Computing (2018) 100:839–860
https://doi.org/10.1007/s00607-018-0633-6

Computing efficient features using rough set theory combined with ensemble classification techniques to improve the customer churn prediction in telecommunication sector

J. Vijaya[1] · E. Sivasankar[1]

Received: 20 February 2018 / Accepted: 5 June 2018 / Published online: 11 June 2018

Abstract Rough set theory (RST) can be viewed as one of the classical set theory for handling with imprecision knowledge. The theory has discovered applications in numerous areas, for example, engineering, industries, environment and others. Churn in telecommunication sector, customer switching from one service provider to another. Predicting telecom customer churn is challenging due to the huge and inconsistent nature of the data. Churn prediction is crucial for telecommunication companies in order to build an efficient customer retention plan and apply successful marketing strategies. In this article, a methodology is proposed using RST to identify the efficient features for telecommunication customer churn prediction. Then the selected features are given to the ensemble-classification techniques such as Bagging, Boosting, Random Subspace. In this work the duke university-churn prediction data set is considered for performance evaluation and three sets of experiments are performed. Finally the performance of the proposed model is evaluated based on the following metrics such as true churn, false churn, specificity, precision and accuracy and it is identified that Proposed system designed with combining attribute selection with ensemble classification techniques works fine with classification accuracy of 95.13% compared to any single model.

Keywords Churn prediction · Telecommunication · Rough set theory · Feature selection · Ensemble classification · Bagging · Boosting · Random subspace

Mathematics Subject Classification 00-General · 03-Mathematical logic and foundations · 41-Approximations and Expansions

✉ J. Vijaya
406114003@nitt.edu

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

1 Introduction

Customer Relationship Management (CRM) is one of the most excellent ways to effectively administer in any business organization. Nowadays, keeping up clients is one of the major issues for the organizations. This is caused due to liberalization in business marketing, it will increase the disappointment of clients in the present organization and choose the best service provider among all service providers. In any organization, customer retention is one of the significant key factors for successful business operation. It has been distinguished that getting new clients is five to six times more costly than holding old clients [1]. Further, loss of clients will lead to income loss, as well as reduce the company good name, product loyalty and the organization morale. Churn in business sector means customer switching from one service provider, withholding existing clients of a firm is an essential thing to increase the economic growth of the firm and holds the good name of the firm in the related competitive business market [2]. There are several reasons for a customer going to churn from their parent organization, this reasons are the major issue for predicting churn in business sector and also identifying that reasons are very difficult. Because it is depend on the individual client interest and their economic level and services they want. The main aim of the organizations is to identify the main reasons of churn for each and every customer before it happens and based on the identified reasons the company will take the necessary steps for the particular customer. Predicting customer churn is challenging due to the huge and inconsistent nature of the data. Churn is a serious problem of banking, internet service provider, newspaper, insurance companies and telecommunication sector [3–7].

The development of any business organization is based on the efficient churn prediction model and mainly used in telecommunication fields [8]. The telecommunication sector providers struggle against competition to keep customers. There is a tremendous growth in the telecommunication sector year by year. 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 [9]. Nowadays, along with voice services Telecommunication industry provide data services, online gaming, e-tickets booking, online purchasing, online banking, entertainments, educational services and many more. Many customers are highly using these provisions that are provided by the telecommunication industries, which help the customer in many ways [10]. Along with the various services provided there are evolving of technology like a shift from 4G to 5G technology also. Because of this, the customers are searching for better service providers and technologies day by day. After identifying a better service provider the customers could easily shift to another service provider even without changing their existing mobile numbers. Hence, in order to retain their customer churn prediction becomes a highly wanted technique for every telecommunication service provider [11]. In the literature, different classical machine learning algorithm was applied and investigated for this identifying churn task, such as statistical classifier (KNN), decision tree, logistic regression, artificial neural networks, random forest, naive bayes, support vector machines [12–14]. Predicting telecommunication customer churn is challenging due to the huge and unwanted features. The important features in the telecommunication sector are customer satisfaction and dissatisfaction information, customer service characteristics, status of subscriber line, customer demographic characteristics, call details, number of calls, minutes of calls, invoice income information, invoice payment information, change and ratio information in the billing. But there are some lagging in identifying these necessary features. Hybridization of more than one algorithm was also proposed churn prediction in an attempt to outperform the single algorithm approach [15,16]. To overcome the above issues a hybrid methodology is proposed to combine Rough Set Feature Selection (RSFS) with Ensemble Classification techniques. In this article, a methodology is proposed using rough set theory (RST) to identify the efficient features for telecommunication Customer Churn Prediction (CCP). Then the selected features are given to the ensemble-classification techniques such as Bagging, Boosting, Random Subspace. In this work the duke university-churn prediction data set is considered for performance evaluation. Finally the performance of the proposed model is evaluated based on the following metrics such as true churn, false churn, specificity, precision and accuracy.

This paper is organized as follows: Sect. 2 discusses the related work done using ensemble classification and feature selection methodologies present in the telecommunication customer churn prediction. Section 3 describes the rough set theory and how it is used to solve feature selection problems. In Sect. 4 describes the model evaluation and the experimental results. Finally, the findings and remarks of this work conclude in Sect. 5.

2 Literature survey

In earlier days predicting probability of the churn and churn data analysis have been handled by expects manually. There is continuous change and development in the telecommunication sector, i.e., globalization and liberalization of telecommunication industries give different picture, that is manual prediction is not possible. Nowadays, several contributions are available for telecommunication customer churn prediction such as artificial intelligence, soft computing techniques, machine learning, optimization algorithms, deep learning. This literature survey section discusses some of the effective contributions based on feature selection, ensemble techniques and some hybrid learning methods.

In the literature, different classical machine learning algorithm was applied and investigated for this identifying churn task, such as statistical classifier (KNN), decision tree, logistic regression, artificial neural networks, random forest, naive bayes, support vector machines [12–16]. Many ensemble based churn prediction model, which is a predictive model composed of a weighted combination of multiple classification models has performed better than single classification churn prediction model. Abbasimehr et al. [17] implemented four ensemble models such as Bagging, Boosting, Stacking, and Voting based on four known base learners, i.e., C4.5 Decision Tree (DT), Artificial Neural Network (ANN), Support Vector Machine (SVM) and Reduced Incremental Pruning to produce Error Reduction (RIPPER). Their proposed Boosting RIPPER and Boosting C4.5 gives better result compared to single base classifier as well as other ensemble techniques [17]. De Bock et al. [18] proposed two rotations based ensemble classifiers such as RotBoost (Rotation Forest combines with Ada-boost) and Rotation Forest. Their proposed RotBoost gives better result compared to well known base classifiers, Rotation Forest and three feature extraction algorithm such as PCA, ICA, SRP for all 4 real world churn prediction data set [18]. De Bock et al. [19] proposed an ensemble technique called GAMensPlus and it is based on GAMens. The two parameters generalized feature importance scores and bootstrap confidence bands for smoothing splines are added to the existing GAMens parameters. The GAMens technique is a mixture of random subspace method, bagging and semi-parametric GAMs [19]. Kim et al. [20] proposed an ensemble model (USE-SVM) based on a weighted combination of multiple SVM with efficient uniformly sub-sampling techniques. The proposed USE-SVM gives better result compared to well known base classifiers and ensemble models and also handle the highly skewed class distribution problem [20]. Liu et al. [21] proposed an ensemble technique based on Real Ada-boost. They are also generating new weighted mechanism for error samples in each iteration. Data mining techniques have difficulty in predicting rare classes when there are unsymmetrical class distributions. They are using Random under sampling techniques to balance the rare classes which is equal to the major class [21]. Lu et al. [22] proposed a boosting based churn prediction model, in which the entire customers are grouped into two clusters based boosting algorithm giving some weight. The proposed boosting algorithm gives better separation of churn customer from the non churn customer compared to logistic regression [22]. Xiao et al. [23] proposed a one step process based ensemble churn prediction model. Their proposed technique utilizes the multiple classifier ensemble techniques and cost sensitive learning to perform predictions [23].

In this paragraph discuss some of the recent and main studies of churn management, according to the techniques used as feature selection. Droftina et al. [24] proposed a feature selection technique for prepaid mobile social network customer churn prediction. The good features are identified using determined scores from available 74 features. The identified features of mobile social network customer are used for building the churn prediction model. As it can seen from results, reduced data sets results have better predictive accuracy than the original data set alone with classifier [24]. Idris et al. [25] proposed a PSO based under-sampling method as the number of non churner's are very high compared to the number of churner's. The various reduction techniques were evaluated. Finally the balanced data set is modeled using Random forest and K-nearest neighbor classifiers. His approach produced better AUC value [25]. Idris et al. [26] was presented a feature selection technique such as minimum Redundancy Maximum Relevance (mRMR). First of all most important effective features have been determined by using mRMR to estimate the churn of the telecommunication company. Then the ensemble of Random Forest, Rotation Forest, RotBoost and DECORATE techniques is applied to customer data with determined features [26]. Maldonado et al. [27] proposed a good profit based churn prediction approach, in which the classification process and variable selection process done in parallel by using support vector machines. Their proposed profit based churn prediction model gives better result for business -related goals compared to well known base classifiers and conventional techniques for feature selection [27]. Sivasankar and Vijaya [29] proposed a churn prediction model based on feature selection combined with base classification. The dominant features are identified using wrapper and filter based feature selection methods. The system designed with combining balanced data with attribute selection works fine with a classification accuracy of 98. 96% [28]. Vijaya and Sivasankar [29] proposed a telecommunication customer churn prediction using one metaheurstic algorithm like particle swarm optimization (PSO). In this work KDD Cup 2009 churn prediction data set is considered for performance evaluation. They present four different churn prediction models such as single PSO as a classifier, PSO embedded with feature selection, PSO embedded with simulated annealing and finally PSO embedded with feature selection as well as simulated annealing [29]. Xiao et al. [30] proposed a dynamic transfer ensemble model. In this model, the first step optimal features are obtained using GMDH-type neural network in which the original data set is reduced in size to form subgroups. It should ensure that the system is able to achieve the same higher accuracy with the subgroups as achieved by retaining all the features in the attribute set. Thus the attribute selection methodology should reduce miss-classification error and should enhance the accuracy with reduced computational cost. The next step different classifiers are trained and best classifier for the test data is selected iterative dynamic analysis. This model is more time consuming, because the iterative approach for feature selection and also lots of classifiers are used for predicting churn [30].

In this paragraph discuss some of the recent and main studies of churn management, according to the techniques used as hybrid leaning methods. Rajamohamed and Manokaran [31] used the data set of credit card churn prediction from UCI repository for their hybrid model building which uses improved Rough K-Means algorithm for clustering and KNN, DT, SVM, NB and ANN for classification. Hence, they proved their proposed hybrid model to be efficient when compared with a single classifier [31]. To cluster the customers who are having the similar characteristics into segments, Hudaib et al. [32] used the clustering methods like K-Means, Self-Organizing Map (SOM) and Hierarchical clustering algorithms. Then the clustered segments are fed into the Artificial Neural Networks (MLP-ANN) classifier for hybrid model building and the accuracy is evaluated [32]. Huang and Kechadi [33] built a hybrid model on telecommunication data set which uses proposed Weighted K-Means algorithm for clustering and First-Order-Inductive-Learning for classification [33]. From the analyzed techniques it was observed that most of the current churn prediction models utilize hybrid models and also selection of important features and balancing techniques played a key role in telecommunication customer churn prediction.

3 Proposed churn prediction model

In this study, an integrated approach of feature selection and ensemble classification is proposed to handle the high dimensional customer data. Summarize below, our proposed model can be displayed as Fig. 1. In this work the duke university-churn prediction data set is collected first and considered for performance evaluation. The normal preprocessing is performed on the collected orange data, such as missing value elimination, string to numeric conversion, normalization and discretization. Next, the dominant features for telecommunication customer churn prediction are iden-

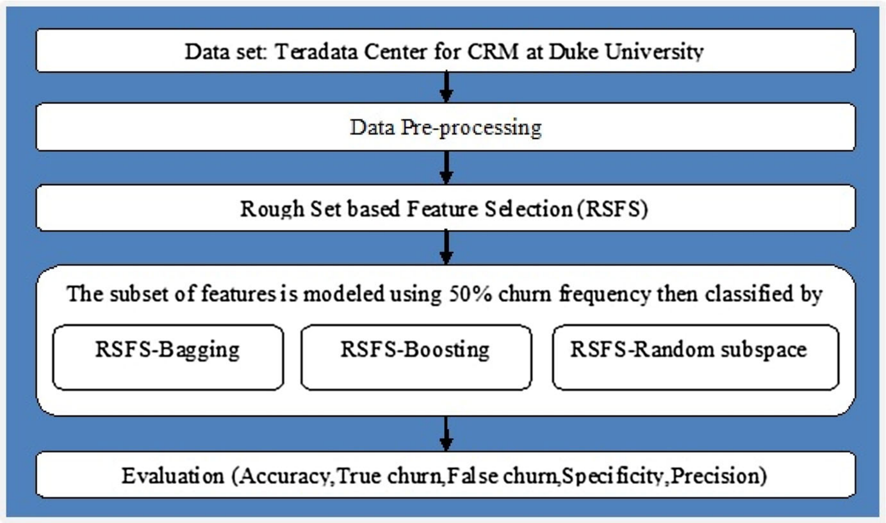
Data set: Teradata Center for CRM at Duke University
Data Pre-processing
Rough Set based Feature Selection (RSFS)
The subset of features is modeled using 50% churn frequency then classified by
RSFS-Bagging
RSFS-Boosting
RSFS-Random subspace
Evaluation (Accuracy,True churn,False churn,Specificity,Precision)

Fig. 1 Customer churn prediction using RSFS with ensemble classification

tified using rough set feature selection (RSFS). Then the selected features are given to the ensemble-classification techniques such as Bagging, Boosting, Random Subspace. This integrated approach for customer churn prediction has three variants. RSFS embedded with Bagging classification is initially analyzed (RSFS-Bagging). This is followed by RSFS embedded with Boosting (RSFS-Boosting) and RSAS embedded with Random Subspace (RSFS-Random Subspace).

3.1 Dataset

Three different groups of data are found in the Tera data Center for CRM at Duke University [34]. One is found to be training data (calibration) and another two are found to be test data (future and score). In this work, the training data part of churn prediction data set (calibration) is taken. There are 100,000 samples and 172 attributes with a class label present in the data set. Out of 172 attributes, 137 attributes are found to be numerical attributes and 35 are found to be nominal attributes. In this data set, nearly 20% of the data set consists of missing values. Nearly 50,438 consumers are non-churner's and 49,562 consumers are found to churn from the subscription.

3.2 Data preprocessing

Important in establishing the prediction model is to obtain the appropriate data. In the considered data set, nearly 20% of the data are missing values. So to increase the accuracy and consistency of the proposed system, standard preprocessing techniques are deployed. In the preprocessing phase last_swap, Customer_ID features are eliminated from the data set and now the size of the features is 170. From the above obtained data set all the missing values of numerical attributes are replaced by mean value and the missing values of nominal attributes are replaced with mode of the attribute values. Seven of the nominal features is showing higher deviation in the attribute values. Hence those features with higher deviations are removed from the data set reducing the features count to 163. Since the statistical classification algorithm handles only numerical values all the string values are converted to numeric. In the next step normalization is carried out using min-max normalization. Then the discretization process is carried out on the normalized data, it partitions the continuous attributes into discretized interval attributes.

3.3 Rough set based feature selection

The extraneous attributes are removed from the data set using a process called feature selection, which also should ensure the classification accuracy of acceptable value. The criteria for selecting the attributes play a key role because it has a direct impact on the effective classification results. In this study, we use the Rough set based feature selection approach for subset generation to specify the feature subset for customer churn prediction.

Rough set theory (RST) is one of the numerical way to deal with defective information or vagueness. The issue of defective learning has been handled previously by statisticians, industries and mathematicians. As of late it turned out to be an important issue in the artificial intelligence area. There are many ways to deal with the issue of how to comprehend and control defective information. Fuzzy set theory and Rough set theory are the most successful concepts proposed by Pawlak [35] to eliminate the vagueness. Rough set theory depends on the assumption that we can relate information or data with each element of the universe of discourse called as U. An RST has an exact idea of lower, upper approximation and boundary region. Vagueness is normally connected with the boundary region. The boundary line is used to separate the lower and upper approximations. The boundary elements can't be characterized with respect to the available learning information about the objects. Therefore the Rough set approx is a combination of lower, upper approximation and vagueness [35].

3.3.1 Information system and decision table

In RS theory, a given data set or knowledge is represented in an InFormation System (IFS), and is referred as the 4-tuples $IFS = (U, F, V, \varphi)$, where, U and F are nonempty finite set. The set of m objects is referred to $U = \{x_1, x_2 \ldots x_m\}$ called the *universal set*. The set of n features or attributes is represented as $F = \{f_1, f_2 \ldots f_m\}$ called a *feature set or attribute set* [36]. Let $F = P \cup Q$, then the tuples $DES = (U, P \cup Q, V, \varphi)$ is referred as *decision table* and P is the condition attribute set and Q is the decision attribute set. Here $P \cap Q = \phi$. In general, $V = U_{f \in F} V_f$ and V_f is the set of values associated with the attribute f or domain of the attribute f. An information function $\varphi : UXF \rightarrow V$ is determined for each attribute $f \in F$, such that $\varphi(x, f) \in V_f, \forall x \in U$.

3.3.2 Indiscernibility relation

The key notion of RS theory is the indiscernibility relation or similarity relation. If two objects are identical or similar feature values, then they are said to be indiscernible (similar) objects [37]. Let us define the indiscernibility relation more precisely:

A binary relation or indiscernibility relation R on the universe U of any subset H of an attribute set *F*, i.e., $H \subseteq F$, denoted by *R(H)*, is defined in the following Eq. (1):

$$R(H) = \left\{(x_i, x_j)\dot{:}(x_i, x_j) \in U^2, \forall_{f \in H}(f(x_i) = f(x_j))\right\} where 1 \leq i, j \leq m \quad (1)$$

where $f(x_i)$ and $f(x_j)$ represents the value of the attribute f for the object x_i and x_j respectively. It is simply noticed that the *R(H)* is defined in this way is referred as an equivalence relation. The collection of the complete equivalence classes of *R(H)*, is represented by the Quotient set of the universe *U*, referred as $U/R(H)$, in short U/H; an equivalence class of *R(H)* and it is denoted by $U/H = \{[x]_H | x \in U\}$.

If $(x_i, x_j) \in R(H)$, then x and y are said to be indiscernible (similar or identical) with respect to *H*, i.e., H-indiscernible. The equivalence classes of the H-indiscernibility relation are represented by $[x]_H$. Hence, the elements in $[x]_H$ are indiscernible by features or attributes from *H*. The equivalence classes of the relation *R(H)* are called H-elementary sets.

3.3.3 Lower and upper approximation

In the preceding section, the indiscernibility relation is depicted and it is used in the definition of another vital concept of RS theory, namely lower and upper approximations. These approximations can be generated by using the indiscernibility relation and they are similar to operations in a topology such as *interior* and *closure* respectively [35,38].

Given any concept $Z \in U$ and $H \in F$, *Z* can be approximated as H-lower and H-upper approximations of *Z*, with the information on *H*. The lower approximation of *Z*, denoted by $\underline{R_H}(Z)$, in Eq. (2) is pointing to the set of objects of *U* that are certainly in *Z*, then:

$$\underline{R_H}(Z) = x \in U | [x]_H \subseteq Z \quad (2)$$

$\overline{R_H}(Z)$ in Eq. (3) is the upper approximation of *Z* and it refers to the set of objects of U that are possibly in *Z*, defined as:

$$\overline{R_H}(Z) = x \in U | [x]_H \cap Z \neq NULL \quad (3)$$

Furthermore, there are three regions is obtained by dividing the universe *U*, namely positive, negative and boundary regions. The regions are disjoint in nature and represented as follows in Eqs. (4)–(6):

$$PR(Z) = \underline{R_H}(Z) \tag{4}$$

$$NR(Z) = U - \overline{R_H}(Z) \tag{5}$$

$$BR(Z) = \underline{R_H}(Z) - \overline{R_H}(Z) \tag{6}$$

From the above definitions, it is make clear that if an object $x_i \in PR(Z)$, then x_i belongs to target set Zdefinitely. Suppose if an object $x_i \in BR(Z)$, then it doesn't belong to Z certainly. Moreover, if $BR(Z) = \phi$, then the set Z is said to be *crisp (or exact)* with reference to H; Otherwise, i.e., if $BR(Z) \neq \phi$, then the set Z is referred to *rough (or inexact)* in respect of H. If an object $x_i \in NR(Z)$, then it's not possible to determine whether the object x_i belongs to a target set Z or not. The pair $\left[\underline{R_H}, \overline{R_H}\right]$ is referred to as the Pawlak's rough set of Z with respect to the set of attributes F.

3.3.4 Accuracy of approximations

Rough set is also characterized by calculating the ratio between lower and upper approximation of the target set Z in terms of numerical. It is known as *Imprecision coefficient*, denoted $\alpha_H(Z)$, is defined in the following Eq. (7),

$$\alpha_H(Z) = \frac{|\underline{R_H(Z)}|}{|\overline{R_H(Z)}|} \tag{7}$$

where $Z \neq \phi$, $|.|$ represents the cardinality of a finite set. Trivially, $0 \leq \alpha_H(Z) \leq 1$. The target set Z, can be classified into two sets with respect to the attributes set $H \subseteq F$ is as follows. If $\alpha_H(Z) = 1$, then **Z** is crisp (or precise) set or also referred as definable set, and otherwise, If $\alpha_H(Z) < 1$, then Z is rough (or vague) set.

In addition, the roughness can be used to measure the uncertainty of a rough set. Formally, the roughness of the set Z with respect to H, denoted $S_H(Z)$, is defined as in Eq. (8)

$$S_H(Z) = 1 - \alpha_H(Z) \tag{8}$$

where $\alpha_H(Z)$ is an imprecision coefficient with $Z \neq \phi$

3.3.5 Dependency of attributes

The concept of dependency presented in RS theory corresponds to that considered in relational databases. In Data analysis, it is a major issue that to discover the dependence between attributes. Formally, the dependency is described as follows. Let Q and P be subsets of F. Let k be the degree of dependency between Q and P, then we state that *Q depends in degree k* on P, in Eq. (9) referred to $p \Rightarrow_k Q$, if

$$k = \gamma(P, Q) = \frac{|PR_p(Q)|}{|U|} \tag{9}$$

where $|U|$ is the cardinality of the universe U and $PR_p(Q) = U_{Z\in U/Q} R_p(Z)$ called a positive region of the partition *U/Q* with reference to *P*. All the elements of *U*, can be uniquely classified to blocks of the partition *U/Q*, by means of *P*. The following assumptions are made from the value of *k*.

- If $k = 1$, then *Q* depends *totally* on *P*, and
- If $k < 1$, then *Q* depends *partially* (in a degree *k*) on *P*.

It can be simply seen that,*Q* depends *totally* on $P \iff R(P) \subseteq R(Q)$ and it states that the partition produced by *P* is better than the partition produced by *Q*. The idea of dependency among attributes is strongly coupled to the concept of consistency of decision table [35]. The coefficient $k^{`} = 1 - k$, is called the inconsistency degree of decision table.

3.3.6 Reduction of attributes

Reduct and core attributes are the most important concepts of RS theory and mainly used for rule discovery from the database. Given an *IS*, some attributes are *redundant* or *superfluous* with respect to classification based on any subset of attribute set *F*. So, these attributes are removed without losing the classification power of the reduced *IS* [35]. The reduct, *Red* is a set of minimal set of features or attributes from *F* that preserves the same classification of objects of the universe *U* with the whole set of attributes. Thus, it preserves partition. In precise, Let $G, H \subseteq F$ and $b \in H$. Then, the following assumptions are hold:

- If $R(H) = R(H - \{b\})$, then *b* is *superfluous* in *H* ; otherwise *b* is *indispensable* in *H*.
- If all the attributes of *H* are indispensable, then set *H* is said to be *independent*.
- Given $G \subseteq H$. If *G* is independent and $R(G) = R(H)$, then the set *G* is the Red of *H*.

The Core of attributes act the very important role, in order to find the most contributing features in *IS*. More formally, Let $H \subseteq F$. The set of all indispensable attributes of *H* is defined as the *Core* of H. The connection between *Red* and *Core* is defined in the following Eq. (10).

$$Core(H) = \cap Red(H) \tag{10}$$

where *Red(H)* represents the set of all reducts of *H*.

From the above property, it is observed that *Core* is the common part (or intersection) of all reducts. Hence, each object of the *Core* certainly belongs to any of reduct. Thus, in a sense, core contains all the attributes that cannot be removed from the set *A* without altering the original classification. Therefore, the rules can be generated by using these concepts of reduct (sufficient information and core (minimum sufficient information) for the discrimination of objects. In practice, *discernibility matrix* is used to compute *Red* and *Core*. There are numerous approaches to compute the same.

3.4 Ensemble classification techniques

More predictor is better than the single one, an ensemble combines many predictors is often some weighted average or weighted combination of predictors. It might be the same type of learners or different type of learners. An ensemble is a supervised learning algorithm, which can be trained and then used to make predictions. Ensemble Size, the number of constituent classifiers of an ensemble involved in generation of the model plays a vital role in performance and accuracy of prediction. There are several ensemble algorithms are available, the most common algorithm for ensemble techniques are Bagging, Boosting and Random Subspace method.

3.4.1 Bagging

Bagging is a method for ensemble learning was proposed by Breiman [39]. Bagging stands for Bootstrap aggregation. In order to use an ensemble of learners and be able to combine the learners and get better accuracy. In the bagging method, we need some set of learners which makes some independent errors. According to the bagging technique, several classifiers are trained independently on different sets of data through bootstrap method. The bootstrap method generates k subsets out of the training data set through sampling with replacement(SWR). Afterward, k classifiers are performed on each subset and the eventual results of these k classifiers are combined. The unknown test data are predicted based on the majority voting for the different k learners.

3.4.2 Boosting

In Boosting [40], is an iterative procedure, we start with uniform probability distribution on the given training instances and we adaptively change the distribution of the training data. Initially all the training instances have equal weights after each round of boosting weight get changed. We assign a strength to each learners and this strength is used to decide the weight of the voting and the final classification is a linear combination of this different hypothesis that weight for each learners. There are several boosting algorithms are available, one most common algorithm for boosting is Ada-Boosting.

3.4.3 Random subspace

In the Random Subspace (RS) method, classifiers are constructed in random subspaces of the data feature space [41]. Afterward, k classifiers are performed on each subset feature space and the eventual results of these k classifiers are combined. Finally, when an unknown instance is presented to each individual classifier, a majority or weighted vote is used to infer the class. Here Bagging, Boosting and Random Subspace method has designed for using Decision Trees (DT). Algorithm 1 that outlines the RSFS-Ensemble classification procedure.

Algorithm 1: Proposed algorithm for RSFS-Ensemble Classification

```
Stage1:Data collection
    The data set is withdrawn from KDD cup 2009
Stage2:Data Preprocessing
    Fill the missing values
    String to numeric conversion
    Normalization and Discretization
Stage3:Feature Selection procedure
    Compute indiscernibility relation
    Compute lower and upper approximations and positive region
    Build the decision-relation discernibility matrix
    Generate all reducts and core
Stage4:Ensemble Classification for subset of features(Reduct)
    If the ensemble algorithm is Bagging
            The Data set consists of K samples and N features
            Generates k subsets from the training through SWR.
            k classifiers are performed on each subset
            Each test instance is predicted based on majority voting by k classifiers.
    Else if the ensemble algorithm is Boosting
            The Data set consists of K samples and N features
            Initially, all objects have equal weights,construct the 1st classifier
            Increase the weight for the prediction error object
            Repeat the steps until maximum accuracy is reached
   Else
            The Data set consists of K samples and N features
            Generates k subsets of feature space from the training through SWR
            k classifiers are performed on each subset
            Each test instance is predicted based on majority voting by k classifiers
   End if
Calculate Accuracy and other metrics.
```

4 Experiments and results

4.1 Performance measures

The confusion matrix is a base element for both comparing and understanding efficiency of the classifier. In Table 1, where F_{11} is the number of samples that both actually positive and positive predicted, F_{22} is the number of samples that both actually negative and negative predicted , F_{12} and F_{21} represent the number of classification errors. The following metrics such as accuracy, true churn, false churn, specificity and precision and is represented in Eqs. (11)–(15)

Table 1 Confusion matrix for customer churn prediction

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

$$Accuracy = \frac{F_{11} + F_{22}}{F_{11} + F_{12} + F_{21} + F_{22}} \tag{11}$$

$$Truechurn = \frac{F_{11}}{F_{11} + F_{12}} \tag{12}$$

$$Falsechurn = \frac{F_{21}}{F_{21} + F_{22}} \tag{13}$$

$$Specificity = \frac{F_{22}}{F_{21} + F_{22}} \tag{14}$$

$$Precision = \frac{F_{11}}{F_{11} + F_{21}} \tag{15}$$

4.2 Experiment setup

Three sets of experiments are performed in this work. Initially a series of experiments are executed to calculate the performance and behavior of the single classsification model like DT, SVM, KNN, NB and ANN and ensemble classification techniques Bagging, Boosting and Random subspace. In the second phase set of research works are performed to evaluate the hybridization behavior and performance of those methods, including rough set, filter and wrapper based attribute selection combined with classification procedures DT, SVM, KNN, NB and ANN and ensemble classification techniques Bagging, Boosting and Random subspace. Finally, the efficiency of proposed techniques compared with existing ensemble and feature selection based techniques

4.2.1 Setup-I

- **Performance based on base classifiers**
  The data set that is being preprocessed consisting of 100,000 samples accounting to 163 attributes with one class prediction label indicating churn or non-churn about the samples, Now the data set is categorized into training and testing data sets. Out of 100,000 samples 50,438 consumers are non-churner's and 49,562 consumers are found to churn from the subscription. On the basis of churn frequency of sampling that amounts to 50% in the single categorizing model, the training data set which the built consist of 24,781 samples that are churned members and 25,219 samples that are non-churn members. The testing data partition consists of the same number of churn and non-churn samples. Table 2 depicts the performance of the base classifier system on experimenting it using parameters like true churn, false churn, specificity, precision and accuracy. It also indicates that the data being preprocessed in this system performs well for the complete classification process, when compared to the classification being carried out normally without preprocessing. Among all the classifiers considered ANN has achieved the highest accuracy of 87.10% and highest true churn of 95.89% attaining the highest objective function values, which is highlighted with bold letters.

Table 2 The performance of base classifiers

Classifier	DT	KNN	SVM	NB	ANN
Accuracy	83.69	72.81	79.21	86.27	**87.10**
True churn	91.58	83.78	86.52	95.35	**95.89**
False churn	12.77	36.96	25.10	12.00	15.69
Specificity	87.23	63.04	74.90	88.00	84.31
Precision	90.42	81.29	89.53	89.86	90.23

Table 3 The performance of ensemble classifiers

Ensemble classifier	Bagging	Boosting	Random subspace
Accuracy	92.66	**93.73**	89.63
True churn	98.74	**99.81**	98.66
False churn	01.45	01.01	02.53
Specificity	98.55	98.99	97.47
Precision	93.76	93.89	90.71

- **Performance based on ensemble classifier**
  The training data set consist of 24,781 samples that are churned members and 25,219 samples that are non-churn members. The testing data partition consists of the same number of churn and non-churn samples. Table 3 depicts the performance of the ensemble classifier system on experimenting it using parameters like true churn, false churn, specificity, precision and accuracy. Here bagging, boosting and the random subspace method has designed for using Decision Trees (DT). Bagging technique is an iterative approach in which it is initiated with the number of bags in the prototype. In each iteration, it is constantly trying to increase the number of bags that shows the best accuracy or improvement in the prototype and continues till an increase the number of bags does not show any improvement in the performance of the prototype. In this experiment the accuracy of bagging techniques is convergence between the number of bags is equal to 82. The accuracy of boosting techniques is convergence between the number of iterations is equal to 78. The accuracy of random subspace techniques is convergence between the number of feature subspace is equal to 81. Table 3 shows that the data set which is ensemble gives better results compared to single classifier. Among all the ensemble classifiers considered Boosting has achieved the highest accuracy of 93.73% and highest true churn of 99.81% attaining the highest objective function values which is highlighted with bold letters. Figure 2 visualizes the accuracy assessment between the base classifier and ensemble classifier.

4.2.2 Setup-II

- **Performance based on base classifier with feature selection approach** For attribute selection, completely preprocessed data are taken as input. In this sec-

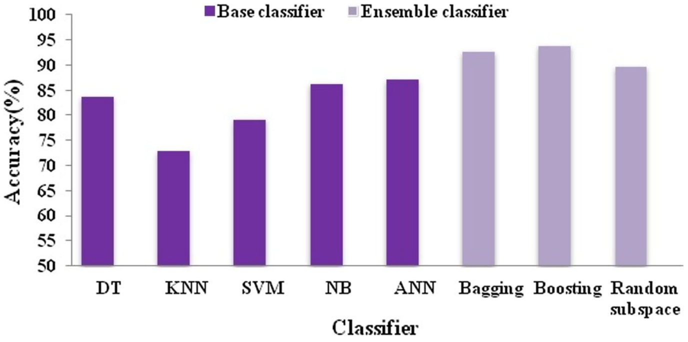

Fig. 2 Accuracy comparison between base classifier and ensemble classifier

tion, three variants of feature selection techniques are considered for comparison such as filter based, wrapper based and rough set based feature selection. The filter based methodologies like Correlation feature selection (CFS) and Information gain (IG) is deployed to select the best attributes. The wrapper based methodologies like forward search (FS) and Backward search (BS) is deployed to select the best attributes. R language for programming is used for churn prediction. Here package called FSelector can be used to filter based and wrapper based feature selection, since it contains the facilities for attribute selection in the data set being considered and a final attribute can also be fixed. In filter based feature selection, initially all the attributes are ranked and then processed to select the best K attributes. In this work K takes the value of 15.In Wrapper based feature selection, classification algorithm provides maximum accuracy by selecting the best attribute subset. The package called RoughSets can be used for rough set based feature selection. The 19 attributes are selected in this process. The designed system works well with 100,000 samples with the best attributes identified and a churn prediction variable. On the basis of churn frequency of sampling that amounts to 50% in the single categorizing model, the training data set which the built consist of 24,781 samples that are churned members and 25,219 samples that are non-churn members. The number of features is based on the outcome of the feature selection techniques. A training data model is created and modeled with the widely considered classifiers, the test samples are evaluated and the prediction is made on the basis of prototype created by the classifiers in consideration. Various classification algorithms are deployed in this work such as DT, KNN, SVM, NB and ANN. During experimentation the performance of the prototype is evaluated using parameters like true churn, false churn, specificity, precision and accuracy and the results are tabulated in Tables 4, 5, 6, 7 and 8 (maximum accuracy value is highlighted in bold letters) and accuracy is graphically depicted in Fig. 3. The proposed feature selection approach, RSFS-SVM performs better than other techniques and achieved the highest accuracy of 92.12%

- **Performance based on ensemble classifier with feature selection approach**
  The designed system works well with 100,000 samples with the best attributes

Table 4 The performance of RSFS along with base classifiers

Classifier	DT	KNN	SVM	NB	ANN
Accuracy	91.04	88.38	**92.12**	90.86	92.10
True churn	95.79	93.95	95.96	98.08	96.04
False churn	08.80	10.86	13.58	02.54	11.11
Specificity	91.20	89.14	86.42	97.46	88.89
Precision	94.79	93.62	95.79	92.49	95.69

Table 5 The performance of CFS along with base classifiers

Classifier	DT	KNN	SVM	NB	ANN
Accuracy	86.49	85.98	87.01	81.25	**88.96**
True churn	92.68	92.45	95.87	89.13	95.02
False churn	11.64	13.68	14.02	16.00	25.93
Specificity	88.36	86.32	85.98	84.00	74.07
Precision	92.68	92.29	90.18	89.78	93.03

Table 6 The performance of IG along with base classifiers

Classifier	DT	KNN	SVM	NB	ANN
Accuracy	88.47	84.36	89.35	87.22	**90.57**
True churn	97.06	91.43	93.82	93.39	94.69
False churn	03.88	14.17	10.14	15.55	13.04
Specificity	96.12	85.83	89.86	84.45	86.96
Precision	90.86	91.36	94.87	92.77	95.35

Table 7 The performance of FS along with base classifiers

Classifier	DT	KNN	SVM	NB	ANN
Accuracy	84.55	87.33	87.99	88.87	**89.77**
True churn	94.63	93.16	93.73	99.54	98.80
False churn	08.00	11.63	14.36	01.89	02.28
Specificity	92.00	88.37	85.37	98.11	97.72
Precision	88.89	93.19	93.33	89.21	90.74

Table 8 The performance of BS along with base classifiers

Classifier	DT	KNN	SVM	NB	ANN
Accuracy	84.73	80.60	**90.49**	88.35	87.97
True churn	92.12	91.74	95.37	97.25	93.23
False churn	14.66	08.45	08.46	03.29	11.24
Specificity	85.34	91.55	91.54	96.71	88.76
Precision	91.10	86.65	94.59	90.57	93.88

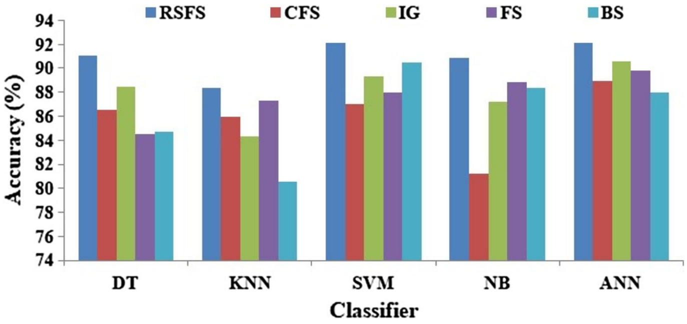

Fig. 3 Accuracy comparison between base classifier with feature selection approach

Table 9 The performance of proposed RSFS along with ensemble classifiers

Ensemble classifier	Bagging	Boosting	Random subspace
Accuracy	95.04	**95.13**	94.39
True Churn	99.72	99.79	98.86
False Churn	01.23	01.67	03.85
Specificity	98.77	98.33	96.48
Precision	95.28	95.31	95.42

identified and a churn prediction variable. On the basis of churn frequency of sampling that amounts to 50% in the single categorizing model, the training data set which the built consist of 24,781 samples that are churned members and 25,219 samples that are non-churn members. The number of features is based on the outcome of the feature selection techniques. The filter based approach number of selected feature is 15 and rough set based feature selection the number of selected feature is 19 and wrapper based method it is depend on the maximum accuracy features. A training data model is created and modeled with the widely considered classifiers, the test samples are evaluated and the prediction is made on the basis of prototype created by the classifiers in consideration. Various ensemble classification algorithms are deployed in this work such as Bagging, Boosting, Random subspace. During experimentation the performance of the prototype is evaluated using parameters like true churn, false churn, specificity, precision and accuracy and the results are tabulated in Tables 9, 10, 11, 12 and 13 and accuracy is graphically depicted in Fig. 4. The proposed feature selection approach, RSFS-Boosting performs better than other techniques and achieved the highest accuracy of 95.13%.

Tables 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13 indicates the performance of well-preprocessed data through single Classifiers, Ensemble Classifiers, Classifier by means of Filter, Wrapper and Rough Set based Feature Selection and Ensemble Classifier constructed by means of Filter, Wrapper and Rough Set based Feature Selection. The tables indicates that ensemble methodology progresses the performance and working of the classifier. The table also indicates that attribute selection methodology pro-

Table 10 The performance of CFS along with ensemble classifiers

Ensemble classifier	Bagging	Boosting	Random subspace
Accuracy	91.98	**92.84**	92.69
True Churn	98.44	98.78	99.37
False Churn	02.77	01.45	00.71
Specificity	97.23	98.55	99.29
Precision	93.32	93.91	92.97

Table 11 The performance of IG along with ensemble classifiers

Ensemble classifier	Bagging	Boosting	Random subspace
Accuracy	93.57	**94.53**	90.81
True Churn	97.50	99.22	97.75
False Churn	02.50	01.85	01.98
Specificity	97.50	98.15	98.02
Precision	93.57	95.24	92.73

Table 12 The performance of FS along with ensemble classifiers

Ensemble classifier	Bagging	Boosting	Random subspace
Accuracy	91.47	**92.62**	90.58
True Churn	96.12	96.82	95.13
False Churn	05.15	02.53	03.23
Specificity	94.85	97.47	96.77
Precision	94.96	95.52	94.97

Table 13 The performance of BS along with ensemble classifiers

Ensemble classifier	Bagging	Boosting	Random subspace
Accuracy	**92.67**	92.58	89.27
True Churn	96.73	99.72	93.95
False Churn	06.45	00.71	19.28
Specificity	93.55	99.29	80.72
Precision	95.64	92.81	94.56

gresses the performance and working of the classifier. The Proposed Rough Set based Feature Selection (RSFS) with Ensemble Classification model perform better than base Classifiers and Ensemble classifier without Feature Selection, base Classifiers and Ensemble Classifier with Filter, Wrapper based Feature Selection. The proposed

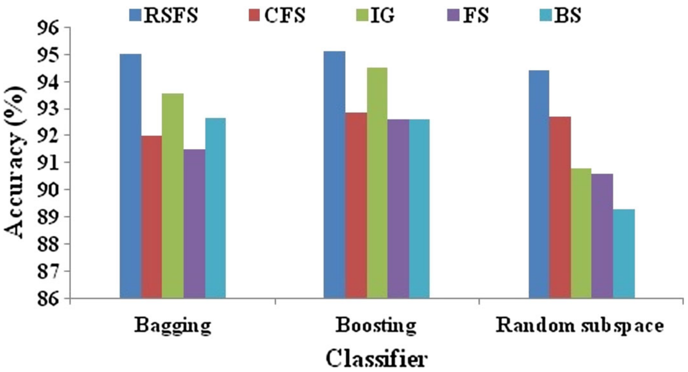

Fig. 4 Accuracy comparison between ensemble classifier with feature selection techniques

feature selection approach, RSFS-Boosting performs better than other techniques and achieved the highest accuracy of 95.13% which is highlighted with bold letters.

4.2.3 Setup-III

- **Performance comparison with other existing approaches**
  The proposed rough set based attribute selection hybrid with the ensemble knowledge system indicates the improved performance, when compared to various hybrid knowledge system designed by preceding research work listed in Table 14. The preprocessed data set (100,000 samples with 163 attributes and 1 class label) is given as input to the following model proposed by Idris et al. [25], Lu et al. [22] and Xiao et al. [30]. Idris et al. [25] proposed a PSO based under-sampling method as the number of non churner's are very high compared to the number of churner's. The various reduction techniques were evaluated. Finally the balanced data set is modeled using Random forest and K-nearest neighbor classifiers. His approach produced better AUC value [25]. Lu et al. [22] proposed a boosting based churn prediction model, in which the entire customers are grouped into two clusters based boosting algorithm giving some weight. The proposed boosting algorithm gives better separation of churn customer from the non churn customer compared to logistic regression [22]. Xiao et al. [30] proposed a dynamic transfer ensemble model. In this model, the first step optimal features are obtained using GMDH-type neural network in which the original data set is reduced in size to form subgroups. It should ensure that the system is able to achieve the same higher accuracy with the subgroups as achieved by retaining all the features in the attribute set. Thus the attribute selection methodology should reduce miss-classification error and should enhance the accuracy with reduced computational cost. The next step different classifiers are trained and best classifier for the test data is selected iterative dynamic analysis. This model is more time consuming, because the iterative approach for feature selection and also lots of classifiers are used for predicting churn [30]. The Table shows that the proposed feature selection approach, RSFS-Boosting

Table 14 Accuracy comparison with existing techniques

Methods	Accuracy
Proposed model RSFS-Boost	**95.13**
Chr-PmRF proposed by Idris et al. [25]	93.33
Ada Boosting proposed by Lu et al. [22]	90.21
GMDH-type neural network proposed by Xiao et al. [30]	88.89

performs better than other existing churn prediction techniques and achieved the highest accuracy of 95.13% which is highlighted with bold letters.

5 Conclusion

In this paper, the duke university data are pre-processed by means of operative data cleansing methodologies. Once cleaning of a data set is completed discretization of data samples are performed. In the next step attributes selection is executed using rough set based method. Using 50% of churn frequency, the sampling techniques are formed and partitioned in the form of training and testing data sets. The ensemble learning techniques are used to model the proposed system. The devised strategy efficiency is evaluated and marked with parameters like true churn, false churn, specificity, precision and accuracy. The following inferences are made out of the model. The data that are ensemble works well compared to the one that is not ensemble. The attribute selection performs effectively. The system designed by combining preprocessing data with attribute selection works fine with ensemble classification accuracy of 95.13%.

Limitation of this combined rough set based feature selection techniques with ensemble classification model is suitable for customer churn prediction data set that contains large amount of features. Classification accuracy is depend on the features in the corresponding data set only. In this work we are using simple feature selection techniques only, For example rough set based feature selection and filter based feature selection method like Correlation Feature Selection, Information Gain and Wrapper based feature selection methods include Forward Search, Backward Search. In the future, to create or add new feature set for corresponding data set in such a way that improve the number of features. For example other attributes like complaint information, fault information, network information are added. Network related information is very informative feature for customer churn prediction because the churn information is spreaded for customers through the communication networks. In addition, more advanced methods for feature selection techniques should be focused in the future.

References

1. Chung BD, Park JH, Koh YJ, Lee S (2016) User satisfaction and retention of mobile telecommunications services in Korea. Int J Hum Comput Interact 32(7):532–543
2. Bose I, Chen X (2009) Hybrid models using unsupervised clustering for prediction of customer churn. J Organ Comput Electron Commer 19(2):133–151

3. Ali OG, Arturk 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
4. Cancho VG, Dey DK, Louzada F (2016) Unified multivariate survival model with a surviving fraction: an application to a Brazilian customer churn data. J Appl Stat 43(3):572–584
5. Gunther CC, Tvete IF, Aas K, Sandnes GI, Borgan Q (2014) Modelling and predicting customer churn from an insurance company. Scand Actuar J 1:58–71
6. Milosevic M, Zivic N, Andjelkovic I (2017) Early churn prediction with personalized targeting in mobile social games. Expert Syst Appl 83:326–332
7. Sankaranarayanan HB, Vishwanath BV, Rathod V (2016) An exploratory analysis for predicting passenger satisfaction at global hub airports using logistic model trees. In: 2016 Second international conference on research in computational intelligence and communication networks (ICRCICN). IEEE, pp 285–290
8. Bose I, Chen X (2009) Hybrid models using unsupervised clustering for prediction of customer churn. J Organ Comput Electron Commer 19(2):133–151
9. Gamulin N, Stular M, Tomazic S (2015) Impact of social network to churn in mobile network. Automatika 56(3):252–261
10. Vafeiadis T, Diamantaras KI, Sarigiannidis G, Chatzisavvas KC (2015) A comparison of machine learning techniques for customer churn prediction. Simul Model Pract Theory 55:1–9
11. Verbeke W, Martens D, Baesens B (2014) Social network analysis for customer churn prediction. Appl Soft Comput 14:431–446
12. Abbasimehr H, Setak M, Soroor J (2013) A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques. Int J Prod Res 51(4):1279–1294
13. Farquad MAH, Ravi V, Raju SB (2014) Churn prediction using comprehensible support vector machine: an analytical CRM application. Appl Soft Comput 19:31–40
14. Huang B, Kechadi MT, Buckley B (2012) Customer churn prediction in telecommunications. Expert Syst Appl 39(1):1414–1425
15. 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
16. 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
17. Abbasimehr H, Setak M, Tarokh MJ (2014) A comparative assessment of the performance of ensemble learning in customer churn prediction. Int Arab J Inf Technol 11(6):599–606
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. De Bock KW, Van den Poel D (2012) Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. Expert Syst Appl 39(8):6816–6826
20. Kim N, Jung KH, Kim YS, Lee J (2012) Uniformly subsampled ensemble (USE) for churn management: theory and implementation. Expert Syst Appl 39(15):11839–11845
21. Liu M, Qiao XQ, Xu WL (2011) Three categories customer churn prediction based on the adjusted real adaboost. Commun Stat Simul Comput 40(10):1548–1562
22. Lu N, Lin H, Lu J, Zhang G (2014) A customer churn prediction model in telecom industry using boosting. IEEE Trans Ind Inf 10(2):1659–1665
23. Xiao J, Xie L, He C, Jiang X (2012) Dynamic classifier ensemble model for customer classification with imbalanced class distribution. Expert Syst Appl 39(3):3668–3675
24. Droftina U, Stular M, Kosir A (2015) Predicting influential mobile-subscriber churners using low-level user features. Automatika 56(4):522–534
25. Idris A, Rizwan M, Khan A (2012) 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
26. Idris A, Khan A, Lee YS (2013) Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. Appl intel 39(3):659–672
27. Maldonado S, Flores A, Verbraken T, Baesens B, Weber R (2015) Profit-based feature selection using support vector machinesGeneral framework and an application for customer retention. Appl Soft Comput 35:740–748

28. Sivasankar E, Vijaya J (2017) A study of feature selection techniques for predicting customer retention in Telecommunication sector. Int J Bus Inf Syst (In press)
29. Vijaya J, Sivasankar E (2017) An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing. Clust Comput 1–12
30. Xiao J, Xiao Y, Huang A, Liu D, Wang S (2015) Feature-selection-based dynamic transfer ensemble model for customer churn prediction. Knowl Inf Syst 43(1):29–51
31. Rajamohamed R, Manokaran J (2017) Improved credit card churn prediction based on rough clustering and supervised learning techniques. Clust Comput 1–13
32. Hudaib A, Dannoun R, Harfoushi O, Obiedat R, Faris H (2015) Hybrid data mining models for predicting customer churn. Int J Commun Netw Syst Sci 8(05):91
33. Huang Y, Kechadi T (2013) An effective hybrid learning system for telecommunication churn prediction. Expert Syst Appl 40(14):5635–5647
34. Duke University Case studies, Presentations and Video modules (2005): dataset available at http://www.fuqua.duke.edu/centers/ccrm/datasets/download.html/data
35. Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11(5):341–356
36. Amin A, Shehzad S, Khan C, Ali I, Anwar S (2015) Churn prediction in telecommunication industry using rough set approach. New trends in computational collective intelligence. Springer, Cham, pp 83–95
37. Inbarani HH, Bagyamathi M, Azar AT (2015) A novel hybrid feature selection method based on rough set and improved harmony search. Neural Comput Appl 26(8):1859–1880
38. Amin A, Anwar S, Adnan A, Nawaz M, Alawfi K, Hussain A, Huang K (2017) Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing 237:242–254
39. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
40. Kearns M, Valiant L (1994) Cryptographic limitations on learning Boolean formulae and finite automata. J ACM 41(1):67–95
41. Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intel 20(8):832–844