Cluster Comput (2019) 22:S10757–S10768
https://doi.org/10.1007/s10586-017-1172-1

An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing

J. Vijaya[1] · E. Sivasankar[1]

Received: 21 April 2017 / Revised: 29 August 2017 / Accepted: 4 September 2017 / Published online: 19 September 2017

Abstract Churn prediction in telecom has gained a huge prominence in the recent times due to the extensive interests exhibited by the stakeholders, large number of competitors and huge revenue losses incurred due to churn. Predicting telecom churn is challenging due to the voluminous and sparse nature of the data. This paper presents a technique for the telecom churn prediction that employs particle swarm optimization (PSO) and proposes three variants of PSO for churn prediction namely, PSO incorporated with feature selection as its pre-processing mechanism, PSO embedded with simulated annealing and finally PSO with a combination of both feature selection and simulated annealing. The proposed classifiers were compared with decision tree, naive bayes, K-nearest neighbor, support vector machine, random forest and three hybrid models to analyze their predictability levels and performance aspects. Accuracy, true positive rate, true negative rate, false positive rate, Precision, F-Measures, receiver operating characteristic and precision-recall plots were used as performance metrics. Experiments reveal that the performance of metaheuristics was more efficient and they also exhibited better predictability levels.

Keywords Churn prediction · Feature selection · Classifier · Metaheuristics · Particle swarm optimization · Simulated annealing

1 Introduction

Customer relationship management (CRM) is the key to any successful management in any organization. Maintaining customers is one if the major issues faced by organizations today. This is caused in part due to increased competition and in major due to dissatisfaction of customers in the current organization. Retaining customers is one of the major key factors in the successful operation of an organization. It has been identified that getting new customers is five to six times more expensive than retaining old customers [1]. Further, loss of customers will lead not only to revenue loss, but also in the reduction of brand loyalty and the company's morale.

Customer churn means customers cancel their service from a firm, withholding existing customers of a firm plays an important role to increase the overall revenue of the firm and retains the good name of the firm in the competitive market [2]. The major issue in predicting churn is that there is no single reason for a customer to churn. Several accumulated reasons, usually lead to customer churn. Identifying these reasons is quite complex, as they depend on both the customer's personal views and the services of the companies that are being utilized by the customers. The major requirement of organizations in this domain is to predict churn before it happens, to identify the major causes of churn and to predict the countermeasures that can be used to avoid churn. Even though all these can be performed with the data available in the organizations, it is the nature of this data that poses a huge drawback to the prediction mechanism. Domains in which churn prediction can be applied include telecom services, product based businesses and tangible or an intangible service oriented areas [3–5].

The remainder of this paper is structured as follows; Sect. 2 presents the literature survey, Sect. 3 presents the issues

✉ J. Vijaya
406114003@nitt.edu

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

related to churn data, Sect. 4 present the advantages of using metaheuristics in such a scenario, Sect. 5 discusses the utility of PSO for churn prediction and analyzes the proposed variants of PSO, Sect. 6 presents the experimental setup and results discussion and Sect. 7 concludes the study.

2 Literature survey

Churn data analysis and churn prediction have been performed by experts manually in earlier times. However the globalization of markets has led to a very different picture, where manual analysis is impossible. Hence the current decade has witnessed several contributions to this domain. This section discusses some of the recent contributions in this domain.

Data mining techniques are used to solve the problem of Customer churn. In earlier days, churn prediction problem was handled by supervised learning techniques like Decision Tree, Logistic Regression, Artificial Neural Networks, Random Forest, Naive Bayes, Support Vector Machines and statistical classifier (KNN) to identify churn [6–12]. Many churn prediction researches show that single model based classification does not produce satisfactory results and hence researchers switched on to hybrid models, which is a combination of two or more classifiers. Here the idea is that, support classifier predicting the customers correctly, which were previously predicted incorrectly by the main classifier. Zhang et al proposed a hybrid model that hybridizes KNN using Logistic Regression [13]. The model evaluated using receiver operating characteristic and accuracy. Tsai et al proposed a hybrid model that hybridizes Neural Networks (NN-NN) [14]. Khashei et al developed a hybrid model by combining Artificial Neural Network and Multiple Linear Regression. It gives better results compared to traditional single classifiers [15].

Current researches utilize both supervised and unsupervised techniques for predicting churn. Initially they segment the customers using clustering techniques and then each cluster is classified using predefined classifier. Indranil Bose et al chosen important churn prediction features like revenue related features and usage related features. Initially they segment the customers using partitional, hierarchical, neural network based clustering and then they developed two classification models. Classification is done by using Boosted decision tree [16]. Ying Huang et al proposed a hybrid model by combining weighted K-Means clustering algorithm and First Order Inductive Learning (FOIL) classification algorithm. They proposed modified K-Means algorithm that uses path analysis [17]. Yeshwanth et al. proposed a hybrid model by combining Evolutionary algorithm GA and Decision Tree [18].

Some of the churn prediction model concentrated on ensemble models and feature selection techniques. A telecom based churn prediction technique employing minimum redundancy maximum relevance (mRMR) was presented by Idris et al. [19]. This technique employs feature selection as a preprocessing component and uses an ensemble of Random Forest, Rotation Forest, RotBoost and DECORATE techniques to predict churn. A dynamic transfer ensemble model that uses feature selection as a part of preprocessing in customer churn prediction was presented by Xiao et al. [20]. This technique incorporates the process of feature selection as a preprocessing phase in-order to eliminate unnecessary entries. Further feature selection also reduces the size of the data, hence making the analysis process simpler for the classifier. Feature selection is actually carried out in two phases. The first phase uses the target domain to perform feature selection using GMDH-type neural network and the next phase iteratively analyses the source domain to shortlist entries and adds it to the feature list. The classifiers are trained and the best classifier for the test data is dynamically selected. The major downsides of this technique are that the model is time consuming due to the process of repeated feature selection and also because of the usage of several classifiers. Further, it assumes that the source and target domains are in the same feature space, which is practically not possible. Another ensemble based technique that performs churn prediction as a one step process was presented by Xiao et al. [21]. This technique utilizes multiple classifier ensemble techniques and cost sensitive learning to perform predictions.

An agent based technique that operates on churn occurring due to unsatisfactory upgrades from 3G to 4G in cellular networks is presented by Alessandro et al. [22]. This technique uses the upgrade data and the satisfactory levels of customers to identify churn. It is also an extended method that recommends countermeasures to avoid churn and hence improve the satisfactory levels of the customer. A real time churn prediction model based on sociometric clique and the social status concept was proposed by Drafting et al. [23]. This technique uses the concept of energy in the diffusion model as an opinion of users. The prediction scheme proposed here can be extended to predict for a smaller subset of the users also rather than constrained to a single user. A study on predicting churn from logistics industry was presented by Chen et al. [24]. This is an analytical model that considers length, regency, frequency, monetary and profit (LRFMP) as the major components in determining churn. It was identified that length, regency and monetary components have a significant effect on churn, while frequency became the top predictor only when the variability of these three values were limited. It was also identified that profit was never a significant predictor. This paper also discusses the managerial implication with additional insights on the predictor variables.

A PSO and neighborhood cleaning rules based technique to perform customer prediction were proposed in [25]. A similar technique using Random Forest and PSO was proposed in [26]. From the analyzed techniques it was observed that most of the current churn prediction models utilize ensemble techniques. These techniques prove to be costly and time consuming, especially when performed on telecom churn data which is huge. This paper presents a technique for the telecom churn prediction that employs particle swarm optimization (PSO) and proposes three variants of the hybrid PSO model for churn prediction namely, PSO incorporated with feature selection as its pre-processing mechanism (PSO-FS), PSO embedded with simulated annealing (PSO-SA) and finally PSO with a combination of both Feature Selection and Simulated Annealing (PSO-FSSA). The proposed classifiers were compared with DT, NB, KNN, SVM, RF and three hybrid models to analyze their predictability levels and performance aspects. Accuracy, TPR, TNR, FPR, Precision, F-Measures, ROC, PR plots were used as performance metrics.

3 Issues with churn data

3.1 Data volume

Churn data is a complete collection of customer history and their churn details. Data used for mining is usually made up of a subset of the collected data. This is because very old data tends to become obsolete. Unlike such requirements, churn data must contain all the historical data. Churn prediction is to be performed on the entire data. Every customer represents a different training prospect; hence eliminating a customer will lead to inappropriate training. Hence, eliminating records on the grounds of its obsolete nature is out of the question. This leads to voluminous data. The volume of a churn data is proportional to the number of customers and the number of product lines associated with it since the establishment of the organization.

3.2 Data imbalance

The data is said to be imbalanced if one class in the data outnumbers the other classes to a very large extent. The dominating class is called the major class and the other classes in the data are called minor classes. Imbalance ratios are usually of the form 100:1, 1000:1 or even 100000:1. Classifiers, by default consider the data provided to it to be balanced. Hence it provides equal prominence to all the classes. Imbalance affects the training phase of the classifiers to a large extent. The major class is appropriately trained due to the sufficient number of training entries. Due to imbalance, there are only a few numbers of minor class entries, hence it gets undertrained. This leads to wrong predictions. The major issue is that the imbalance levels in churn data cannot be determined earlier. It varies largely between organizations. Hence churn detection systems must be capable of identifying the imbalance levels and apply appropriate balancing techniques on the data such that the classifier is sufficiently trained in all the classes.

3.3 High attributes in a dataset

Churn data, as discussed earlier contain customer details representing their product or scheme utilizations that are provided by the organizations. This paper has its major concerns on telecom churn prediction. On examining the telecom data, it was identified that the dataset contains attributes representing every service offered by the organization from the inception of the organization to current date. Representing every service requires several attributes depicting the resource utilization of that service. This leads to a huge increase in the number of attributes in the dataset. Hence every addition of service increases the entire dataset in terms of attribute. This leads to very high attribute levels even if the organization has a medium number of product lines.

3.4 Highly sparse nature of data

The high levels of the attributes discussed in the previous section are the major cause of the high sparse nature of the data. It was previously discussed that the churn dataset contains details about all the customers of an organization. Each customer has the probability of being associated with a minimum of one and a maximum of few products in the organization. But the attributes associated with the customer corresponds to all the products developed by the organization so far. Hence, apart from the associated entries, all the other entries tend to contain null values, leading to the highly sparse nature of the data.

3.5 Improper representation of flat file

Flat file representations are the most common representations of data. However the highly sparse nature of the data makes the churn files contain mostly null values leading to a huge data size with very few legitimate entries.

4 Use of metaheuristics in prediction

Metaheuristic is a higher-level procedure or heuristic that has been formulated to identify or select a heuristic that can provide an optimal solution (not best) to an optimization problem, especially with incomplete or imperfect information or limited computation capacity [27]. The near-optimality of the solution is a major drawback when it comes to the application of prediction. It has been observed for decades that a statistical algorithm overcomes the problem of near-optimality and provides optimal solutions, in other words, the best solution for the problem in hand. The current years witness a huge data explosion due to the automation of nearly every possible process. This data is information rich, hence can be mined to obtain valuable information.

The process of Churn detection falls under this category. The properties of churn discussed in the previous section clearly expose the complexity involved in the process of mining churn data. Applying data mining algorithms on such data tends to be disastrous, due to its inefficiency in handling such data. Not only that data mining techniques require their input data to be in specific formats, they also require the entire data to be present in-order to perform the prediction process. The complexity of such algorithms shoots up as the amount of data increases. Hence, in terms of scaling, most of the statistical prediction techniques fail. Hence opting for meta-heuristic techniques in such scenarios appear to be the best decision. The non-expectancy of complete data not only suits the current Big Data scenario, but also the problem of churn prediction, whose data tends to contain sparse huge data. This resolves the scalability issue associated with statistical data mining techniques.

The near optimality of results in metaheuristics is brought about by their basic nature to accept errors in their operational process. All metaheuristics were observed to embrace errors. This leads them to near optimal, but faster solutions. This is mainly due to the existence of errors and the embracing nature of the metaheuristic techniques to accept the errors [28]. Errors exist in three basic forms; approximation error, estimation error and optimization error [29]. Estimation error occurs during the model building phase. It is the error that occurs as a result of training the model with a finite amount of data. However, it is believed that with infinite data, every feature could be exactly mapped for accurate predictions. This error is common to all the prediction techniques. While the statistical techniques strive to overcome it, metaheuristics embraces this error in its basic architecture by accepting a data that is incomplete. The approximation error arises from the compromises that the user makes while building the model and the optimization error arises due to the choice of algorithm. Approximation error is built into metaheuristics, as the entire operation operates on probability. It has been identified that when it comes to huge data, conscious introduction of optimization error [30] leads to better results as opposed to small data. These properties of metaheuristics gravitated the choice of our prediction algorithm towards metaheuristics

Several metaheuristic techniques exist in literature and several other techniques are also being proposed. This section justifies the use of PSO for the process of churn prediction and how other metaheuristic algorithms are not appropriate candidates for churn prediction. Ant colony optimization (ACO) is a metaheuristic technique proposed by Dorigo in [31,32]. This technique is based on the movement of ants to the food source. Hence it has very high space complexities with low scalability levels. Churn data being implicitly large would not be the best operating scenario for ACO. Firefly algorithm [33] is a metaheuristic technique working on the flashing behavior of fireflies. This technique considered a single shared parameter, firefly intensity, hence it exhibits very low space complexities. However, in every iteration the intensity of a firefly needs to be compared with every other firefly to define the movement. Large datasets require a large number of fireflies, and large number of fireflies increases the number of comparisons exponentially. Churn data would increase the computational complexity of firefly algorithms to a huge extent; hence firefly algorithm is not a suitable metaheuristic technique. Bee colony optimization [34] and several of its variants were also analyzed and was identified to be very similar to ACO with similar space complexities.

PSO being a metaheuristic technique operates directly on the data provided and does not require substantial intermediate computations or memory requirements. Further, comparisons are also limited to a constant value (two comparisons). Hence PSO was identified to be the best technique suitable for operating on churn data.

5 Proposed churn prediction model

Figure 1 describes our proposed model for customer churn prediction. Using widely available data cleansing and preprocessing methods the collected orange data set is processed. Churn data being customer based data, has very high probabilities of containing imbalance nature. Imbalance in data tends to reduce the reliability of classifiers to a large extent. Random under sampling was used to reduce the size of the data and to eliminate the imbalance levels. The three data sets

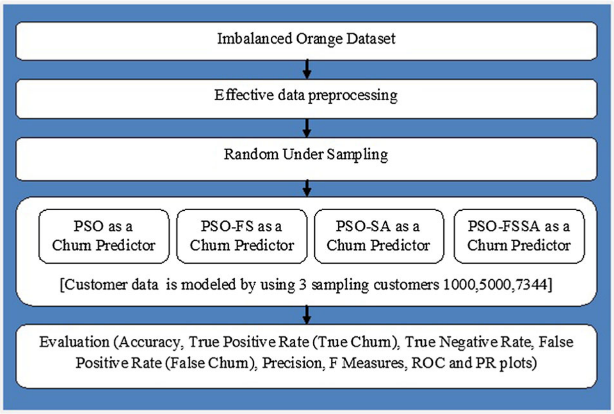
Imbalanced Orange Dataset
Effective data preprocessing
Random Under Sampling
PSO as a Churn Predictor
PSO-FS as a Churn Predictor
PSO-SA as a Churn Predictor
PSO-FSSA as a Churn Predictor
[Customer data is modeled by using 3 sampling customers 1000,5000,7344]
Evaluation (Accuracy, True Positive Rate (True Churn), True Negative Rate, False Positive Rate (False Churn), Precision, F Measures, ROC and PR plots)

Fig. 1 Block diagram of proposed churn prediction model

(1000 customers, 5000 customers, 7344 customers) obtained from sampling were utilized for churn prediction. The preprocessed data were passed to PSO and the three variants PSO-FS, PSO-SA, PSO-FSSA for predicting the customer churn. Finally the performance of the proposed model was analyzed in terms of Accuracy, TPR, TNR, FPR, Precision, F-Measures, ROC and PR plots. This approach of PSO based churn prediction is made up of three variants. PSO in its basic form is initially analyzed in terms of accuracy. This is followed by PSO incorporated with feature selection as its pre-processing mechanism (PSO-FS), PSO embedded with simulated annealing (PSO-SA) and finally PSO with a combination of both feature selection and simulated annealing (PSO-FSSA).

5.1 PSO as a churn predictor

Particle swarm optimization (PSO) [35,36] is a mathematical based compute methodology that solves a difficulty of recursively working to advance an applicant result with a view to an assured parameter called its measure of suitability. PSO a great multitude based metaheuristic methodology finds the solutions for the predictive difficulty by means of a group of applicant solutions called particles. Using a component called velocity the units entitled particles are motivated in a broader exploration space. This task is recursively carried out to make out the best result in the exploration space. This work makes use of such a particle based optimizer for the development of churn forecast. The process of PSO is carried out in three major phases, namely particle initialization, particle movement and convergence.

5.1.1 Particle initialization

Particles are given with initial values by particularly finding out the position using a methodology called function of constant dispersal. Particle dispersal is defined within the exploration place limitations. The count of the particles that can be used for a defined problem itself stands as a problem of optimization. By means of continuous checking the particle count for a given problem is fixed. Then the particle rapidity is found out in a random manner with the help of Eq. (1)

$$V_i \sim U(-|b_{up} - b_{lo}|, |b_{up} - b_{lo}|) \tag{1}$$

Here b_{up} is the upper bounds of the search space U and b_{lo} is the lower bounds of the search space U.

5.1.2 Particle movement

The rapidity credentials lay out the initiation of particles drive. Each and every particle drives conferring to its rapidity measure. Because of the random nature of rapidity, the particles are widely distributed in the exploration place. The particle paramount *(pbest)* and the global paramount *(gbest)* values are marked. Particle paramount marks the excellent solution found out by a particle from the recursive iterate values global paramount marks the excellent solution found out from the swarm that proves to be the dominating particle paramount values of all. The aptness of the particle provides the value of dominating particle paramount value. The current aptness fount out value is compared every time with the past paramount value. If it surpasses the previous value new value will be updated in the place of older paramount value. The new aptness is then compared with the present global paramount and the dominant value will be taken as the present global paramount.

5.1.3 Convergence

The particle new paramount *(pbest)* and the global best *(gbest)* values are triggered for adding velocity. The velocity is calculated based on Eq. (2)

$$V_{i,d} \leftarrow \omega V_{i,d} + \phi_p r_p \left(p_{i,d} - X_{i,d}\right) + \phi_g r_g \left(g_d - X_{i,d}\right) \tag{2}$$

Here r_p and r_g are the random numbers, $P_{i,d}$ and g_d are the new paramount *(pbest)* and the global paramount *(gbest)*, $X_{i,d}$ is the value current particle position, and the parameters ω, ϕ_p, ϕ_g are selected by the practitioner.

5.1.4 PSO classifier

The process of churn prediction is internally a classification system that identifies the current state of a customer using the properties of the customer. In general this can be expressed as follows. Given a database with two classes (representing churn and no churn) and N attributes, the process of classification is to identify the optimal centroid C in an N dimensional space for each class. A particle i is represented in Eq. (3) and a velocity component is represented in Eq. (4).

$$P_i = \left\{P_{i,1}, P_{i,2}, P_{i,3} - - - - P_{i,N}\right\} \tag{3}$$

With a velocity component

$$V_i = \left\{V_{i,1}, V_{i,2}, V_{i,3} - - - - V_{i,N}\right\} \tag{4}$$

The fitness of particle is computed by identifying the Euclidean distance between the training data and the current position of the particle using Eq. (5). Particle exhibiting the closest distance with the centroid in the N dimensional region is considered as the best solution for the current training data and hence the churn value of the best solution is identified to be the prediction of the training data.

$$\psi_i = \sum_{K=1}^{N} \sqrt{\left(P_{i,k} - \tau_k\right)^2} \tag{5}$$

where $P_{i,k}$ refers to the current position of particle I in the kth dimension and τ_k represents the value of the training data in its kth dimension.

5.2 PSO-FS as a churn predictor

This variant incorporates feature selection/elimination as a pre-processing component prior to the utilization of PSO for classification. Incorporating attribute elimination into the classification process has several benefits. The major benefit being reduction of data, which leads to faster predictions. Churn data, being a huge data can benefit from this process. Further, the sparse nature of the churn data makes several attributes useless. Such attributes can be effectively identified and eliminated by the system. Not all attributes contribute to the final solution. Some attributes even have a negative impact on the solution. Another advantage of a feature selection technique is that it can effectively identify such attributes and eliminate them, which in turn affects the accuracy of the algorithm in a positive note. Attribute selection is carried out using PSO based feature selection. The fitness of particle is computed by identifying the Euclidean distance between the training data and the current position of the particle.Fitness function is used to find the *pbest* value. The velocity of ith particle $V_{id}(t)$ is computed using the above Eq. (2)

$$S\left(V_{i,d}(t+1)\right) = \frac{1}{1 + e^{v_{i,d}(t+1)}} \tag{6}$$

$$X_{i,d}(t+1) = \begin{cases} 0 : \mathit{if}\ \mathit{randoms}() \geq S\left(V_{i,d}(t+1)\right) \\ 1 : \mathit{if}\ \mathit{randoms}() < S\left(V_{i,d}(t+1)\right) \end{cases} \tag{7}$$

The above Eqs. (6) and (7) explains the feature selection conditions. First step the velocity of each attribute is calculated. After that we have to apply the sigmoid function on the calculated velocity of each attribute. Suppose a sigmoid function is less than a random number generated between 0 to 1 then the corresponding attribute is not selected, otherwise the corresponding attribute is selected. Finally, we have select most informative attribute based on selection frequency.Algorithm 1 explains the steps involved in PSO based feature selection.The 19 attributes are selected in this process,which includes $Var24$, $Var35$, $Var57$, $Var65$, $Var73$, $Var126$, $Var143$, $Var144$, $Var173$, $Var189$, $Var196$, $Var203$, $Var205$, $Var208$, $Var210$, $Var211$, $Var218$, $Var221$, $Var227$.

Algorithm 1: Proposed Algorithm for PSO based feature selection

```
1.Attribute selection and evaluation using PSO based feature selection.
2.Attribute filtering to generate pruned dataset.
3.Initialize particle location and velocity.
4.Triggering particle acceleration.
5.For each particle
    a.Discretize particle coordinates to identify the nearest available node.
    b.Identify node fitness with respect to particles initial location
    c.Compare current fitness with pbest of particle.
        If particle best < current fitness
            Current fitness is consider as particle best
    d.Compare particle best and global best
        If global best < particle best
            Assign pbest to be the latest gbest
    e.Generate new particle velocity using

$V_{i,d} \leftarrow \omega V_{i,d} + \phi_p r_p \left(p_{i,d} - X_{i,d}\right) + \phi_g r_g \left(g_d - X_{i,d}\right)$
6.If the termination condition is not reached go to step4.
```

5.3 PSO-SA as a churn predictor

PSO-SA is a variant that hybridizes PSO to incorporate Simulated Annealing (SA) in the process of selecting the best solutions. PSO has the major shortcoming of getting struck in local optima. It was identified from the operating behavior of PSO that after every iteration, every particle compares its fitness value with its own *(pbest)* and the *(gbest)* to update itself and the swarm of the best solution. This process tends to gravitate to the algorithm towards local optima (if *(gbest)* contains a local optimal solution). In-order to eliminate this shortcoming and avoid frequent comparisons, SA is incorporated into PSO. Simulated Annealing is a probabilistic technique to identify the global optimum from the given set of solutions. This is a metaheuristic technique that performs global optimization on a large search space.

Operating nature of PSO-SA is similar to that of the regular PSO until the selection of *(pbest)*. This is usually followed by the selection of *(gbest)*. This variant modifies this process by identifying the *(pbest)* of all the particles and then utilizing SA to identify the *(gbest)*. This technique is particularly useful in swarms containing large number of particles. Due to the probabilistic nature of SA, the probability of getting struck in local optima reduces considerably in PSO-SA when compared to other techniques. Algorithm 2 explains the steps involved in PSO hybrid with SA and Algorithm 3 explains the function of Simulated Annealing

5.4 PSO-FSSA as a churn predictor

The final variant combines both Feature Selection and Simulated Annealing to provide a classifier with a reduced

Algorithm 2: Proposed Algorithm for PSO-SA

```
1.Search space boundary identification using foot data
2.For each particle j=1....p
    a.Initialize particle location using a uniform distribution
    b.Initialize particle velocity using the boundaries of the search space
    c.Initialize particle best and global best
3.Until the termination criterion is met perform the following
    a.For each particle j=1....p
        i.Generate random numbers $r_p$ and $r_g$ using the normal distribution
        ii.Identify the particle velocity using
$V_{j,d} \leftarrow \omega V_{j,d} + \phi_p r_p \left(p_{j,d} - x_{j,d}\right) + \phi_g r_g \left(g_d - X_{j,d}\right)$
        iii.Update the particle's position to P'
        iv.If particle best < current fitness
            Current fitness is considered as particle best
    b.global best ← SimulatedAnnealing(global best,particle best,p)
4.The best solution is global best
```

Algorithm 3: Simulated Annealing (global best,particle best,p)

```
    1.Let z = globalbest
    2.For j = 1 through p :
        a.T1 ← particle best$_j$
        b.Pick a random particle best ($p_b$), $z_{new} \leftarrow p_b$
        c.If P1(E(z), E(znew), T1) ≥ random(0, 1), shift to the new state:
            • $z \leftarrow z_{new}$
    3.Output: the final state z
P1 (e,e',T) was defined as 1 if e' < e and exp(-(e'-e)/T) otherwise.
```

selection rate of local optima and faster processing with accurate results.

6 Experimental setup and results

PSO and its variants were implemented in MATLAB version (2105 b) and experiments were conducted on DELL Precision T7600 workstation with 64 bit window machine consuming Intel core i5 processor along with speed 1.6 GHz and 8 GB RAM.

6.1 Data set

Orange Small and Orange Large data sets correspond to the French Telecom companys churn data [37]. This is a benchmark data, also provided as a part of the KDD 2009 challenge. A description of the Orange dataset is presented in Table 1. The major use of orange dataset is used to predict the probability of a customer to switch providers. It is also used to predict the propensity of customers to buy new products or services or upgrades.

Table 1 Data set description

Property	OrangeSmall	OrangeLarge
No. of attributes	230	15,000
No. of records	50,000	50,000
Missing values	60%	60%
No. of numerical attributes	190	14,740
No. of categorical attributes	40	260

6.2 Data preprocessing

Experiments were conducted on the Orange dataset. The Orange data have an imbalance ratio of 12.6. The majority of the customers about 46,328 are non-churn customers and about 3672 customers are churn customers.It was observed from the data set that contain 60% of missing values and 40 attributes are categorical. Since the fitness function operates only on numerical values, these values need to be normalized. Equilateral normalization was used to normalize the data.The missing values are replaced with the mean and mode of the corresponding attribute. Hence the data set was sampled and used for experiments. Random Sampling was used to reduce the size of the data and to eliminate imbalance levels. Logarithmic reduction was performed on the data and the data set was reduced to 1000 and 5000 records. An imbalance level of 12.6 was maintained in both the data sets. Random Under sampling was carried out on the Orange data to reduce the imbalance levels. The imbalance level of 1 was achieved by under sampling and the data set size was reduced to 7344 records, with equal number of records in each of the classes. These three datasets obtained from sampling were utilized for churn prediction. This marks the end of the preprocessing phase.

6.3 Evaluation criteria

Confusion matrix describes the performance of the classifier. This is shown in Table 2, where D_{11}—Churn customer predicted as Churn, D_{12}—Churn customer predicted as Non-Churn, D_{21}—Non-Churn customer predicted as Churn, D_{22}—Non-Churn customer predicted as Non-Churn. From the confusion matrix the following performance measures are find out that is represented in Eqs. (8)–(13)

Table 2 Confusion matrix for customer churn prediction

Actual	Predicted	
	Churn	NonChurn
Churn	D_{11}	D_{12}
NonChurn	D_{21}	D_{22}

$$\text{Accuracy} = \frac{D_{11} + D_{22}}{D_{11} + D_{12} + D_{21} + D_{22}} \tag{8}$$

$$\text{Truepositiverate} = \frac{D_{11}}{D_{11} + D_{12}} \tag{9}$$

$$\text{Falsepositiverate} = \frac{D_{21}}{D_{21} + D_{22}} \tag{10}$$

$$\text{Truenegativerate} = \frac{D_{22}}{D_{21} + D_{22}} \tag{11}$$

$$\text{Precision} = \frac{D_{11}}{D_{11} + D_{21}} \tag{12}$$

$$\text{F} - \text{Measure} = 2.\frac{\text{Precision.Recall}}{\text{Precision+Recall}} \tag{13}$$

6.4 Result and discussion

The pre-processed data were passed to PSO and the three variants and their performance were analyzed in terms of Accuracy, TPR, TNR, FPR, Precision, F-Measures, ROC and PR plots. These metrics were chosen to identify the performance of the classifiers from the perspective of both majority and minority classes. Comparison of the current technique with DT, NB, KNN, SVM, RF and three hybrid models proposed by Indranil et al. [16], Khashei et al. [15] and Ying et al. [17] were performed to measure the difference in performance levels of the metaheuristic classifier (PSO) and the variants proposed in this paper. A test data was selected based on hold out method to analyze the performance of the classifiers. This section observes the classifiers and their performance levels from several perspectives, namely data size, data imbalance levels, true prediction rates in terms of positive and negative classes and false prediction rates for each of the algorithm being used, and finally provides a consolidated view of each classifier.

6.4.1 Performance on Orange-1000

This section discusses the performance of the classifiers on the Orange dataset with 1000 records and the imbalance level of 12.6. Performance exhibited by algorithms on orange 1000 is presented in Table 3. It could be observed that the accuracy levels of PSO and its variants vary between 84.02 and 90.65%. The best accuracy levels were exhibited by PSO-FSSA, closely followed by PSO-FS and then by PSO-SA, PSO marked in bold letters. The performance of the base classifiers and the hybrid models are less compared to the proposed models, so the statistical models are not suitable for the small and imbalanced data set.

6.4.2 Performance on Orange-5000

This section discusses the performance of the classifiers on the Orange dataset with 5000 records and the imbalance level of 12.6. Performance exhibited by algorithms on orange 5000 is presented in Table 4. It could be observed that the accuracy levels of PSO and its variants vary between 86.74 and 94.08%. The best accuracy levels were exhibited by PSO-FSSA, closely followed by PSO-FS and then by PSO-SA, PSO marked in bold letters. The performance of the Orange 5000 has improved compared to the Orange 1000.The performance of the base classifiers and the hybrid models are less compared to the proposed models, so the statistical models are not suitable the imbalanced data set.

6.4.3 Performance on Orange-7344

This section discusses the performance of the classifiers on the Orange dataset with 7344 records and the imbalance level of 1. Performance exhibited by algorithms on orange 7344 is presented in Table 5. It could be observed that the accuracy levels of PSO and its variants vary between 89.51 and 96.33%. The best accuracy levels were exhibited by PSO-FSSA, closely followed by PSO-FS and then by PSO-SA, PSO marked in bold letters. The performance of the Orange 7344 has improved compared to the Orange 5000 and Orange 1000.Because the Orange 7344 consist of 7344 customers and the number of churn and non-churn customers are balanced. Accuracy comparison between different churn pre-

Table 3 Performance measure (Orange 1000)

	Proposed				Classifiers					Hybrid		
Performance metric (%)	PSO	PSO-FS	PSO-SA	PSO-FSSA	DT	NB	KNN	SVM	RF	K-MEANS-DT	WK-FOIL	ANN-MLR
Accuracy	**84.02**	**89.23**	**85.79**	**90.65**	76.45	79.66	79.42	81.08	80.07	78.57	77.81	81.10
TPR	87.69	90.00	88.23	92.92	76.98	82.10	79.52	81.60	80.29	79.44	78.21	81.15
TNR	71.79	85.51	75.75	77.41	69.49	62.50	78.57	76.47	78.35	70.93	74.07	80.73
FPR	28.20	14.48	24.24	22.58	30.50	37.50	21.42	23.52	21.64	29.06	25.92	19.26
Precision	91.20	96.77	93.75	95.98	97.09	93.90	87.43	96.87	87.70	96.01	96.63	96.63
F-measure	89.41	93.26	90.90	94.43	85.87	87.60	79.42	88.58	80.07	86.95	86.45	88.22

Table 4 Performance measure (Orange 5000)

	Proposed				Classifiers					Hybrid		
Performance metric (%)	PSO	PSO-FS	PSO-SA	PSO-FSSA	DT	NB	KNN	SVM	RF	K-MEANS-DT	WK-FOIL	ANN-MLR
Accuracy	**86.74**	**92.89**	**89.11**	**94.08**	75.75	80.02	80.37	83.43	82.24	84.26	83.07	85.44
TPR	86.26	93.75	88.70	93.84	77.65	80.93	81.01	83.95	82.96	84.96	84.48	85.57
TNR	89.74	90.24	91.24	94.87	78.43	73.58	75.72	80.32	78.04	80.71	75.73	84.42
FPR	10.25	09.75	08.75	05.12	21.56	26.41	24.27	19.67	21.95	19.28	24.26	15.17
Precision	98.12	96.77	98.12	98.38	96.32	95.56	96.03	96.19	95.68	95.68	94.77	90.68
F-measure	91.81	95.23	93.17	96.06	85.99	87.64	87.88	89.66	88.87	90.00	89.33	85.44

Table 5 Performance measure (Orange 7344)

	Proposed				Classifiers					Hybrid		
Performance metric (%)	PSO	PSO-FS	PSO-SA	PSO-FSSA	DT	NB	KNN	SVM	RF	K-MEANS-DT	WK-FOIL	ANN-MLR
Accuracy	**89.51**	**95.47**	**92.84**	**96.33**	81.53	82.01	84.02	85.32	85.56	89.94	93.01	91.36
TPR	89.00	96.14	93.15	97.20	81.32	81.76	83.89	85.44	85.22	90.68	94.78	92.94
TNR	91.56	93.51	91.83	91.22	83.01	83.63	84.73	84.72	86.90	84.61	83.82	81.96
FPR	08.43	06.48	08.16	08.80	16.98	16.36	15.26	15.27	13.09	15.38	16.17	18.03
Precision	97.17	97.71	97.39	98.45	97.09	97.09	96.76	96.45	96.32	97.67	96.89	96.82
F-measure	93.15	96.92	95.23	97.83	88.51	88.77	89.87	90.62	90.43	89.94	95.79	94.84

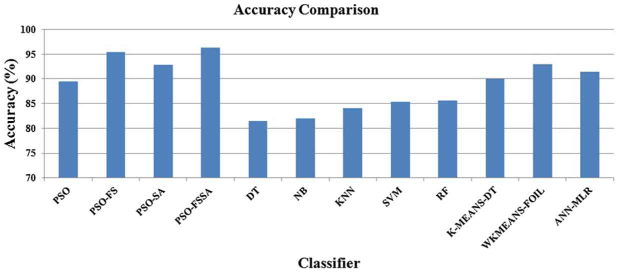

Fig. 2 Accuracy comparison between different churn predictor (PSO,PSO-FS,PSO-SA,PSO-FSSA,DT, NB, KNN, SVM, RF, K-Means-DT, WK-FOIL, ANN-MLR) based on Orange-7344

dictor (PSO,PSO-FS,PSO-SA,PSO-FSSA,DT, NB, KNN, SVM, RF, K-Means-DT, WK-FOIL, ANN-MLR) based on Orange-7344 is presented in Fig. 2.

6.5 Comparison analysis- ROC and PR plots (Dataset based)

This section presents the ROC,PR comparison of DT, NB, KNN, SVM, RF and three hybrid models proposed by Indranil et al. [16], Khashei et al. [15], Ying et al. [17] and variants of PSO proposed in this paper namely, PSO, PSO-FS, PSO-SA and PSO-FSSA. Performance observations are performed on dataset size.

ROC and PR plots for sampled Orange data containing 1000 records and data imbalance of 12.6 are presented in Fig. 3a, b. It can be observed from the ROC Plot that the true positive rate remained moderate and a medium level of false positives were exhibited by base classifier and hybrid classifiers compared to PSO and their variants and also observed that almost all the algorithms except Naive Bayes exhibits low false positive levels. In terms of true positive levels, it is observed that Naive Bayes exhibits high TPR levels (82.10%), but its reliability is abridged by its false positive levels. The PSO-FS and PSO-FSSA exhibit high true positive levels of 90.00, 92.92%, however, they exhibit false positive levels of 14.48, 22.58% making them effective for imbalanced data. The PR Plot exhibits very low levels of precision and recall. These measures indicate very low performance levels of 1000 customers. ROC and PR plots for sampled Orange data containing 5000 records and data imbalance of 12.5 are presented in Fig. 4a, b. It was observed from Fig. 4a that the regular PSO exhibits low true positive rates with moderate to high false positive levels compared to PSO embedded with PSO-FS, PSO-SA, PSO-FSSA. ROC and PR plots for sampled Orange data containing 7344 records and the data imbalance level of 1 are presented in Fig. 5a, b. It could be observed that all the statistical classifier exhibit an almost similar performance (attributed to the huge data size, affecting their predictability) while a performance boost

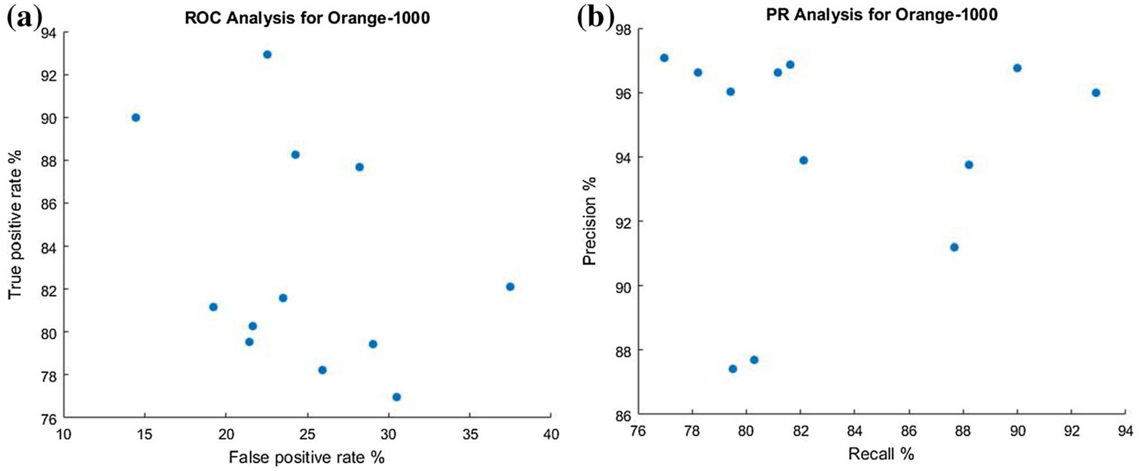

Fig. 3 Compare the ROC curve and PR curve for different churn predictor (PSO,PSO-FS,PSO-SA,PSO-FSSA,DT, NB, KNN, SVM, RF, K-Means-DT, WK-FOIL, ANN-MLR) on orange data set with 1000 customers. **a** ROC PLOT(Orange-1000), **b** PR PLOT(Orange-1000)

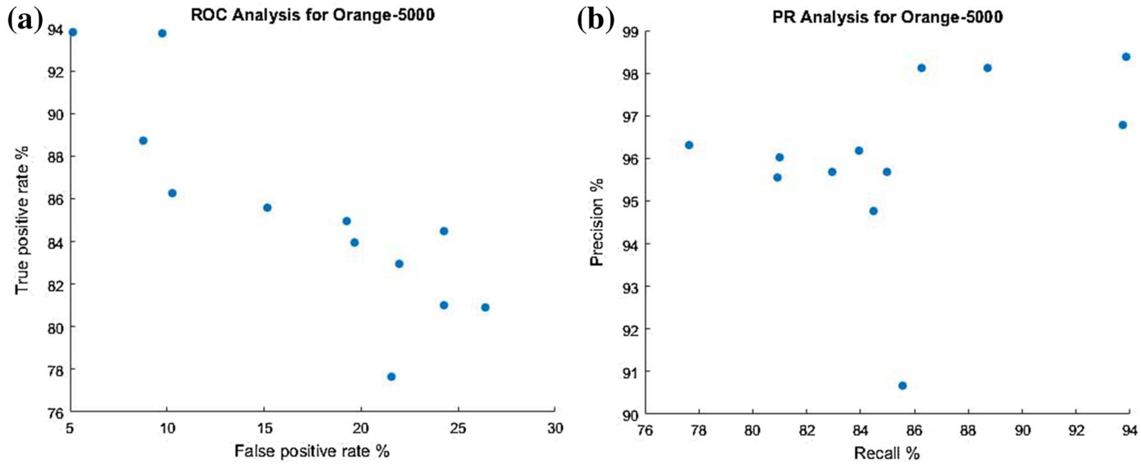

Fig. 4 Compare the ROC curve and PR curve for different churn predictor (PSO,PSO-FS,PSO-SA,PSO-FSSA,DT, NB, KNN, SVM, RF, K-Means-DT, WK-FOIL, ANN-MLR) on orange data set with 5000 customers. **a** ROC PLOT(Orange-5000), **b** PR PLOT(Orange-5000)

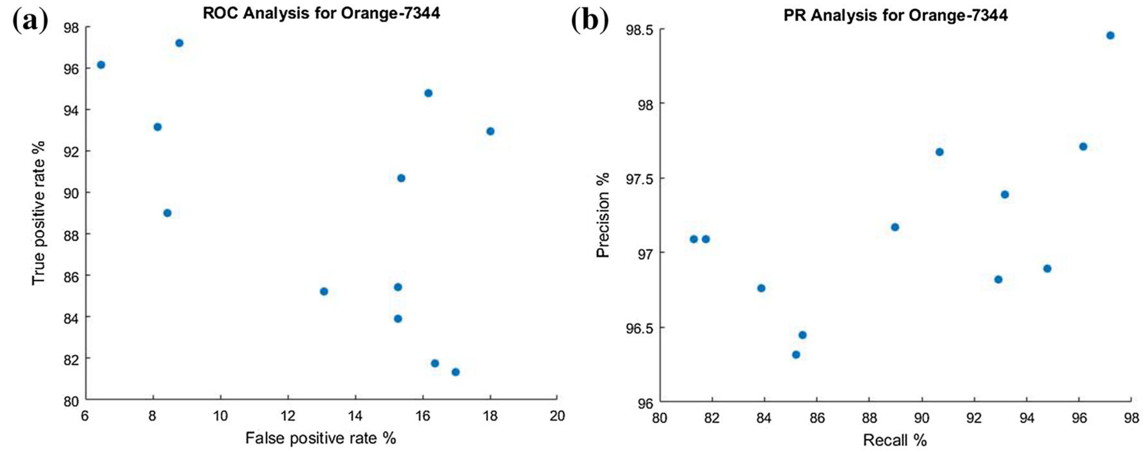

Fig. 5 Compare the ROC curve and PR curve for different churn predictor (PSO,PSO-FS,PSO-SA,PSO-FSSA,DT, NB, KNN, SVM, RF, K-Means-DT, WK-FOIL, ANN-MLR) on orange data set with 7344 customers. **a** ROC PLOT(Orange-7344), **b** PR PLOT(Orange-7344)

Table 6 Compare the Accuracy values (%) of the base classifier and PSO-FSSA techniques based on the benchmark datasets

Data set	# Customers	# Features	KNN	SVM	ANN	NB	DT	PSO-FSSA
Australian	690	15	68.51	87.66	44.21	73.50	89.78	91.28
Breast	699	11	45.80	30.25	76.89	99.57	95.93	97.15
German-credit	1000	25	83.95	96.23	79.41	74.12	100	98.47
Credit card-churn	30000	24	76.52	85.3	45.71	54.04	75.49	90.59
Japan-credit	690	16	49.36	86.10	64.04	70.40	86.92	89.99

was observed in PSO and the proposed PSO variants. This improved performance is attributed to the reduced imbalance levels. Scalability levels of PSO and its variants can also be observed from the above plots. Even with the increase in the data size, the robustness of PSO could be well observed here.

6.6 Other application prediction

So as to demonstrate that the proposed PSO-FSSA modelling technique is more broad and can be connected to numerous other classification or prediction applications, we connected this framework on 5 benchmark datasets, which are gathered from the UCI machine learning repository. The outcomes are organized in Table 6, demonstrate that the benchmark datasets pick up the maximum accuracy while applying the proposed model.

7 Conclusion

This paper presents an analysis of several statistical data mining algorithms for their use in churn prediction. The analysis was carried out in terms of Accuracy, TPR, TNR, FPR, Precision, F-Measures, ROC and PR plots. The major drawbacks of a churn data are its voluminous nature and imbalance levels. This paper analyzes the impact of these properties on the accuracy of the classifiers. It also discusses the importance of metaheuristic algorithms on such data states. An analysis is carried out using the PSO as a churn predictor. Three variants of PSO have been proposed and analyzed on the same criteria. It was observed that PSO and its variants perform well on huge imbalanced data. It was also observed that the variant incorporating Feature Selection as a pre-processing mechanism and Simulated Annealing as the local search algorithm performs well in imbalanced scenarios. Further the scalability associated with this algorithm makes it the best candidate for churn prediction. However, it was observed that this technique is more conservative in identifying classes. Future works will be based on enhancing this algorithm for faster and more efficient performance.

References

1. Bhattacharya, C.B.: When customers are members: customer retention in paid membership contexts. J. Acad. Marketing Sci. **26**(1), 31–44 (1998)
2. Dyche, J.: The CRM handbook: a business guide to customer relationship management. Addison-Wesley Professional, Boston (2002)
3. Nie, G., Rowe, W., Zhang, L., Tian, Y., Shi, Y.: Credit card churn forecasting by logistic regression and decision tree. Expert Syst. Appl. **38**(12), 15273–15285 (2011)
4. Gunther, C.C., Tvete, I.F., Aas, K., Sandnes, G.I., Borgan, O.: Modelling and predicting customer churn from an insurance company. Scand. Actuar. J. **2014**(1), 58–71 (2014)
5. Khan, A.A., Jamwal, S., Sepehri, M.M.: Applying data mining to customer churn prediction in an internet service provider. Int. J. Comput. Appl. **9**(7), 8–14 (2010)
6. Huang, S.Y., Yen, D.C., Wang, H.Y.: Applying data mining to telecom churn management. J. Expert Syst. Appl. **31**(3), 515–524 (2006)
7. Xia, G.E., Jin, W.D.: Model of customer churn prediction on support vector machine. J. Syst. Eng. Theor. Pract. **28**(1), 71–77 (2008)
8. Larivire, B., Poel, D.V.D.: Predicting customer retention and profitability by using random forests and regression forests techniques. J. Expert Syst. Appl. **29**, 472–484 (2014)
9. Burez, J., Van den Poel, D.: Handling class imbalance in customer churn prediction. J. Expert Syst. Appl. **36**(3), 4626–4636 (2009)
10. Huang, B.Q., Kechadi, M.T., Buckley, B.: Customer churn prediction in telecommunications. J. Expert Syst. Appl. **39**(1), 1414–1425 (2012)
11. Hwang, H., Jung, T., Suh, E.: An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. J. Expert Syst. Appl. **26**(2), 181–188 (2004)
12. Ruta, D., Dymitr, Nauck, D., Azvine, B.: K nearest sequence method and its application to churn prediction. In: Intelligent Data Engineering and Automated Learning IDEAL, pp. 207–215. Springer, Berlin (2006)
13. Zhang, Y., et al.: A hybrid KNN-LR classifier and its application in customer churn prediction. In: IEEE International Conference on Systems, Man and Cybernetics-ISIC 2007. IEEE (2007)
14. Tsai, C.F., Lu, Y.H.: Customer churns prediction by hybrid neural networks. J. Expert Syst. Appl. **36**(10), 12547–12553 (2012)
15. Khashei, M., Hamadani, A.Z., Bijari, M.: A novel hybrid classification model of artificial neural networks and multiple linear regression models. J. Expert Syst. Appl. **39**(3), 2606–2620 (2012)
16. Bose, I., Chen, X.: Hybrid models using unsupervised clustering for prediction of customer churn. J. Organizational Comput. Electron. Commer. **19**(2), 133–151 (2009)
17. Huang, Y., Kechadi, T.: An effective hybrid learning system for telecommunication churn prediction. J. Expert Syst. Appl. **40**, 5635–5647 (2013)

18. Yeshwanth, V., Raj, V.V., Saravanan, M.: Evolutionary churn prediction in mobile networks using hybrid learning. In: Proceedings of the twenty-fourth international Florida artificial intelligence research society conference, (FLAIRS), Palm Beach, Florida, USA, May 1820. AAAI Press (2011)
19. 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)
20. Xiao, Jin, et al.: Feature-selection-based dynamic transfer ensemble model for customer churn prediction. Knowl. Inf. Syst. **43**(1), 29–51 (2015)
21. Xiao, J. et al.: One-step classifier ensemble model for customer churn prediction with Imbalanced Class. In: Proceedings of the Eighth International Conference on Management Science and Engineering Management. Springer Berlin, vol. 281, pp. 843–854 (2014)
22. Alessandro, D., Johnson, L., Gray, D.: Consumer satisfaction versus churn in the case of upgrades of 3G to 4G cell networks. Marketing Lett. **26**(4), 489–500 (2015)
23. Droftina, U., Stular, M., Kosir, A.: A diffusion model for churn prediction based on sociometric theory. Adv. Data Anal. Classif. **9**(3), 341–365 (2015)
24. Chen, K., Hu, Y.H.: Predicting customer churn from valuable B2B customers in the logistics industry: a case study. Inf. Syst. e-Bus. Manag. **13**(3), 475–494 (2015)
25. Faris, H., et al.: Neighborhood cleaning rules and particle swarm optimization for predicting customer churn behavior in telecom industry. Int. J. Adv. Sci. Technol. **68**, 11–22 (2014)
26. Idris, A., Rizwan, M., Khan, A.: Churn prediction in telecom using random forest and PSO based data balancing in combination with various feature selection strategies. Comput. Electr. Eng. **38**(6), 1808–1819 (2012)
27. Bianchi, L., et al.: A survey on metaheuristics for stochastic combinatorial optimization. Nat. Comput. **8**(2), 239–287 (2009)
28. Dalessandro, B.: Bring the noise: Embracing randomness is the key to scaling up machine learning algorithms. Big Data **1**(2), 110–112 (2013)
29. Bousquet, O., Bottou, L.: The tradeoffs of large scale learning. Adv. Neural Inf. Process. Syst. **20**, 161–168 (2008)
30. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT-2010, pp. 177–186 Physica, Heidelberg (2010)
31. Dorigo, M., et al.: Ant system: an autocatalytic optimizing process (1991)
32. Dorigo, Marco, Gambardella, Luca Maria: Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Trans. Evolut. Comput. **1**(1), 53–66 (1997)
33. Yang, X.-S.: Firefly algorithm. Nat. Inspir. Metaheuristic Algorithms **20**, 79–90 (2008)
34. Pham, D.T., et al.: The Bees Algorithm, Technical note. Manufacturing Engineering Centre, Cardiff University, UK, pp. 1–57 (2005)
35. Kennedy, J.: Eberhart. RC: particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, pp. 1948–1942 (1995)
36. Shi, Y., Eberhart, R.: A modified particle swarm optimizer. Evolutionary Computation. In: The 1998 IEEE International Conference on Proceedings 1998. IEEE World Congress on Computational Intelligence. IEEE (1998)
37. http://www.kdd.org/kdd-cup/view/kdd-cup-2009/Data

J. Vijaya obtained her Bachelors degree in Computer Science and Engineering from MS University-Tirunelveli, India. Then she obtained her Masters degree in Computer Science and Engineering from Anna University-Tiruchirappalli, India and perusing her PhD scholar in Computer Science and Engineering majoring in Data Mining and Soft computing Techniques from National Institute of Technology-Tiruchirappalli, India. Her current research interests are Data Mining, Machine Learning Concepts, Metaheuristic and swarm Intelligence Algorithms, Soft computing Techniques, Fuzzy logic, Evolutionary computation and Big Data Analytics. She is a Member of Indian Society of Technical Education (MISTE).

E. Sivasankar obtained his PhD degree in Computer Science and Engineering majoring in Data Mining from MS University- Tirunelveli, India. He is currently an Assistant Professor in the Department of Computer Science and Engineering from National Institute of Technology-Tiruchirappalli, India. His current research interests are Data Warehousing and Data Mining, Web Technology, Data Base Management System, Big Data Analytics, Sentimental Analysis, and Text Mining.