Neural Computing and Applications (2019) 31:7181–7200
https://doi.org/10.1007/s00521-018-3548-4

ORIGINAL ARTICLE

Hybrid PPFCM-ANN model: an efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network

E. Sivasankar[1] · J. Vijaya[1]

Received: 8 November 2016 / Accepted: 11 May 2018 / Published online: 28 May 2018

Abstract

A vital issue for customer correlation management and consumer conservation in the telecommunications business is increased customer churn. The data mining approaches can aid in the prediction of churn behavior of consumers. This article aims to propose a system to predict customer churn through hybrid probabilistic possibilistic fuzzy C-means clustering (PPFCM) along with artificial neural network (PPFCM-ANN). This paper comprises of two modules: (1) proposing clustering component on the basis of PPFCM and (2) churn prediction component on the basis of ANN. The input dataset is gathered into clusters, with the help of probabilistic possibilistic fuzzy C-means clustering algorithm in the clustering module. The obtained clustered information is used in the artificial neural network, and this hybrid construction is further used in the churn prediction module. During the testing process, the clustered test data select the most accurate ANN classifier which corresponds to the closest cluster of the test data, according to minimum distance or similarity measures. Finally, to predict the churn customer the output score value is utilized. Three sets of experiments are carried out: the primary set of experiments comprises PPFCM clustering algorithm, the secondary set assesses the classification result, and the third set authenticates the proposed hybrid model presentation. The proposed hybrid PPFCM-ANN model provides maximum accuracy when compared to any single model.

Keywords Customer churn prediction · Telecommunication · Hybrid CCP · Clustering · Fuzzy C-means · Possibilistic fuzzy C-means · Probabilistic possibilistic fuzzy C-means · Classification · Levenberg–Marquardt classifier · Artificial neural network

1 Introduction

Due to liberalization and globalization, the size of the information and communication technology (ICT) market is increasing as numerous service providers are entering the market. This increase provides innovative and efficient service to customers, which, in turn, tempts a customer to shift from one supplier to another [1]. When a customer closes his/her business from a service provider, it is called ‘churning’. The retention of customers depends on customer satisfaction and maintenance. This is a keystone in customer relationship management (CRM) [2, 3]. Customer retention is only possible by calculating the various set of attributes of the customer going to churn, but such a calculation is very difficult because in manual recording of data, accuracy is not effective and also it is time-consuming [4]. The continuous and persistent churning of customers heavily damages business. The mobile telecommunication industry would do better if it could forecast the churning customers. The industry tries to make peripheral changes and retain the customers and income [5]. Forecasting the clients that are likely to break their contract and shift to competitors (churn) is a necessity. If such churning customers are identified in advance, then effective steps can be taken to retain them. This will increase the profit of the company [6].

In any business scenario, it is tragic to lose an existing customer. There are three ways of customer churning.

✉ J. Vijaya
406114003@nitt.edu; vijayacsedept@gmail.com

1 Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, India

Some customers may leave because of problems with the service provider. These may be network failure, high tariffs, billing issues, etc. Secondly, there are some customers whose nature is to change from one supplier to another. Thirdly, some customers tend to change the service provider for reasons that elude the communication industry. This third type of customers and their reasons must be identified in advance, and their churning must be prevented [7]. The enrollment of a new customer is difficult and expensive. So, churn prediction and churn prevention are the only way of success for the telecommunication company [8]. Once a possible churn customer is identified, the company can offer concessions and gifts to retain such customer. The attraction of customers through coupons and e-mail marketing is followed by major telecommunication companies. The trends of the churn customers are analyzed, and suitable offers are announced for different groups according to their classification. Some may get call concessions, and others may get data benefits. During this classification, customers who are valuable to the company/industry are identified and suitable high value offers are extended to them [7].

If the customer churn prediction corporation has several methodologies in data mining techniques to predict the customers churn well ahead, then the company can stand well balanced in terms of customer base, market share and market price [9]. The data mining method can even access hidden information such as customer behavior [10]. The telecommunication companies always aim at retaining the existing customers, which in turn will reduce the cost, time and manpower involved to build up an entirely new customer base. In order to predict customer churn, wide number of machine learning methods are in practice, which include decision trees (DT), K-nearest neighbor (KNN), naive Bayes (NB), neural network (NN), regression analysis-logistic regression (LR), rule-based classification first-order inductive learning (FOIL) and support vector machine (SVM) [11–14]. These play a key role in predicting customer churn in banking, marketing, pay-TV, insurance, textile industries, electronic commerce, newspaper publishing, medical and Internet service providers. In the past, there were only classification algorithms based on a single model. Recent researchers focused on ensemble-based classification techniques. These recent techniques deploy hybrid algorithms combining many single model classifications, and their forecasts are merged into one aggregated outcome with the help of a fusion rule [15]. A scrutiny of the literature confirms that hybrid multi-classification techniques are more effective than single classification techniques [16]. The hybrid model is a proper substitute for enhanced classification function. This technique replaces the single model classification and leads to higher accuracy of the hybrid model and is adaptable for all methods. Earlier, the whole training information was utilized to shape prediction models. But breaking down the entire training data into sets delivers good results in comparison with the usage of whole training data samples. This information instance is carried forward for grouping similar customers.

In this work, the CCP on the basis of hybrid model completely integrating clustering and classification learning algorithms is proposed. Primarily, the training information is grouped on the basis of probabilistic possibilistic fuzzy C-means clustering algorithm. Subsequently, all cluster information is assigned to the artificial neural network classifier to gather consumer information. In testing procedure, the information to be assessed depends on the greatest suitable classifier that is evolved from the most related cluster of the test examples. The main contribution of the projected hybrid model is given as follows:

- *Clustering* The proposed probabilistic possibilistic fuzzy C-means clustering is used in clustering stage. This will group the data into U number of clusters. The PPFCM algorithm uses the membership and typicality matrix with probability values to cluster the unlabeled data. The novel method of PPFCM overcomes various problems of the fuzzy C-means (FCM) algorithm, possibilistic C-means (PCM) algorithm, fuzzy possibilistic C-means (FPCM) and possibilistic fuzzy C-means (PFCM) algorithm.
- *Churn data prediction* In churn prediction stage, all test information is calculated with the help of selecting the most appropriate classifier, which resembles the neighboring cluster to the test information based on the least distance or resemblance measures. Levenberg–Marquardt (LM), a gradient descent method for training neural nets by backpropagation is used for the process of classification.

The rest of the paper is organized as follows: in Sect. 2, work done by researchers in this same field (churn prediction) is presented, and in Sect. 3, the proposed customer churn prediction method is described. In Sect. 4, the experimental results and performance evaluation are explained, and in Sect. 5, the concluding part is discussed.

2 Review of related works

There are four levels in the management of customer relationship with the organization, i.e., (1) identifying the customers, (2) attracting the customers, (3) retention of the existing customers and (4) development of customers. These four levels aim at understanding the customers so that long-term retention of the customers can be achieved which maximizes the benefits of both customer and company. Churn prediction helps the organization to enhance profit and maintain a steady customer base. Churn prediction can be done using various information mining and soft computing techniques. These methods can be utilized at various stages of the information mining procedures like eradicating the noise and outliers, decreasing the feature space by choosing most applicable attributes, sampling selection. In this segment, the methods that use single classification model to predict the customer churn are discussed. Huang et al. discovered a new set of features for landline telecom customers. Based on these features, they developed customer churn prediction models. For prediction models, they used seven classification techniques such as NB, DT, SVM, LR, multilayer perceptron NN, linear classification and evolutionary data mining algorithm. Of these seven classifications, evolutionary data mining algorithm did had not give good results, but the other six classifications produced better results [17].

Koen et al. developed an ensemble of classification models. They named it as a rotation-based classification model. They implemented two algorithms such as Rotation Forest and Rot-Boost with AdaBoost. They compared Rotation Forest (RF) and Rot-Boost with other classifiers such as Bagging, Random Forest, RSM, CART and C4.5. Also they compared Rotation Forest and Rot-Boost with the feature extraction techniques of principal component analysis (PCA), sparse random projections (SRP) and independent component analysis (ICA). After all the experiments, they declared Rot-Boost to be better than Rotation Forest. They also found that ICP-based RF is the best of all the existing classifiers [18]. Ozden et al. applied the churn prediction model to banking customers in a dynamic manner rather than statically, in which all the existing works have been done. They did it in three sets of customer data. The first set of data was not based on time, and it was a single observation for each customer (SPTD). The second set of data was time based, and multiple observations were made for each customer at different times (MPTD). In the third type, lag was added to SPTD and the customer churn was predicted. After creating the three datasets, they classified it under logistic regression and decision tree [19]. Verbeke et al. used multiple classifiers such as tree induction-based M5, CART, logistic regression, neural network-based LS-SVM, multilayered perceptron NN, MARS, radial basis NN and case-based learning approach. They used wrapper backward attribute selection approach to identify the best features [20]. Keramati et al. proposed the hybrid model using the dataset collected from the Iranian mobile company. In the proposed model field, 95% of accuracy was achieved with good precision and recall. The accuracy of the projected model is compared with existing learning algorithms such as DT, KNN, ANN and SVM [21].

Kim et al. proposed an approach for customer churn prediction which examines the communication network of the particular customer. To better predict the customer behavior, a customer who uses their model and spreads churning information is studied as a part of the network instead of studying his/her behavior alone [22]. Data mining by evolutionary learning (DMEL) is a rule-based model proposed by Chang et al. This algorithm generates first-order rule at a lower level; then, higher-order rules are generated at the next level. Churn prediction is done based on these generated rules with improved accuracy [23]. Churn prediction model for predicting the churning customer in newspaper subscription has done by Poel et al. using SVM algorithm (grid search and cross-validation). They compared SVM with logistic regression and random forest and finally concluded that SVM outperforms only logistic regression [24]. Mozart et al. implemented logistic regression, decision tree, neural networks and boosting in a churn prediction dataset that contained 47,000 US domestic subscribers and included features like billing, credit, complaint history and application [25]. The importance of interpersonal influence is projected by Zhang et al. Traditional and network attributes are incorporated in the churn prediction model. The proposed model was interpersonal influence and outperforms the traditional model [26]. Verbeke et al. proposed the model for predicting the churning customers using the customers social networking information. A significant improvement is achieved in predicting the churning customers using Markovian network, which includes both network-based and non-network-based attributes [27].

This paper discusses a hybrid model which combines one or more classifiers or clusters combined with a classifier. Medhi et al. proposed a model, which combines artificial neural network and multiple linear regressions. The accuracy of the projected model is compared with existing knowledge and the learning algorithms such as LDA, QDA, DT and SVM [28]. Yeshwanth et al. proposed the model, which combines tree induction system and genetic algorithm for customer churn prediction [29]. Indranil et al. selected 14 attributes from a collection of 271 attributes related to the customer's revenue contribution and the minutes spent on calls. They used five clustering techniques, namely FCM, K-means, balanced iterative reducing and clustering using hierarchies, self-organizing, K-medoids to segment the customers. Each cluster is modeled by boosting decision tree, and the system uses a sensitivity top decile lift to evaluate and measure the performance of the combined clustering and boosted decision tree [30]. Mai et al. proposed two new hybrid algorithm sequential k-nearest neighbor and an extension of the Markov model, in which sequence alignment is combined with original KNN and higher-order

Markov model. The above models were tested on four different types of telecommunication churn prediction dataset (DS1, DS2, DS3 and DS4). Among the four considered datasets, DS1 and DS2 were tested using the proposed Markov sequence alignment (MSV), random model (RM), all kth Markov models (AllKth), hidden Markov model (HMM) and comparisons were made with the percentage of accuracy considering as the parameter. It is proposed from the obtained results that MSV outperforms all the other techniques. The percentage of correct prediction for DS1 is 27% and DS2 is 70%. In DS1, the second-order MSA performs better than other orders, and in DS2, the third-order MSA performs better than the other orders. The next two datasets DS3 and DS4 were tested using the K-nearest sequence with sequence alignment (KnsSA), random model (RM) and original KNN. It is observed from the obtained comparisons that KnsSA proved to outperform the other techniques [31]. Other hybrid models such as WKM-IL, K-means-DT, DT-ANN, ANN-ANN and KNN-LR are discussed in Sect. 4.4.3.

3 Proposed customer churn prediction system

The intention of this research is to design CCP model based on probabilistic fuzzy C-means and artificial neural network. Figure 1 shows how the proposed prediction system can be processed. The dataset is collected by the Tera data Center at Duke University, USA [32]. The collected data must be preprocessed to improve the accuracy. The proposed approach consists of two prediction phases: training phase and test phase. In the training phase, the proposed PPFCM clustering algorithm divides all the training data into U number of cluster training set (CTR). Because of CTR, the size and difficulty of every training subset will decrease and also the efficiency of succeeding ANN will increase. Each CTR is trained by using separate ANN. In the test process, all the test data are clustered into U number of cluster test sets (CTS) using the proposed PPFCM. The CTS uses the nearest ANN classifier of the training dataset and develops churn prediction model. Since the entire dataset is not used for clustering, only the subset of the dataset is used for model building.

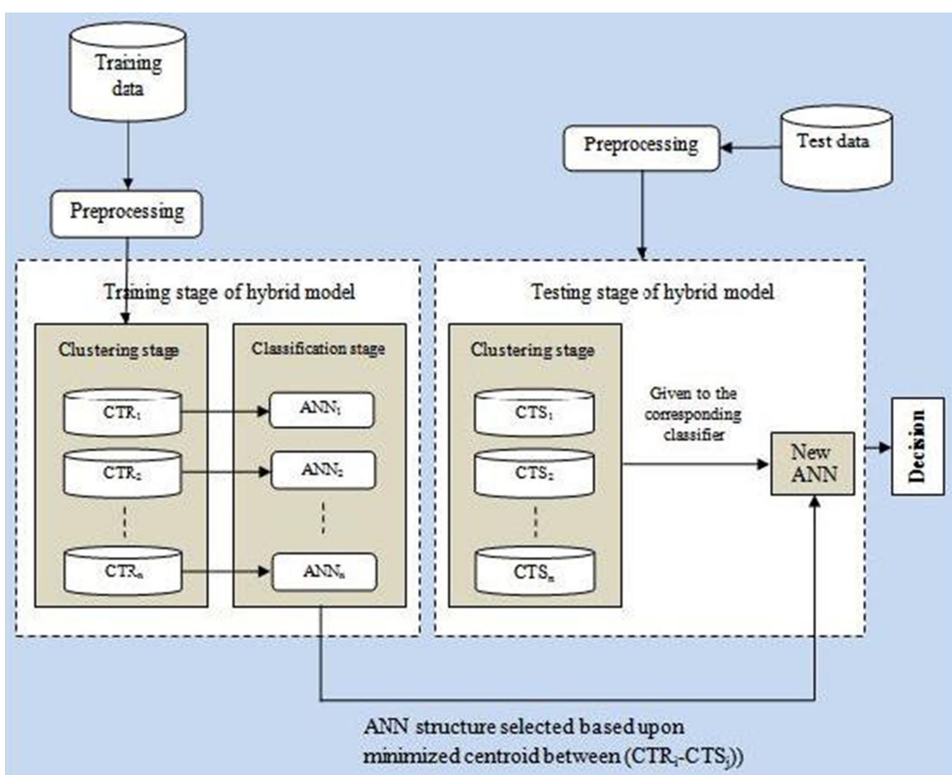


Fig. 1 Proposed hybrid churn prediction model (PPFCM-ANN)

3.1 Proposed probabilistic possibilistic fuzzy C-means (PPFCM) clustering algorithm

'*D*' represents the Duke University Telecommunication dataset, and '*N*' represents the number of customers. The dataset shows whether a customer has churned or not. If the customer is churned, 1 will be shown. If not churned, 0 will be shown. '*C*' represents the number of churned customers, and '*NC*' represents the number of non-churned customers. As there are only two options, churned and not churned, this can be treated as a binary classification problem. In this churn prediction model, clustering is very important. If the clustering is done properly, the accuracy of prediction will be very high [39, 41, 43], because similar type of customers is grouped within the cluster and analyzed for better customer segmentation. The customers having similar behavior patterns (characteristics) will be more likely to behave the same in the future if an unlabeled instance is predicted using partial training instance, which has similar characteristics with the tested instance, rather than the whole data. This can be achieved by dividing the training data into clusters, and the test instance is assigned to the closest cluster to it. The fuzzy C-means [33] clustering is the parent model of the proposed PPFCM. The FCM handles larger data and gives a quantified set of cluster information. But the proposed PPFCM does the same thing in a better manner. First the user is given the number of clusters (U); suppose the number of clusters is equal to two, it will form two clusters. The FCM then select U random customers and treat them as cluster center. It can find the fuzzy membership M_{ki} distance between each cluster center and each customer. Depending on the membership function which could be forty percentage or sixty percentage, the customer will be attached to the cluster of maximum membership value. The total of membership value for all the clusters is equal to 1. The membership function is calculated using Eq. (1).

$$M_{ki} = \frac{1}{\sum_{j=1}^{U}\left[\frac{\|x_k - v_i(y)\|_A}{\left\|x_k - v_j(y)\right\|_A}\right]^{\frac{2}{z-1}}} \tag{1}$$

Using this membership function, new cluster centers are formed, using Eq. (2).

$$v_i(y) = \frac{\sum_{k=1}^{N}(M_{ki})^z x_k}{\sum_{k=1}^{N}(M_{ki})^z} \tag{2}$$

When the objective function found through Eq. (3) becomes minimum, the algorithm will stop; otherwise the membership function and the cluster center will continue to be updated. When the objective function is minimized, the cluster quality becomes high because the objects within the cluster are of similar nature.

$$J_M = \sum_{k=1}^{N}\sum_{i=1}^{U}(M_{ki})^z\|x_k - v_i(y)\|_A^2 \tag{3}$$

where N is the number of information points, U is the number of clusters, M_{ki} is the fuzzy membership performance of x_k in a class i and z is the degree of fuzziness of the algorithm $(1 \leq z < \infty)$.$\|x_k - v_i(y)\|_A$ is the Euclidean distance between kth data point and ith cluster center with respect to the number of attributes A. The main objective of PPFCM is to improve the objective function of Eq. (3). The gradual techniques of the proposed PPFCM algorithm are comprehensively described in the following stages;

Step 1: Calculation of distance matrix In the clustering process, first the number of cluster U is defined by the user and the centroid $v_i(y)$ is chosen in a random fashion , which is equal to the number of clusters U. The Euclidean distance measure method is chosen for calculating the distance between data points and cluster centers. The Euclidean distance is calculated using Eq. (4).

$$d(x_k, v_i(y)) = \|x_k - v_i(y)\|_A, \quad 1 \leq i \leq U, \quad 1 \leq k \leq N$$
$$\|x_k - v_i(y)\|_A = \sqrt{\sum_{A=1}^{F}\left[(x_k)_A - (v_i(y))_A\right]^2} \tag{4}$$

$\|x_k - v_i(y)\|_A$ is the Euclidean distance between kth data point and i^{th} cluster center with respect to the number of attributes A. Here F is the number of features or attributes.

Step 2: Calculation of typicality matrix In this step, the typicality matrix T_{ki} is calculated using Eq. (5). It is derived from the probabilistic C-means clustering algorithm [34].

$$T_{ki} = \frac{1}{1 + \left[\frac{\|x_k - v_i(y)\|_A}{\gamma_i}\right]^{\frac{1}{(z-1)}}}, \quad 1 \leq i \leq U, \quad 1 \leq k \leq N \tag{5}$$

Step 3: Calculation of probability matrix Once the distance matrix and the typicality matrix is computed, then the probability matrix is calculated [35]. Let the cluster centers $v_i(y) = (v_1, v_2, \ldots, v_U)$, x be the data point and $\{d_i(x_k) : i = 1, \ldots, U\}$ be its distance from the given centers. The membership probability of x is denoted by P_{ki} in Eq. (6). $d_i(x_k)$ is an Euclidean between a given cluster center $v_i(y)$ and given data point x_k.

$$P_{ki} = \frac{\prod_{j \neq i} d_j(x_k)}{\sum_{t=1}^{i}\prod_{j \neq t} d_j(x_k)}, \quad i = 1, \ldots, U, \quad k = 1, \ldots, N \tag{6}$$

Step 4: Calculation of membership matrix The proposed membership matrix is the product of probability matrix P_{ki} and the membership matrix M_{ki} of FCM which is used in Eq. (1). The proposed membership matrix U_{ki} is calculated using Eqs. (7) and (8). The clusters are made based on the obtained maximum new membership value U_{ki}

$$U_{ki} = M_{ki} \times P_{ki} \tag{7}$$

$$U_{ki} = \frac{1}{\sum_{j=1}^{U}\left[\frac{\|x_k - v_i(y)\|_A}{\left\|x_k - v_j(y)\right\|_A}\right]^{\frac{2}{z-1}}} \times \frac{\prod_{j \neq i} d_j(x_k)}{\sum_{t=1}^{i}\prod_{j \neq t} d_j(x_k)} \tag{8}$$

Step 5: Update the centroid Once the clusters are made, the next step is to update the centroid based on Eq. (9).

$$v_i^l(y) = \frac{\sum_{k=1}^{N}(U_{ki} + T_{ki})x_k}{\sum_{k=1}^{N}(U_{ki} + T_{ki})}, \quad 1 \leq j \geq U \tag{9}$$

Once the centroid is upgraded for every cluster, the above steps are performed all over again using the new centroid and it continues up to the calculation of updating the new centroid. This procedure is rehashed until the updated centroids of every cluster are similar in consecutive iterations. Algorithm 1 explains the step by step procedures followed in the proposed PPFCM clustering.

3.2 Prediction based on artificial neural network (ANNLM)

In hybrid models, unsupervised learning techniques play a vital role for predicting better results (based on literature review). Each of the obtained clusters from the PPFCM clustering algorithm (Sect. 3.1) is trained using an U number of artificial neural network classifier. For every cluster, separate neural networks are used in the training process. The ANN indicates the biologically aggravated type of distributed evaluation. The network system contains an input layer and an output layer with at least one hidden layer in the middle of the input and output layers. Each originated consequence clusters from the PPFCM are experienced and trained in U number of artificial neural network classifier(s). In this article, the classical artificial neural network (Levenberg–Marquardt algorithm) is used to predict the churning customer. The Algorithm 2 explains the comprehensive Levenberg–Marquardt (LM) training procedure.

Algorithm 1: Proposed PPFCM clustering algorithm

Input: Database D
Output: Clustered data
Parameters:
D= Database $N \times M$
U = Total number of clusters
$v_i(y)$= Set of centroids
$d(x_k, v_i(y))$= Distance of k^{th} data point with respect to the i^{th} centroid
P_{ki}= Probability of $(x_k, v_i(y))$
T_{ki}= Typicality value of $(x_k, v_i(y))$
U_{ki}= Membership function
begin
  1.Define the number of cluster U
  2.Select $v_i(y)$
  3.Calculate the distance matrix $d(x_k, v_i(y))$ according to Eq. (4)
  5.Calculate the typically matrix T_{ki} according to Eq. (5)
  4.Calculate the probability matrix P_{ki} according to Eq. (6)
  6.Construct cluster based membership matrix U_{ki} according to Eq.(8)
  7.Update $v_i(y)$ according to Eq .(9)
    if all $v_i^l(y) = v_i^{l+1}(y)$
      Terminate
    else
      Go to step (3)
    end if
end

Algorithm 2: LM training process

1.Initialize the weight of input, output layers and learning rate (μ)
2.Calculate the Mean squared error over inputs Q(x) according to Eq. (10)

$$Q(x) = EE^T \tag{10}$$

$x = [x_1, x_2, ..., x_n]$ comprise of all weights of the network
E is an obtained Error vector
3.Resolve Eq.(11) to attain increase of weight $\Delta(x)$

$$\Delta x = [M^T M + \mu I]^{-1} M^T E \tag{11}$$

M is represented as Jacobian matrix
I is an Identity matrix
T is an Target value
4.Recomputed Mean squared error Q(x) with the help of $x + \Delta x$ by ways of the trail Q(x) and JUDGE
if trail $Q(x) < Q(x)$in step (2)
$x = x + \Delta x$
$\mu = \mu.\delta(\delta = 1)$
Satisfied means Go back to step (2)
else
$\mu = \frac{\mu}{\delta}$
Satisfied means Go back to step (4)
end if

3.3 Testing process

After the training process, the testing process is performed. The testing set (TS) is first given into a preprocessing stage. The preprocessed data are given to the clustering process. Here, the data are grouped based on the PPFCM clustering algorithm $(\mathrm{CTS}_1, \mathrm{CTS}_2, \ldots, \mathrm{CTS}_n)$. All test information is forecasted using the most appropriate classifier, where the neighboring cluster resembles the test information based on the minimum distance, i.e., $(\mathrm{CTR}_i - \mathrm{CTS}_i)$. Finally, the given data are checked to find out whether it falls under the churn or non-churn category based on the threshold value. If the score value is above the threshold, the data are said to be churn; otherwise, the data are said to be non-churn. Thus, the obtained score value is assessed for the purpose of categorizing data fulfilling the condition in Eq. (12).

$$\text{Decision} = \begin{cases} T_h \geq \text{score}; & \text{data are non-churn} \\ T_h < \text{score}; & \text{data are churn} \end{cases} \tag{12}$$

.

4 Experiments and results

In this section, the outcome is deliberated from the proposed hybrid consumer churn prediction model. For applying the suggested method, MATLAB version (2015b) is utilized. The proposed method is performed on a windows installed machine that has an Intel Core i3 processor along with 8 GB RAM and a speed of 2 GHz.

4.1 Dataset description

Three different versions of consumer churn datasets, i.e., calibration, current score and future score data, were collected from the Tera data Center at Duke University, USA. In the USA, these client churn datasets were collected from July 2001–December 2001. In this article, the data part of calibration set is taken. There are 100,000 samples and 172 attributes with 1 churn attribute present in this dataset. Out of 100,000 samples, 50,438 consumers are found to be non-churners and 49,562 consumers are found to be churn from the subscription. Out of 172 attributes, 137 attributes are found to be numerical attributes and 35 are found to be nominal attributes. Nearly 20% of the data consist of missing values. In this article, three kinds of attributes such as interactive qualities, company communication qualities and client household attributes are utilized. Table 1 illustrates the different samples of training and testing data that are used in the experiment. In Table 1, the third and fourth sections demonstrate the amount of churn and non-churn data which are utilized in the training process and the final column shows the amount of test samples to be utilized in the test process.

4.2 Dataset Preprocessing

Data preprocessing is the most essential and fundamental notion in the field of data mining. There are mainly three processes involved in data preprocessing such as missing value elimination, conversion of string to numerical value and normalization. Data preprocessing also means that the source database is changed in to a format which is

Table 1 Samples used for training and testing process

Sample (S)	Churn rate $\%(C^R)$	#Train churn (T^C)	#Train non-churn (T^{NC})	#Test sample
S1	5	2478	47,916	49,606
S2	10	4956	45,394	49,650
S3	15	7434	42,872	49,694
S4	20	9912	40,350	49,738
S5	25	12,391	37,828	49,781
S6	30	14,869	35,306	49,825
S7	35	17,347	32,785	49,868
S8	40	19,825	30,263	49,912
S9	45	22,303	27,741	49,956
S10	50	24,781	25,219	50,000

suitable for mining. In first step, the unwanted features (last_swap, Customer_ID) are eliminated from the dataset and reduced to 170 features (136 numeric and 34 nominal). Then, the missing data present in the nominal attributes are filled with the mode value of the particular attributes and the missing data present in the numerical attributes are filled with the mean value of the particular attributes. The string values are changed to numeric values because the clustering algorithms are capable of handling only the numerical type of data. A continual min-max normalization process is carried out to normalize the value.

4.3 Evaluation matrix

Clustering The quality of clustering techniques is measured by using sum of square error (SSE). The fundamental goal of clustering algorithm is to decrease the SSE value. When the SSE function is minimized, the cluster quality is high because the objects within the cluster are of similar nature. SSE is calculated using Eq. (13).

$$\mathrm{SSE} = \sum_{i=1}^{U} \sum_{j=1}^{N} ||C_i - O_{ij}||^2 \tag{13}$$

where U is the number of clusters and N is the number of information points in cluster C_i. For all U clusters, the Euclidean distance is calculated between every information point O_{ij} within a cluster to cluster center C_i and then is summarized.

Classification Two sets of classification techniques were evaluated on the basis of accuracy, true churn (TC), false churn (FC), precision (PPV) and negative predicted value (NPV). Any model performs better if it has high accuracy and greater true churn value and low false churn value. Table 2 indicates confusion matrix where Z_{11} indicates the number of churn customers that were correctly predicted, Z_{12} indicates the number of churn customers incorrectly predicted , Z_{21} and Z_{22} indicate the number of non-churning customers that were predicted incorrectly and correctly, respectively. Performance metrics were calculated using following Eqs. (14)–(18):

$$\mathrm{TC} = \frac{Z_{11}}{Z_{11} + Z_{12}} \tag{14}$$

$$\mathrm{FC} = \frac{Z_{21}}{Z_{21} + Z_{22}} \tag{15}$$

$$\mathrm{Precision} = \frac{Z_{11}}{Z_{11} + Z_{21}} \tag{16}$$

$$\mathrm{NPV} = \frac{Z_{22}}{Z_{12} + Z_{22}} \tag{17}$$

$$\mathrm{Accuracy} = \frac{Z_{11} + Z_{22}}{Z_{11} + Z_{12} + Z_{21} + Z_{22}} \tag{18}$$

Chi-square test Chi-square test is a nonparametric test used for two specific purpose: (a) to test the hypothesis of no association between two or more groups, population or criteria (i.e., to check independence between two variables) (b) and to test how likely the observed distribution of data fits with the distribution that is expected (i.e., to test the goodness-of-fit). Chi-square statistic is calculated using Eq. (19).

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{19}$$

Table 2 Confusion matrix

	Predict	
	Churn (C)	Non-churn (NC)
Actual		
Churn (C)	Z_{11}	Z_{12}
Non-Churn (NC)	Z_{21}	Z_{22}

where O_i is the number of observed value in the ith information point and E_i is the number of expected value in the ith information point.

4.4 Experiment setup

The function of the projected method is examined on the basis of three sets of experiments. In the first set of experiments, the proposed PPFCM clustering algorithm is compared with other related clusters algorithm. In the second experiment, the proposed CCP based on a hybrid model classifier is compared with other existing single model-based classifiers. Finally, the proposed CCP based on a hybrid model is compared with other existing hybrid model approaches.

4.4.1 Setup I

The performance of the proposed CCP method mostly depends on the clustering module. A good clustering method is used to increase the prediction accuracy. Therefore, in the experiments, the proposed clustering results are compared with the existing clustering algorithms such as K-means, FCM, PCM and PFCM.

K-means clustering K-means clustering is a clustering examination algorithm which divides objects (here customers) on the basis of their quality value into U disjoint clusters [41]. Primarily, the number of clusters U is described. Consequently, the initial cluster centroids are prepared by randomly selecting U objects from the entire objects. Then, the Euclidean distance is calculated between the chosen cluster centroids and each objects. Objects which have the minimum Euclidean distance are combined to form one cluster. Next the new cluster centroids for these newly formed clusters are calculated based on the means value. The process is repeated based on the new cluster centroids until the centroids do not alter any more.

Fuzzy C-means clustering The fuzzy clustering [33] is a clustering algorithm that can separate the information elements into U clusters so that item in the similar class is the same in nature. The FCM find out the membership value for each information point, which shows its association of the point with each cluster whose summation is equal to 1. Consider the unlabeled dataset $D = \{D_1, D_2, \ldots, D_N\} \subset R^P$. In clustering, this dataset D is partitioned into $1 < U < N$ subgroups. The objective function and fuzzy membership of the FCM clustering are furnished by means of Eqs. (20) and (21), as follows:

$$J_M = \sum_{k=1}^{N}\sum_{i=1}^{U}(M_{ki})^z \|x_k - v_i(y)\|_A^2 \tag{20}$$

$$M_{ki} = \frac{1}{\sum_{j=1}^{U}\left[\frac{\|x_k - v_i(y)\|_A}{\left\|x_k - v_j(y)\right\|_A}\right]^{\frac{2}{z-1}}} \tag{21}$$

where U signifies the number of clusters, N represents the amount of information points, M_{ki} conforming fuzzy membership performance of x_k in class i and z is the degree of fuzziness of the algorithm. $v_i(y) = (v_1, v_2, \ldots, v_u)$ characterizes a matrix of anonymous cluster centers (prototypes) $v_i(y) \in R^P$ and $\|.\|$ personifies the Euclidean norm.

Possibilistic C-means clustering In fuzzy C-means clustering, primarily the number of clusters (U) is described. For instance, if the number of cluster is equal to two, it forms two clusters. Then, FCM selects two random customers and treats them as cluster center. It finds the fuzzy membership (M_{ki}) distance between each cluster center and each customer. Depending on the membership function, there may be fifty percentage of it to first cluster and fifty percentage of it to the second cluster. This prompts the issue of outlier focuses that are equidistant from the two clusters and can be given equal participation in both clusters. When it is required, such focuses must be given low or no enrollment in both the clusters. To beat this issue of outlier focuses, Krishnapuram and Keller [34] proposed another clustering algorithm known as possibilistic C-means (PCM), which resolves this issue. The ordinary FCM could be enhanced. The objective function and fuzzy membership (typicality matrix) of the PCM clustering are furnished by means of Eqs. (22) and (23), as follows: .

$$J_M = \sum_{k=1}^{N}\sum_{i=1}^{U}(T_{ki})^z \times \|x_k - v_i(y)\|_A^2 + \sum_{i=1}^{U}\gamma_i \sum_{k=1}^{N}(1 - T_{ki})^z \tag{22}$$

$$T_{ki} = \frac{1}{1 + \left[\frac{\|x_k - v_i(y)\|_A}{\gamma_i}\right]^{\frac{1}{(z-1)}}}, \quad 1 \le i \le U, \quad 1 \le k \le N \tag{23}$$

$z > 0$ is a user defined constant. The value of $D^2(x_k, v_i(y))$ may even be zero here, unlike FCM making PCM overcome the problem of singularity that FCM suffers from. Therefore with appropriate values of γ_i , PCM can significantly resolve the issue of outliers and noise points.

Possibilistic fuzzy C-means clustering The hybridization of fuzzy C-means clustering algorithm (FCM) and possibilistic C-means clustering algorithm (PCM) is possibilistic fuzzy C-means algorithm. Pal and Bezdek [36] utilized the fuzzy values of the FCM and the characteristic value of the PCM to design become an improved clustering algorithm named FPCM (fuzzy possibilistic C-means). Conversely, the restraint consistent to the amount of typicality values of total information to a cluster must be equivalent to that which causes glitches, principally for a large information set. For evading this issue, Pal et al. [37] proposed another clustering algorithm known as possibilistic fuzzy C-means (PFCM) which reduce this restraint. The objective function of the PFCM clustering is furnished by means of Eq. (24), as follows .

$$J_M = \sum_{k=1}^{N}\sum_{i=1}^{U}(aM_{ki}^{z} + bT_{ki}^{z} \times \|x_k - v_i(y)\|_A^2 + \sum_{i=1}^{U}\gamma_i\sum_{k=1}^{N}(1 - T_{ki})^{\eta} \tag{24}$$

Subject to the constraints $\sum_{i=1}^{U} M_{ki} = 1, \ \forall k,$ and $0 \le T_{ki} \le 1$. Here, $a > 0, b > 0, \eta > 1, z > 1$ and J_M is the objective performance.

4.4.2 Setup II

Researches confirm that the hybrid clustering- classification technique is more effective than single classification techniques [39, 41, 43]. Therefore, in the experiments, the proposed CCP results are compared with existing classification algorithms such as DT, KNN, SVM, NB and ANN. Classification is a technique in which similar information is grouped together and represented with the help of class labels. The initial process involved in classification is dividing the entire dataset into testing and training partitions. The rules for the classification are framed using the training datasets, and the test is carried out on the testing dataset.

Decision tree The decision tree classification algorithm proposed by Quinlan in 1993 is based on the divide and conquer strategy [21, 38]. The algorithm consists of the following steps. First the information gain for the entire attributes set is calculated; then, a binary tree begins from the attribute gaining maximum information gain. The tree creation breaks up until there are no more attributes. After creating a tree, the classification protocols are put together based on every path laid between the root and every leaf.

K-nearest neighbor K-nearest neighbor is one of the most convenient classification methods in data mining [21, 38]. The K-Nearest Neighbor cataloging procedure includes the following steps. First initialize the count of adjoining neighbor, that is K. Then, discover the distance between every indefinite test instance and all identified training instances. Every indefinite test instance value is projected by a majority poll by its K adjoining neighbors. In this effort, K is taken to be 3.

Support vector machine SVM supports Gaussian radial centered kernel functions that are proposed by Boser, Vapnik and Guyon in 1992. The nonlinear and linear data models are classified by SVM [21, 38]. The churn prediction problem considered in this work is linearly distinguishable, since it follows a binary level classification simply stating whether subscribers or customers churn or non-churn from the subscription of a company's plan or from the usage of a product. These two classes representing the customers are partitioned by means of a hyperplane. The classification task of SVM is evaluated as using Eq. (25)

$$d(X^T) = \sum_{i=1}^{l} y_i\alpha_i X_i X^T + b_0 \tag{25}$$

X^T is a notation in a test case, where y_i is the class tag belonging to support vector X_i; b_0 and α_i are numeric constants. l indicates the quantity of evaluating support vectors.

Naive Bayes Naive Bayes classifier designed on the basis of Bayes' theorem is a statistical classifier [21, 38]. It calculates the posterior probability $P(H/Y)$ using posterior likelihood trained on $Y[P(Y/H)]$, prior probability of $H[P(H)]$ and prior probability of $Y[P(Y)]$ as represented in Eq. (26)

$$P(H/Y) = \frac{P(Y/H)P(H)}{P(Y)} \tag{26}$$

The above equation can be written as in Eq. (27)

$$\text{Posterior} = \frac{\text{likelihood} \times \text{Prior}}{\text{evidence}} \tag{27}$$

Naive Bayes assumes all evidences to be conditionally independent, because the denominator does not depend on H and the value of feature Y is given, so that the evidence is effectively constant. The numerator is equivalent to the posterior probability $P(H/Y)$ as represented in Eq. (28)

$$P(H/Y) = P(Y_1|H) \times P(Y_2|H) \times \cdots \times P(Y_n|H) \times P(H) \tag{28}$$

4.4.3 Setup III

In this experiment, the proposed hybrid model is compared with other existing hybrid classification algorithms. The projected technique of PPFCM-ANNLM, PPFCM-ANNSCG, PPFCM-ANNRP, PFCM-ANNLM, FCM-ANNLM is associated with the existing hybrid models WKM-IL (weighted K-means-inductive learning) [39], K-means-DT [41], DT-ANN [42], SOM-ANN [43] and KNN-LR [44]. In this hybrid model, all fuzzy clustering algorithms are hybrid with the classifiers. Three sets of analysis are carried out on this proposed work using setup III.

1. Analysis of this proposed PPFCM algorithm hybrid with other artificial neural networks like scaled conjugate gradient backpropagation (SCG), resilient backpropagation (RP).
2. An analysis in which clustering algorithm (PPFCM, PFCM, FCM) hybrid with Levenberg–Marquardt (LM) produces a better churn prediction
3. The proposed hybrid model is compared with existing hybrid models [39, 41–44].

WKM-IL In [39], the scheme shows an weighted K-means clustering algorithm and a classic rule inductive method (FOIL) [10, 40]. The weighted K-means algorithm is based on path analysis. The model has three important terms; close instances, close cluster and close subclassifier. In the first step, the data are grouped together using weighted K-means algorithm. Then, each test is predicted based on the number of close instances which is dependent on minimum distance between test instance and samples. If the cluster possesses maximum number of close instances, then the cluster is called close cluster. If all the close cluster samples are churned, then the test instance outcome is churn; if all the close cluster samples are non-churn, then the test instance outcome is non-churn. When all the close cluster samples are a combination of churn and non-churn, then the close cluster data are given to the subclassifier (first-order inductive rule learning). It generates some positive rules if the test instance satisfies that rule; then, the test instance is churn; otherwise, it is non-churn.

K-means-DT Hung et al. [41] created two hybrid models which depend on three existing strategies such as K-means, decision tree and backpropagation neural network. The main arrangement shows K-means clustering and decision tree consolidated, and in the second model, decision tree and backpropagation neural networks were joined. Finally, they assessed the above models utilizing Hit ratio and LIFT. The outcome showed that the decision tree and the neural system procedures can convey precise forecast which is demonstrated by utilizing client attributes like interactive qualities, company communication qualities and client household attributes [41].

DT-ANN (SePI) Lee and Lee [42] proposed SePI (Segmentation by Performance Information), which includes three models: main model, discrimination model and support model. The decision tree is considered as the main model, which out performs in single-level classification. The discrimination model demonstrates and utilizes the result of the main model. The support model in which artificial neural network is used make use of the information which is incorrectly predicted by main model. Hence, the key idea of their work is the data which are incorrectly predicted by main model can correctly predicted by the support model [42].

SOM-ANN Tsai et al. [43] created two hybrid models, which depend on two existing strategies such as self-organizing map and backpropagation artificial neural network. In the first model, self-organizing map clustering and backpropagation artificial neural network were consolidated, and in the second model, backpropagation neural networks were joined itself. The three sets of test samples are carried out based on the American telecom churn prediction dataset. One is found to be general, and the other two are fuzzy based. Churn prediction is done based on these hybrid SOM neural network with good accuracy[43].

KNN-LR Zhang et al. [44] proposed a hybrid model which joins the K-nearest neighbor and logistic regression, and they assessed result utilizing two measurements: accuracy and receiver operating characteristic (ROC) curve. The proposed model yields 92.55% of accuracy with good area under curve (AUC) value. The accuracy of the projected model is compared with existing learning algorithm such as C4.5 and RBF [44].

4.5 Results and discussion

4.5.1 Result for setup I

The main objective of clustering algorithm is to decrease the sum of square error (SSE). Consequently, SSE is utilized to assess the clustering functions. 10 samples are used to train the models, and all the samples are used in functional assessment. It is used for associating the clustering methods reflected in this work. Figure 2 displays various clustering solutions generated using the proposed and

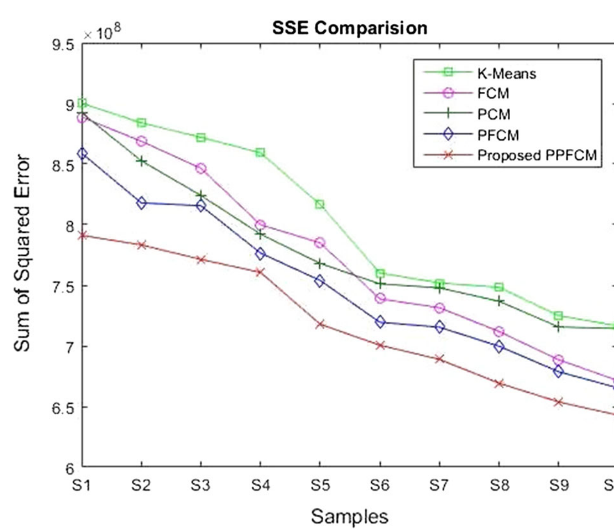


Fig. 2 SSE performance of clustering process ($U = 2$)

prevailing algorithm. It also displays the clustering functions on the basis of different statistical investigation. In this figure, the horizontal axis represents the samples based on various churn rates, whereas the vertical axis represents the SSE.

In clustering module, the probabilistic possibilistic fuzzy C-means (PPFCM) clustering algorithm has been used. The centroid of the information is obtained using the PPFCM algorithm. There are only two options such as churned or not churned. Hence CCP can be looked upon as a binary classification problem ($U = 2$) The SSE values of the five clustering methods are compared. One can comprehend from Fig. 2 that this projected method yields improved clustering solutions than the ones formed with the help of prevailing approaches. Additionally, the figure also illustrates that the solution produced with the help of all fuzzy clustering has lower SSE value than K-means clustering. The minimum SSE value elucidated the effectiveness of the clustering procedure

Table 3 displays the TC value of all samples after altering the number of clusters (U). The high TC value is acquired, which reflects the consistency and solidity of the anticipated model.

The clustering result thus obtained by the existing PFCM and the proposed PPFCM is shown in Fig. 3a, b, respectively. As the churn forecasting is foreseen as a binary classification problem, all the customers are assumed to be a part of the two separated clusters (cluster1 and cluster2). The Duke university dataset has 170 attributes, and to better represent the quality of clustering, the information gain technique is utilized to identify the two best features (da_Mean, drop_blk_mean). These two features are represented by X-axis and Y-axis. The first cluster is represented in red and second cluster is represented in blue. The center of first cluster is denoted as 'green + symbol,' and the center of second cluster is denoted as 'black + symbol'. Figure 3b demonstrates high accuracy in the separation of customers compared to the existing PFCM as shown in Fig. 3a.

4.5.2 Result for setup II

In this segment, the experimental outcomes of classification algorithm are discussed. The investigation setup is discussed in Sect. 4.4.2. Figure 4 displays the comparisons of the classification functions of projected classifier along with the various entrenched classification modeling performances. The probabilistic possibilistic fuzzy C-means clustering (PPFCM) together with an LM training algorithm has been used for foreseeing the churn and non-churn information in classification module. The number of cluster $U = 2$ is set. The horizontal axis characterizes the samples based on various churn rates, and the vertical axis characterizes the overall predicted accuracy. It is comprehended from Fig. 4 that the projected method yields results with improved accuracy than the prevailing approaches. Additionally, the figure also illustrates that the solution produced with the help of decision tree and artificial neural network possesses higher accuracy value than the k-nearest neighbor algorithm and the support vector machine. The maximum accuracy value elucidated the effectiveness of the classification procedure.

Moreover, Table 4 shows the performance analysis of the various classification techniques based on the accuracy. In this projected method, an improved accuracy of 94.62% is obtained when compared with the existing ones which were found to be 83.28% for decision tree, 64.45% for KNN classifier, 83.63% for exhausting SVM classifier, 85.55% for NB classifier and 86.26% for single-based

Table 3 True churn performance using proposed PPFCM by varying cluster size ($U = 2-10$)

Sample	Performance metric (%)	#Cluster (U)								
		2	3	4	5	6	7	8	9	10
S1	TC	0.6190	0.6022	0.6417	0.6264	0.6428	0.6315	0.6610	0.6162	0.6195
S2	TC	0.6527	0.6505	0.6111	0.6333	0.6082	0.6329	0.6129	0.6282	0.6428
S3	TC	0.6530	0.6895	0.6968	0.6730	0.6505	0.6627	0.6521	0.6469	0.6875
S4	TC	0.6632	0.7150	0.6944	0.7232	0.6888	0.7578	0.7708	0.6700	0.6931
S5	TC	0.7010	0.7401	0.7500	0.7510	0.7785	0.8032	0.7586	0.7401	0.7157
S6	TC	0.8260	0.7792	0.7727	0.7960	0.8041	0.8300	0.8615	0.7702	0.7714
S7	TC	0.8292	0.8823	0.8666	0.8492	0.8301	0.8708	0.8045	0.8644	0.8103
S8	TC	0.9215	0.9016	0.9491	0.9326	0.8627	0.8857	0.9250	0.8751	0.8623
S9	TC	0.9318	**0.9787**	0.9572	0.9214	0.8888	0.8974	0.9596	0.8810	0.9029
S10	TC	**0.9486**	0.9742	**0.9765**	**0.9548**	**0.9565**	**0.9642**	**0.9600**	**0.9389**	**0.9790**

Bold values indicate the maximum true churn values obtained by varying cluster size (U = 2–10)

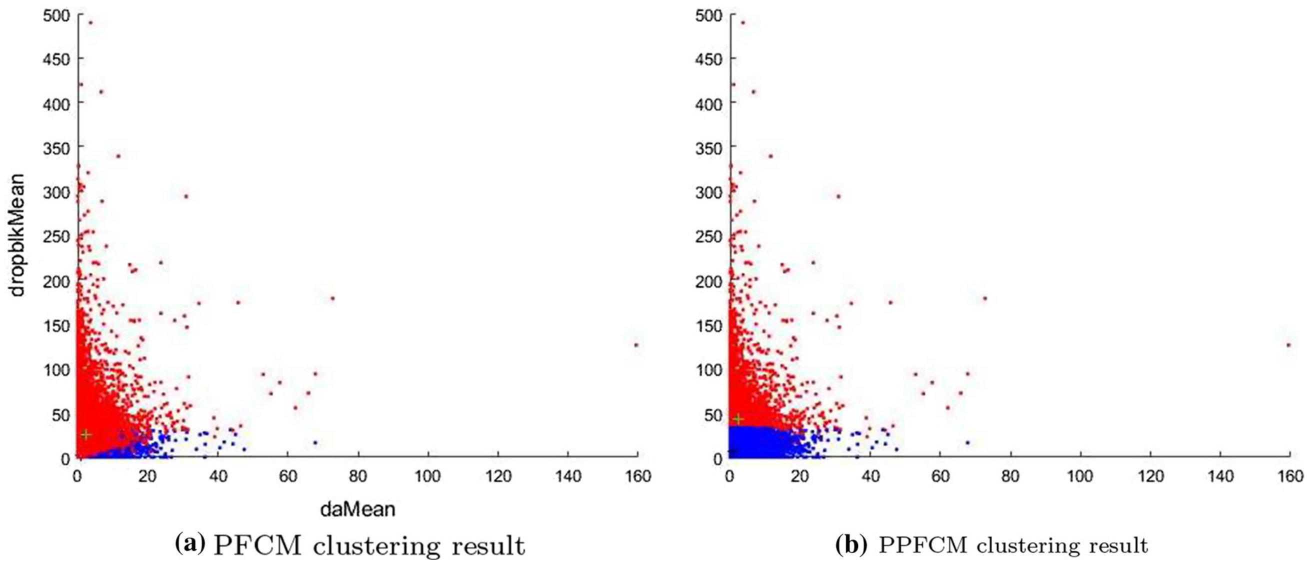


Fig. 3 Clustering results comparison of PFCM with proposed PPFCM

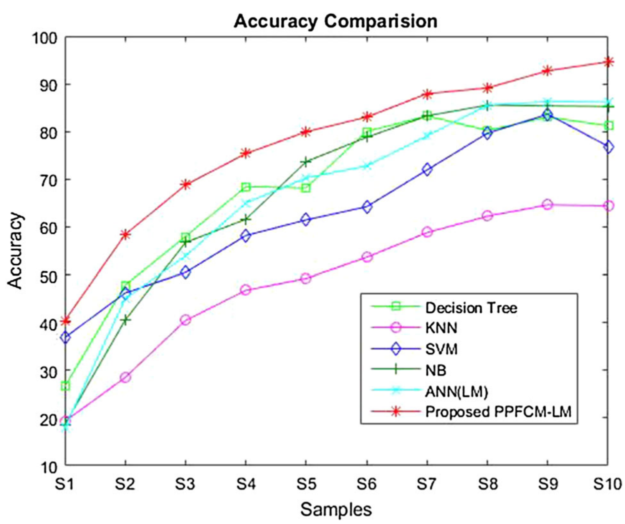


Fig. 4 Comparative study of classifiers based on Accuracy

ANN. The sample S10 is found to have better accuracy than other 9 samples.

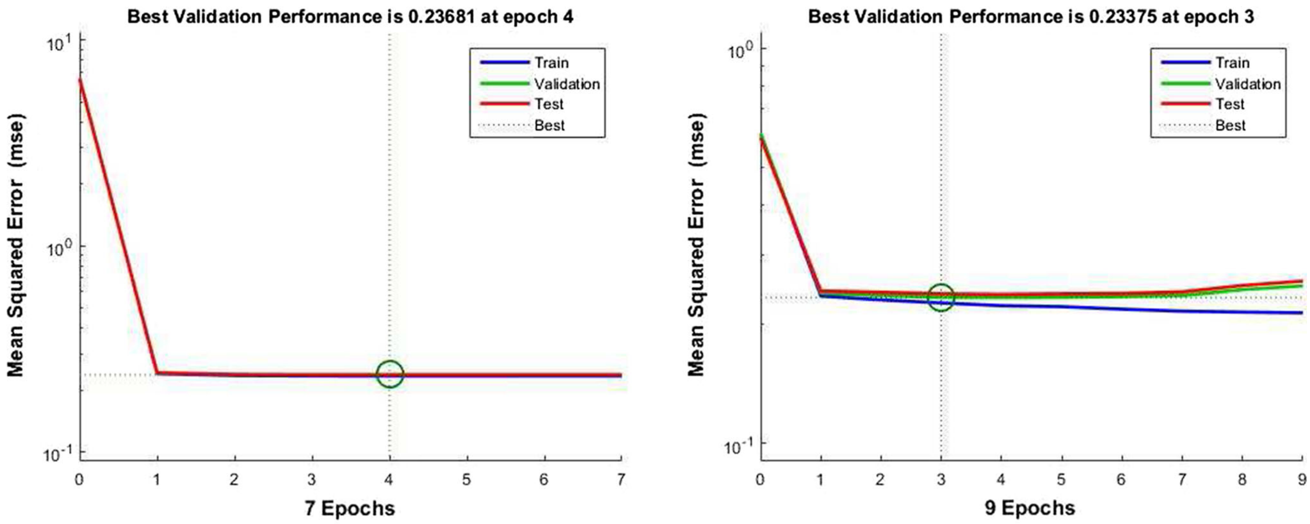
The neural network learning process generated by single ANN and the proposed PPFCM-ANN is shown in Fig. 5a, b, respectively. The number of epochs is represented along the *X*-axis, and mean squared error is represented along *Y*-axis. Figure 5b delivers less mean squared error result compared to the existing single ANN as shown in Fig. 5a.

Further, Tables 5, 6, 7 and 8 illustrate the numerical functions of classification performance of projected classifier along with other entrenched classification technique on the basis of the true churn, false churn, positive predicted value and negative predicted value. The prognostic accuracy is on the basis of the high true churn with low false churn. The true churn designates the proportion of churn instances that were appropriately prophesied as churn and false churn, which is the amount of non-churn

Table 4 Classification accuracy (CA) of proposed PPFCM-LM with other classifiers

Sample	Performance metric (%)	Classifiers					
		DT	KNN	SVM	NB	ANN(LM)	PPFCM-LM ($U = 2$)
S1	CA	26.60	19.29	36.86	18.59	17.95	**40.44**
S2	CA	47.78	28.48	46.20	40.51	44.94	**58.40**
S3	CA	57.99	40.44	50.47	56.74	53.92	**68.79**
S4	CA	68.42	46.75	58.20	61.61	65.02	**75.37**
S5	CA	68.20	49.24	61.47	73.70	70.28	**79.95**
S6	CA	80.00	53.64	64.24	78.79	72.78	**83.02**
S7	CA	**83.28**	58.81	71.94	83.28	79.09	**87.93**
S8	CA	80.24	62.24	79.65	**85.55**	85.54	**89.14**
S9	CA	83.04	64.62	**83.63**	85.38	**86.26**	**92.74**
S10	CA	81.21	**64.45**	76.88	85.26	86.13	**94.62**

Bold values indicate the maximum true churn values obtained by varying cluster size (U = 2–10)



(a) Single ANN error performance

(b) PPFCM-ANN error performance

Fig. 5 The neural network learning process generated by single the ANN and the proposed PPFCM-ANN

Table 5 Comparative study of classifiers based on true churn value

Sample	Performance metric (%)	Classifiers					
		DT	KNN	SVM	NB	ANN (LM)	PPFCM-LM ($U = 2$)
S1	TC	21.58	13.70	32.53	13.01	12.33	**27.21**
S2	TC	41.16	18.77	38.63	32.13	37.18	**45.62**
S3	TC	49.04	28.35	40.23	47.13	55.28	**58.27**
S4	TC	56.76	32.93	42.65	49.59	62.20	**69.37**
S5	TC	59.28	35.50	46.32	64.50	63.20	**72.59**
S6	TC	73.02	39.53	46.51	70.23	71.02	**84.28**
S7	TC	73.50	46.00	57.50	82.50	73.16	**95.43**
S8	TC	79.51	49.73	74.05	89.73	88.89	**94.85**
S9	TC	82.07	54.44	84.02	**91.12**	91.12	**95.68**
S10	TC	**84.42**	**60.39**	**84.98**	90.26	**91.96**	**96.84**

Bold values indicate the maximum true churn values obtained by varying cluster size (U = 2–10)

Table 6 Comparative study of classifiers based on false churn value

Sample	Performance metric (%)	Classifiers					
		DT	KNN	SVM	NB	ANN (LM)	PPFCM-LM ($U = 2$)
S1	FC	**00.00**	**00.00**	**00.00**	**02.35**	**00.00**	**00.56**
S2	FC	01.13	02.56	00.00	04.13	00.00	**00.86**
S3	FC	01.73	05.17	03.45	03.55	03.90	**01.15**
S4	FC	03.90	09.09	04.28	04.71	03.90	**01.74**
S5	FC	03.12	17.71	05.08	04.17	04.17	**01.99**
S6	FC	06.96	20.00	05.61	05.62	06.69	**02.73**
S7	FC	11.11	22.22	06.67	15.56	06.09	**03.95**
S8	FC	11.69	22.73	13.64	19.48	16.15	**05.15**
S9	FC	12.72	25.43	16.77	20.23	18.50	**06.22**
S10	FC	21.35	32.69	19.27	18.75	21.43	**07.57**

Bold values indicate the maximum true churn values obtained by varying cluster size (U = 2–10)

Table 7 Comparative study of classifiers based on positive predicted value

Sample	Performance metric (%)	Classifiers					
		DT	KNN	SVM	NB	ANN (LM)	PPFCM-LM ($U = 2$)
S1	PPV	**100.0**	**100.0**	**100.0**	**98.34**	87.36	**100.0**
S2	PPV	98.28	98.11	100.0	97.15	**91.81**	**100.0**
S3	PPV	99.22	96.10	98.13	97.38	82.45	**97.84**
S4	PPV	98.00	92.05	73.42	96.73	86.86	**100.0**
S5	PPV	97.74	82.83	98.17	97.39	82.65	**97.33**
S6	PPV	95.15	78.70	97.09	96.18	78.97	**95.73**
S7	PPV	91.38	75.41	92.74	88.71	75.93	**95.63**
S8	PPV	88.31	72.44	86.71	84.69	73.64	**93.74**
S9	PPV	85.81	67.65	83.04	81.48	72.82	**92.50**
S10	PPV	76.02	60.00	75.00	79.43	74.15	**92.63**

Bold values indicate the maximum true churn values obtained by varying cluster size (U = 2–10)

Table 8 Comparative study of classifiers based on negative predicted value

Sample	Performance metric (%)	Classifiers					
		DT	KNN	SVM	NB	ANN (LM)	PPFCM-LM ($U = 2$)
S1	NPV	08.03	07.35	09.22	07.30	40.06	**06.35**
S2	NPV	18.50	14.45	18.66	17.18	58.46	**18.31**
S3	NPV	30.00	22.73	26.42	29.59	69.16	**40.22**
S4	NPV	42.77	29.79	16.28	38.31	76.19	**44.31**
S5	NPV	47.94	34.65	43.12	52.87	81.11	**51.98**
S6	NPV	64.85	41.44	49.34	63.01	84.83	**65.05**
S7	NPV	74.53	49.30	59.72	76.51	87.80	**63.53**
S8	NPV	73.51	56.13	73.48	86.71	88.83	**88.97**
S9	NPV	80.75	62.62	**84.21**	90.20	91.51	**90.38**
S10	NPV	**86.29**	**68.06**	78.28	**91.23**	**93.10**	**92.43**

Bold values indicate the maximum true churn values obtained by varying cluster size (U = 2–10)

instances that were incorrectly projected as churn. The positive predicted value is a proportion of the number of true positive and summation of true positive and false positive. The negative predicted value is a proportion of the number of true negative and summation of true negative and false negative. In this projected method, the improved true churn rate of 96.84% is acquired with less false churn rate of 07.57% which is 84.42% and 21.35% for consuming decision tree, 60.39% and 32.69% by means of KNN classifier, 84.98% and 19.27% for exhausting SVM classifier, 91.12% and 18.75% for NB and 91.96% and 21.43% for single-based ANN.

Chi-square test for Independence Chi-square is an evaluation performed to calculate how carefully the experimental data suitable the anticipated data. If small the chi-squared value, null hypothesis fixed by us can be well accepted. If the value is found to be large, our null hypothesis can be negated and recognize that some key factor plays a role. In this section explains the Chi-square test for the proposed (PPFCM-LM) algorithm with the number of cluster $U = 2$ and the churn rate between the training and test data equal to 50% (S10). The null and alternative hypothesis will be: H0: There is no association between actual and predicted value. H1: There is an association between actual and predicted value. Tables 9 and 10 show the observed and expected value for calculating the Chi-square test for the proposed model. To

Table 9 Observed values for calculating Chi-square statistics

	Predict		
	Churn (C)	Non-churn (NC)	Total
Actual			
Churn (C)	24,000	781	24781
Non-churn (NC)	1909	23,310	25,219
Total	25,909	24,091	50,000

Table 10 Expected values for calculating Chi-square statistics

	Predict		
	Churn (*C*)	Non-churn (*NC*)	Total
Actual			
Churn (*C*)	12,841.02	11,939.98	24,781
Non-Churn (*NC*)	13,067.98	12,151.02	25,219
Total	25,909	24,091	50,000

calculate the expected value from the observed value for a particular cell, the following formula is used: [(corresponding row total customers × corresponding column total customers)/total no. of customers]

When a comparison is made between one sample and another, the DOF equals (number of columns − 1) × (number of rows − 1) excluding the rows and column containing the total, where DOF is the degrees of freedom. In this churn prediction problem, $\text{DOF} = (2 - 1)(2 - 1) = 1$. Chi-square value can be calculated by Eq. (19) described above, and the value is 39903.1292. Excel has a function CHIDIST (Chi-square value, DOF) that calculates the probability from the calculated chi-square value and the given DOF; in this problem, DOF is equal to 1. The original *P* value found from the Chi-square value is equal to 0.000000. If we want to test our hypothesis at 1% level of significance, then our predetermined alpha level of significance is 0.01. Looking into the Chi-square distribution table with 1 degree of freedom in the row and 0.01 significance level in the column shows the critical χ^2 (6.635) value which is lesser than our calculated χ^2 (39903.1292) value. Since a *P* value of 0.000000 is lesser than the conventionally accepted significance level of 0.01 (i.e., $P < 0.01$), the null hypothesis is rejected or in other words the alternate hypothesis is accepted and it is concluded that the model is very significant and there is a strong association between actual and predicted value.

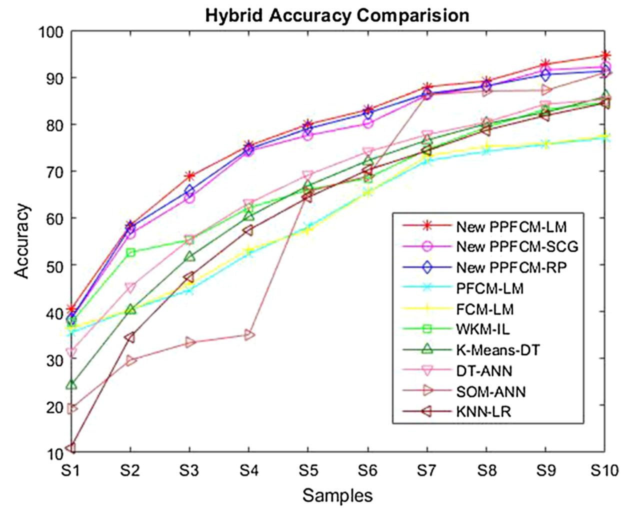


Fig. 6 Comparative study of hybrid classifiers based on accuracy ($U = 2$)

4.5.3 Result for setup III

In this segment, the investigation outcomes are elucidated on the basis of the hybrid model. Figure 6 associates the classification accuracy function among the anticipated hybrid model and other hybrid classifiers as designated in Sect. 4.4.3. The number demonstrates that the projected hybrid model is functioning better when compared to the other four proposed hybrid classification approaches such as PPFCM-ANNSCG, PPFCM-ANNRP, PFCM-ANNLM, FCM-ANNLM with existing hybrid models WKM-IL (weighted K-means inductive learning) [38], K-means DT [39], DT-ANN [40], SOM-ANN [10] and KNN-LR [41]. In hybrid module, proposed PPFCM clustering technique and

Table 11 Classification accuracy (CA) of proposed PPFCM-LM with other proposed 4 hybrid models ($U = 2$)

Sample	Performance metric (%)	Hybrid models				
		PPFCM-ANNLM	PPFCM-ANNSCG	PPFCM-ANNRP	PFCM-ANNLM	FCM-ANNLM
S1	CA	**40.44**	38.33	38.42	35.46	36.54
S2	CA	**58.40**	56.58	57.98	40.38	40.35
S3	CA	**68.79**	64.28	65.82	44.59	45.89
S4	CA	**75.37**	74.23	74.67	52.34	53.27
S5	CA	**79.95**	77.59	79.03	58.09	57.26
S6	CA	**83.02**	80.06	82.29	65.46	65.42
S7	CA	**87.93**	86.12	86.46	72.15	73.15
S8	CA	**89.14**	87.99	88.11	74.18	75.25
S9	CA	**92.74**	91.53	90.54	75.65	75.81
S10	CA	**94.62**	92.15	91.28	77.00	77.44

Bold values indicate the maximum true churn values obtained by varying cluster size (U = 2–10)

Table 12 Classification accuracy (CA) of proposed PPFCM-LM with other existing hybrid models ($U = 2$)

Sample	Performance metric (%)	Hybrid models					
		PPFCM-ANNLM	WKM-IL	K-MEANS-DT	DT-ANN	SOM-ANN	KNN-LR
S1	CA	**40.44**	37.58	24.23	31.49	19.30	11.03
S2	CA	**58.40**	52.65	40.29	45.30	29.67	34.57
S3	CA	**68.79**	55.29	51.72	55.42	33.37	47.52
S4	CA	**75.37**	62.13	60.28	63.12	35.09	57.43
S5	CA	**79.95**	65.96	66.89	69.15	65.62	64.38
S6	CA	**83.02**	68.47	72.18	74.07	68.89	70.21
S7	CA	**87.93**	74.52	76.56	77.73	86.22	74.29
S8	CA	**89.14**	79.31	80.06	80.39	87.04	78.69
S9	CA	**92.74**	83.12	82.46	84.24	87.16	81.81
S10	CA	**94.62**	84.57	85.84	85.22	90.96	84.47

Bold values indicate the maximum true churn values obtained by varying cluster size (U = 2–10)

Table 13 Classification accuracy (CA) of proposed PPFCM-LM by varying cluster size ($U = 2-10$)

Sample	Performance metric (%)	# Cluster (U)								
		2	3	4	5	6	7	8	9	10
S1	CA	40.44	41.61	41.03	42.30	**44.21**	43.25	40.22	41.57	42.95
S2	CA	58.40	61.29	61.13	**63.33**	60.32	63.29	58.05	62.82	59.28
S3	CA	68.79	65.21	**69.68**	67.40	65.95	66.27	68.95	64.69	68.75
S4	CA	75.37	**77.08**	73.52	72.35	71.98	75.78	76.50	69.00	73.31
S5	CA	**79.95**	75.86	75.04	75.01	77.75	80.32	77.01	74.01	79.57
S6	CA	**83.02**	81.15	79.27	82.60	80.41	83.00	82.56	79.02	80.14
S7	CA	87.93	**88.23**	86.66	84.93	86.01	87.08	82.45	86.44	81.03
S8	CA	89.14	**92.50**	91.91	88.26	88.07	88.57	90.16	87.51	86.23
S9	CA	**92.74**	92.35	91.72	92.14	87.88	89.74	91.85	88.10	90.29
S10	CA	**94.62**	93.70	92.64	91.48	91.65	93.02	93.47	93.89	91.90

Bold values indicate the maximum true churn values obtained by varying cluster size (U = 2–10)

ANNLM training neural network are used for the prediction procedure. The horizontal axis and the vertical axis signify the charges for samples and accuracy individually. At this time, this projected method attains the supreme accuracy of 94.62%, which is relatively elevated in comparison with other five hybrid classifiers. (1) It is observed from Fig. 6 that prediction based on PPFCM along with artificial neural network like scaled conjugate gradient backpropagation (PPFCM-SCG), resilient backpropagation (PPFCM-RP) performed better than other techniques and only a slight variation was observed compared to PPFCM-LM. (2) Levenberg–Marquardt (LM), a gradient descent method for training the neural nets by backpropagation, hybrid with PPFCM predicts churn better than FCM, PFCM. (3) This proposed approach produces better results than WKM-IL, K-means-DT, DT-ANN, SOM-ANN and KNN-LR. Tables 11 and 12 shows the performance analysis of the hybrid model based on accuracy. It is comprehended from the three investigation consequences that the

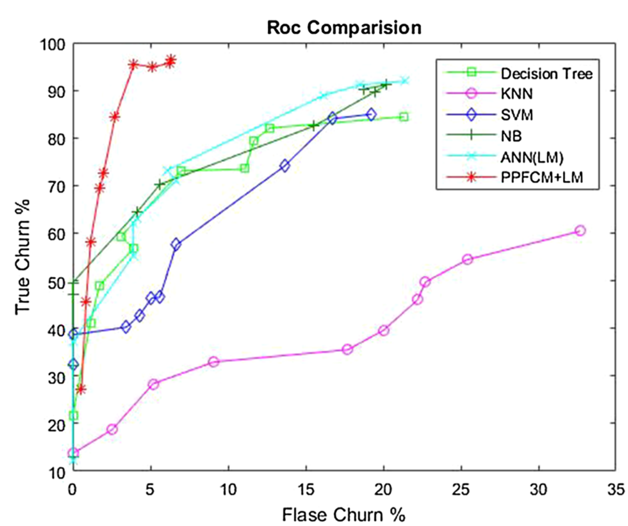


Fig. 7 Receiver operation characteristic (ROC) curve for classification

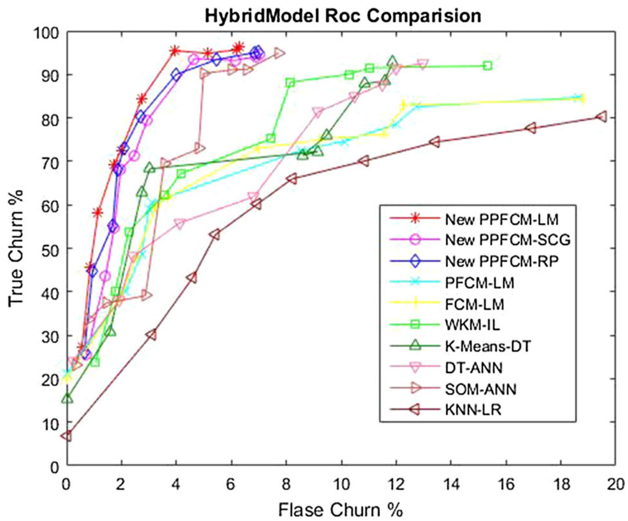


Fig. 8 Receiver operation characteristic (ROC) curve for hybrid models

hybrid model is an improvement over the solitary model. Table 13 displays the prediction functions of all samples on the basis of altered number of cluster (U). The high accuracy value which reflects the consistency and solidity of the anticipated model is acquired.

Figure 7 displays the ROC curve for the existing classifier and the proposed hybrid approaches, which with the help of a TC-FC pairs to establish the prognostic function. If the amount of TC is superior and the rate of FC is less, it will improve the consequences. The horizontal axis and the vertical axis characterize the rates of false positive and true positive correspondingly. The best prediction method has a point in upper left corner. The ROC curve of the proposed hybrid fuzzy clustering and the artificial neural network are present in upper left corner unlike other classifiers.

Figure 8 displays the ROC curve for the existing hybrid approaches and the proposed hybrid approaches, which with the help of a TP-FP pairs to establish the prognostic function. The horizontal axis and the vertical axis characterize the rates of false-positive and true-positive rate correspondingly. (1) It is observed from the figure that PPFCM along with artificial neural network like scaled conjugate gradient backpropagation (PPFCM-SCG), resilient backpropagation (PPFCM-RP) and Levenberg–Marquardt (PPFCM-LM) are closer to each other. (2) Levenberg–Marquardt (LM) hybrid with PPFCM gives the ROC value better than FCM, PFCM. (3) This proposed approach produces better ROC result than WKM-IL [38], K-means DT [39], DT-ANN [40], SOM-ANN [10] and KNN-LR [41].

4.6 Time complexity and speed of convergence

CPU time is the total time taken for executing the instructions of a program by the CPU. Space complexity is the amount of storage required for an algorithm. Table 14 compares the performance of the various algorithms discussed previously in Sects. 4.4.1, 4.4.2, 4.4.3 in terms of

Table 14 Time, space and speed of convergence for all algorithm used in three setups

S. No.	CCP models	DM techniques	CPU time (s)	Memory (MB)	Speed of convergence
1	Clustering	K-means	41.1	2828.9	14
		FCM	67.6	2594.3	66
		PCM	58.2	2597.1	50
		FPCM	66.8	2659.0	53
		PPFCM	112.9	2659.1	82
2	Classification	DT	94.8	2696.9	–
		KNN	550.4	2693.6	–
		SVM	729.8	2982.6	–
		NB	44.2	2782.5	–
		ANN	52.1	2627.2	–
3	Hybrid models	PPFCM-LM	874.8	2693.2	87
		PPFCM-SCG	899.0	2396.8	91
		PPFCM-RP	882.6	2458.9	93
		PFCM-LM	950.3	2587.4	71
		FCM-LM	972.5	2697.0	83
		WKM-IL	1095.7	2430.0	29
		K-means DT	203.9	2473.4	13
		DT-ANN	160.8	2696.9	–
		SOM-ANN	789.4	2693.2	–
		KNN-LR	785.2	2693.6	–

CPU time measured in seconds, space complexity measured in MB and the convergence iteration of the algorithm (speed convergence). As seen from the comparison, the hybrid approach takes more time than the single approach due to the added overhead of performing clustering before classification. But this is only a necessary trade-off, in order to achieve a better accuracy.

5 Conclusion

In this study, the customer churn prediction (CSP) is explained on the basis of the probabilistic possibilistic fuzzy C-means clustering (PPFCM) and artificial neural network (ANN). The proposed hybrid model comprises of two modules clustering and classification. In the clustering module, the data are grouped based on the PPFCM algorithm, and for the classification module, the artificial neural network (ANN) is used. The three groups of experimentations were approved on telecom datasets. First set of experimentations confirmed that the PPFCM clustering can maintain an improved information partitioning consequence; the second group of experimentations assessed the classification consequences and associated it with other renowned classification methods; the last set of research associates the projected hybrid model scheme with numerous other lately projected hybrid organization methods. All the results consequently display that the hybrid model-based learning scheme is very encouraging and that it overtakes the current models.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Mattison R (2001) Telecom churn management: the golden opportunity. APDG Publishing, Fuquay Varina
2. Payne A, Frow P (2005) A strategic framework for customer relationship management. J Mark 69(4):167–176
3. Reinartz W, Krafft M, Hoyer WD (2004) The customer relationship management process: its measurement and impact on performance. J Mark Res 41(3):293–305
4. Neslin SA, Gupta S, Kamakura W, Lu J, Mason CH (2006) Defection detection: measuring and understanding the predictive accuracy of customer churn models. J Mark Res 43(2):204–211
5. Van den Poel D, Lariviere B (2004) Customer attrition analysis for financial services using proportional hazard models. Eur J Oper Res 157(1):196–217
6. El-Zehery AM, El-Bakry HM, El-Ksasy MS (2013) Applying data mining techniques for customer relationship management: a survey. Int J Comput Sci Inf Secur 11(11):76
7. Reinartz WJ, Kumar V (2003) The impact of customer relationship characteristics on profitable lifetime duration. J Mark 67(1):77–99
8. Lin SC, Tung CH, Jan NY, Chiang DA (2011) Evaluating churn model in CRM: a case study in Telecom. J Converg Inf Technol 6(11):192200
9. Sharma A, Panigrahi D, Kumar P (2013) A neural network based approach for predicting customer churn in cellular network services. arXiv preprint arXiv:1309.3945
10. Huang Y, Huang B, Kechadi MT (2011) A rule-based method for customer churn prediction in telecommunication services. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, pp 411–422
11. Hwang H, Jung T, Suh E (2004) An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Syst Appl 26(2):181–188
12. Larivire B, Van den Poel D (2005) Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Syst Appl 29(2):472–484
13. Wei CP, Chiu IT (2002) Turning telecommunications call details to churn prediction: a data mining approach. Expert Syst Appl 23(2):103–112
14. Xia GE, Jin WD (2008) Model of customer churn prediction on support vector machine. Syst Eng Theory Pract 28(1):71–77
15. Dietterich TG (2000) Ensemble methods in machine learning. Mult Classif Syst 1857:1–15
16. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley, New York
17. Huang B, Kechadi MT, Buckley B (2012) Customer churn prediction in telecommunications. Expert Syst Appl 39(1):1414–1425
18. De Bock KW, Van den Poel D (2011) An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. Expert Syst Appl 38(10):12293–12301
19. Ali G, Artrk U (2014) Dynamic churn prediction framework with more effective use of rare event data: the case of private banking. Expert Syst Appl 41(17):7889–7903
20. Dejaeger K, Verbeke W, Martens D, Baesens B (2012) Data mining techniques for software effort estimation: a comparative study. IEEE Trans Softw Eng 38(2):375–397
21. Keramati A, Jafari-Marandi R, Aliannejadi M, Ahmadian I, Mozaffari M, Abbasi U (2014) Improved churn prediction in telecommunication industry using data mining techniques. Appl Soft Comput 24:994–1012
22. Kim K, Jun CH, Lee J (2014) Improved churn prediction in telecommunication industry by analyzing a large network. Expert Syst Appl 41(15):6575–6584
23. Au WH, Chan KC, Yao X (2003) A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Trans Evol Comput 7(6):532–545
24. Coussement K, Van den Poel D (2008) Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques. Expert Syst Appl 34(1):313–327
25. Mozer MC, Wolniewicz R, Grimes DB, Johnson E, Kaushansky H (2000) Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. IEEE Trans Neural Netw 11(3):690–696
26. Zhang X, Zhu J, Xu S, Wan Y (2012) Predicting customer churn through interpersonal influence. Knowl Based Syst 28:97–104
27. Verbeke W, Martens D, Baesens B (2014) Social network analysis for customer churn prediction. Appl Soft Comput 14:431–446
28. Khashei M, Hamadani AZ, Bijari M (2012) A novel hybrid classification model of artificial neural networks and multiple linear regression models. Expert Syst Appl 39(3):2606–2620

29. Yeshwanth V, Raj VV, Saravanan M (2011) Evolutionary churn prediction in mobile networks using hybrid learning. In: Twenty-fourth international FLAIRS conference
30. Bose I, Chen X (2009) Hybrid models using unsupervised clustering for prediction of customer churn. J Organ Comput Electr Commer 19(2):133–151
31. Le M, Gabrys B, Nauck D (2014) A hybrid model for business process event and outcome prediction. Expert Syst 34(5):e12079
32. Duke University (2005) Case studies, presentations and video modules. http://www.fuqua.duke.edu/centers/ccrm/datasetsdownload.html#data
33. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers, New York
34. Krishnapuram R, Keller JM (1993) A possibilistic approach to clustering. IEEE Trans Fuzzy Syst 1(2):98–110
35. Iyigun C (2007) Probabilistic distance clustering. Wiley, New York
36. Pal NR, Pal K, Bezdek JC (1997) A mixed c-means clustering model. In: Proceedings of the sixth IEEE international conference on fuzzy systems, no 1. IEEE, pp 11–21
37. Pal NR, Pal K, Keller JM, Bezdek JC (2005) A possibilistic fuzzy c-means clustering algorithm. IEEE Trans Fuzzy Syst 13(4):517–530
38. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, Zhou ZH (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37
39. Huang Y, Kechadi T (2013) An effective hybrid learning system for telecommunication churn prediction. Expert Syst Appl 40(14):5635–5647
40. Quinlan JR (1990) Learning logical definitions from relations. Mach Learn 5(3):239–266
41. Hung SY, Yen DC, Wang HY (2006) Applying data mining to telecom churn management. Expert Syst Appl 31(3):515–524
42. Lee J, Lee J (2006) Customer churn prediction by hybrid model. Adv Data Min Appl 4091:959–966
43. Tsai CF, Lu YH (2009) Customer churn prediction by hybrid neural networks. Expert Syst Appl 36(10):12547–12553
44. Zhang Y, Qi J, Shu H, Cao J (2007) A hybrid KNN-LR classifier and its application in customer churn prediction. In: IEEE International conference on systems, man and cybernetics, 2007, ISIC. IEEE, pp 3265–3269