samples and 49 attributes with one churn attribute. This data set is input into the ensemble classifier model and the performance is evaluated using the accuracy, TPR, and FPR. Holdout method is used to divide the preprocessed data into training part and the test part. The obtained results are tabulated in Table 2.

4.4 Setup 3: Performance of Proposed Fuzzy Clustering with Ensemble Techniques

In the proposed hybrid model, fuzzy-based clustering methods such as Fuzzy C-Means (FCM), Possibility C-Means (PCM), and Possibility Fuzzy C-Means (PFCM) are used for grouping the customers into clusters with different *C* values, where *C* is the number of clusters to be generated. Those clusters are partitioned into training and testing data using holdout the method. Later, the training data are given to ensem-

Table 3 Performance of ensemble classifiers with FCM using various *C* values

Clustering technique	FCM								
Ensemble techniques	Bagging			Boosting			RS		
Number of clusters	C = 2	C = 4	C = 6	C = 2	C = 4	C = 6	C = 2	C = 4	C = 6
Accuracy	92.09	**93.52**	91.44	95.47	95.38	**96.50**	93.86	93.50	**94.54**
TPR	96.53	97.42	97.06	96.48	100.0	97.53	98.20	99.87	99.91
FPR	04.24	07.40	05.97	02.73	00.00	01.17	06.06	00.97	01.23

Table 4 Performance of ensemble classifiers with PCM using various *C* values

Clustering technique	PCM								
Ensemble techniques	Bagging			Boosting			RS		
Number of clusters	C = 2	C = 4	C = 6	C = 2	C = 4	C = 6	C = 2	C = 4	C = 6
Accuracy	93.14	**93.27**	91.67	96.20	**97.03**	96.68	94.91	94.43	**95.20**
TPR	97.23	97.59	96.12	96.65	97.24	96.88	95.33	94.42	95.11
FPR	07.81	06.60	05.15	13.51	06.66	06.97	16.66	05.26	02.77

ble models like bagging, boosting, and random subspace algorithms for building the model. The test data are predicted based on the majority voting, provided by the ensemble techniques. The efficiency is measured using the performance metrics such as accuracy, TPR, and FPR, and the results are tabulated in Tables 3, 4 and 5. Figure 2 depicts the accuracy comparison between the proposed fuzzy clustering with ensemble classification techniques and single ensemble classification techniques. Figure 2 shows that the Possibility Fuzzy C-Means (PFCM) hybrid with boosting produce maximum accuracy of 97.86%. Table 6 result shows the performance comparison between the proposed churn prediction model and existing techniques.

4.5 Prediction of Various Other Applications

To test the capability of our propound fuzzy clustering with ensemble classification technique, we have considered other predicting applications like bank marketing data set, credit approval data set, heart disease predicting data set, and telecommunication churn prediction data set from UCI repository for our analysis [18]. The results are discussed in Tables 7 and 8, which show that our propound model could predict the other benchmark applications efficiently. First, the collected data sets are input

Table 5 Performance of ensemble classifiers with FPCM using various *C* values

Clustering technique	FPCM								
Ensemble techniques	Bagging			Boosting			RS		
Number of clusters	C = 2	C = 4	C = 6	C = 2	C = 4	C = 6	C = 2	C = 4	C = 6
Accuracy	94.11	**94.20**	93.13	96.95	97.75	**97.86**	**96.68**	96.09	95.61
TPR	99.77	99.78	99.89	96.90	97.74	98.36	97.13	96.53	96.05
FPR	01.12	00.23	00.78	04.68	02.12	09.80	12.19	13.88	15.62

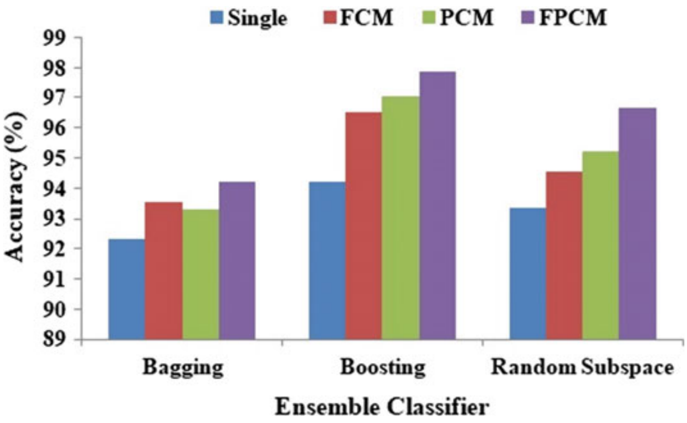

Fig. 2 Accuracy comparison between hybrid ensemble classifier and single ensemble classifier

Table 6 Accuracy comparison with existing techniques

Churn prediction models	Accuracy
Proposed clustering with ensemble (FPCM+Boosting)	**97.86**
Single classifier (SVM) [14]	92.55
Ensemble of classifier (Boosting)	94.17
Clustering with classifier (K-Means+KNN) [14]	94.72
Clustering with classifier (K-Means+BoostedC5.0) [14]	92.84

Table 7 Predicted accuracy of single classifiers versus ensemble classifiers for the collected four benchmark data sets

Data set information			Single classifiers					Ensemble classifiers		
Data sets	# Samples	#Features	DT	KNN	SVM	NB	LDA	Bagging	Boosting	RS
Bank marketing	45211	17	77.62	77.11	79.78	52.63	65.75	78.10	83.49	82.22
Credit approval	690	15	85.54	84.73	85.33	84.80	82.19	86.66	89.22	88.71
Heart disease	303	75	82.01	81.70	83.24	81.88	72.65	84.15	82.70	83.97
Telecom churn	5000	21	91.01	91.94	92.04	92.75	92.23	93.37	94.30	93.90

Table 8 Predicted accuracy of proposed fuzzy clustering with ensemble classifiers for the collected four benchmark data sets

			Number of clusters C = 2								
Data set information			FCM			PCM			FPCM		
Data sets	# Samples	#Features	Bagging	Boosting	RS	Bagging	Boosting	RS	Bagging	Boosting	RS
Bank marketing	45211	17	82.20	83.74	83.72	83.88	83.87	86.34	87.29	**88.75**	88.69
Credit approval	690	15	92.66	92.72	93.43	92.79	93.70	92.15	93.18	94.42	**95.25**
Heart disease	303	75	85.30	85.42	84.24	87.32	87.14	87.32	**88.74**	88.59	88.44
Telecom churn	5000	21	94.46	95.32	96.11	96.79	95.53	94.65	97.53	**98.40**	97.09

into the single classifier model and ensemble classifier model then the performance is evaluated using the accuracy. Holdout method is used to divide the data into training part and the test part. The obtained results are tabulated in Table 7. Next, the collected data sets are input into the proposed hybrid model, fuzzy-based clustering methods such as Fuzzy C-Means (FCM), Possibility C-Means (PCM), and Possibility Fuzzy C-Means (PFCM) with ensemble models like bagging, boosting, and random subspace algorithms. The efficiency is measured using the performance metrics such as accuracy, and the results are tabulated in Table 8.

5 Conclusion

Nowadays, effective churn prediction for an organization has become an essential process to withstand its position in the market. In this paper, we have deployed a hybrid fuzzy clustering with an ensemble classification model for telecommunication firm. Through this experimentation, it can be concluded that (1) the result produced by the proposed hybrid fuzzy clustering with ensemble technique is providing better performance than the single classifier and the ensemble classifier. (2) Among the hybrid fuzzy clustering with ensemble technique, boosting with FPCM produces a better performance compared with the others. (3) FPCM performs better than FCM and PCM because clustering is done efficiently. This proposed hybrid model can be extended for any firm for their churn prediction.

References

1. Huang, B., Kechadi, M.T., Buckley, B.: Customer churn prediction in telecommunications. Expert Syst. Appl. **39**(1), 1414–1425 (2012)
2. Web page reference. https://gadgets.ndtv.com/telecom/opinion/reliance-jio-business-model-how-can-it-make-money-1454531
3. Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, K.C.: A comparison of machine learning techniques for customer churn prediction. Simul. Model. Pract. Theory **55**, 1–9 (2015)
4. Hudaib, A., Dannoun, R., Harfoushi, O., Obiedat, R., Faris, H.: Hybrid data mining models for predicting customer churn. Int. J. Commun. Netw. Syst. Sci. **8**(05), 91 (2015)
5. Bose, I., Chen, X.: Hybrid models using unsupervised clustering for prediction of customer churn. J. Organ. Comput. Electron Commer. **19**(2), 133–151 (2009)
6. Xiao, J., Xiao, Y., Huang, A., Liu, D., Wang, S.: Feature-selection-based dynamic transfer ensemble model for customer churn prediction. Knowl. Inf. Syst. **43**(1), 29–51 (2015)
7. Idris, A., Khan, A., Lee, Y.S.: Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. Appl. Intell. **39**(3), 659–672 (2013)
8. Rajamohamed, R., Manokaran, J.: Improved credit card churn prediction based on rough clustering and supervised learning techniques. Clust. Comput. 1–13 (2017)
9. Runge, J., Gao, P., Garcin, F., Faltings, B.: Churn prediction for high-value players in casual social games. In: Computational Intelligence and Games (CIG), pp. 1–8. IEEE, Aug 2014