CrossMark

# Customer churn prediction in the telecommunication sector using a rough set approach

Adnan Amin[a], Sajid Anwar[a,*], Awais Adnan[a], Muhammad Nawaz[a], Khalid Alawfi[b],
Amir Hussain[c], Kaizhu Huang[d]

[a] Center for Excellence in Information Technology, Institute of Management Sciences, Peshawar, 25000 Pakistan
[b] College of Computer Science and Engineering, Taibah University, Saudi Arabia
[c] Department of Computing Science and Maths, University of Stirling, Scotland, UK
[d] Department of EEE, Xi'an Jiaotong-Liverpool University, Suzhou, China

## ARTICLE INFO

## ABSTRACT

Customer churn is a critical and challenging problem affecting business and industry, in particular, the rapidly growing, highly competitive telecommunication sector. It is of substantial interest to both academic researchers and industrial practitioners, interested in forecasting the behavior of customers in order to differentiate the churn from non-churn customers. The primary motivation is the dire need of businesses to retain existing customers, coupled with the high cost associated with acquiring new ones. A review of the field has revealed a lack of efficient, rule-based Customer Churn Prediction (CCP) approaches in the telecommunication sector. This study proposes an intelligent rule-based decision-making technique, based on rough set theory (RST), to extract important decision rules related to customer churn and non-churn. The proposed approach effectively performs classification of churn from non-churn customers, along with prediction of those customers who will churn or may possibly churn in the near future. Extensive simulation experiments are carried out to evaluate the performance of our proposed RST based CCP approach using four rule-generation mechanisms, namely, the Exhaustive Algorithm (EA), Genetic Algorithm (GA), Covering Algorithm (CA) and the LEM2 algorithm (LA). Empirical results show that RST based on GA is the most efficient technique for extracting implicit knowledge in the form of decision rules from the publicly available, benchmark telecom dataset. Further, comparative results demonstrate that our proposed approach offers a globally optimal solution for CCP in the telecom sector, when benchmarked against several state-of-the-art methods. Finally, we show how attribute-level analysis can pave the way for developing a successful customer retention policy that could form an indispensable part of strategic decision making and planning process in the telecom sector.

## 1. Introduction

Customer churn is one of the mounting issues of today's rapidly growing and competitive telecom sector. The focus of the telecom sector has shifted from acquiring new customers to retaining existing customers due to the associated high cost [1]. The retention of existing customers also leads to improved sales and reduced marketing cost as compared to new customers. These facts have ultimately resulted in customer churn prediction activity to be an indispensable part of telecom sector's strategic decision making and planning process.

Customer retention is one of the main objectives of Customer Relationship Management (CRM) and its importance has led to the development of various tools that support some important tasks in predictive modeling and classification. In recent decades, organizations are increasingly focusing on long-term relationships with their customers and observing a customer's behavior from time to time. They use various applied knowledge discovery in database (KDD) techniques [2–5] to extract hidden relationships between different entities and attributes in a flood of data banks. These facts have attracted many companies to invest in CRM to maintain customer information. Customer centric approach is very common, particularly, in the telecommunication sector for predicting customers' behavior based on historical data stored in CRM. To handle the mounting issue of customer churn, data maintained in such CRM systems can be converted into meaningful information that will help to identify customer's churn activities before the customers are lost; thereby, increasing customer strength [6]. Customer churn prediction modeling has been widely studied in various domains; such as financial services,

social network services, telecommunication, airlines, online gaming, and banking [7].

Researchers have also used various machine learning (ML) techniques (e.g. Random Forest, Balanced Random Forest, Rotation Forest and RotBoost) for dealing with the problem of customer churn prediction, but these ML techniques lack the required effectiveness for predicting customer churn [8]. On the other hand, the RST proposed by Pawlak [9] is an effective technique for discovering hidden rules, handling uncertainty and dealing with the unknown distribution of data [6]. However, RST application has not been widely studied in customer churn prediction, specifically in the telecommunication sector. Therefore, this study is an initial exploration of RST for CCP in the telecommunication sector by constructing more appropriate predictive classifiers that forecast churn prediction based on accumulated knowledge.

More specifically, an RST based benchmarking and empirical approach is proposed in this study, and compared with previous state-of-the-art machine learning based methods, in order to address the challenging CCP problem. As part of the aim to develop an improved CCP technique, we evaluate the performance of four well-known rule extraction algorithms (EA, GA, CA, and LA), in order to empirically determine which is most suitable for use with RST, and to extract more useful rules from hidden patterns in the telecom sector.

The extraction of decision rules based on RST, coupled with a best performing rule-generation algorithm, followed by appropriate attribute level analysis, can lead to a more strategic decision making and efficient planning process within the telecom industry. For example, based on the extracted rules, decision makers can readily devise and adopt new retention policies, and improve the overall performance of the organization. Finally, with attribute level analysis, decision makers can identify reasons for customer churns and develop an appropriate retention policy.

The rest of the paper is organized as follows: the next section presents customer churn and a review of CCP approaches; the preliminary study on RST is explored in Section 3 and evaluation methods are described in Section 4; the evaluation setup and experiments are discussed in Section 5 followed by results and comparisons in Section 6; the paper is concluded in the last Section 7.

## 2. Related work

### 2.1. Customer churn

Customer churn— shifting from one service provider to the next competitor in the market, is a key challenge in highly competitive markets and is very much observed in the telecommunication sector [5–7,10–12]. Customer churns are those targeted customers who have decided to leave a service provider, product, or even a company and shifted to a competitor in the market. Literature reveals the following three types of customer churns [13]:

1. **Active churner (volunteer):** those customers who want to quit the contract and move to the next provider.
2. **Passive churner (non-volunteer):** when a company discontinues service to a customer.
3. **Rotational churner (silent):** those customers who discontinue the contract without the prior knowledge of both parties (customer and company), where each party (e.g. customer or company) may suddenly terminate the contract without any notification.

The first two types of churns can be predicted easily with the help of traditional approaches in terms of the Boolean class value; however, the third type of churn is difficult to predict since there is a possibility of customers who might churn in the near future, for a variety of reasons that are either not known or difficult to predict. It should be the goal of the decision maker to decrease the churn ratio, for it is a well-known fact that existing customers are the most valuable assets for companies as compared to acquiring new ones [1]. Customer churn behavior has certain impacts on the company's performance which are summarized as follows: (i) a negative impact on the overall performance of the company, (ii) a potential cause for low sales because new/short-term customers buy fewer services, (iii) helps competitors to gain dissatisfied customers with business promotion(s), (iv) leads to revenue losses, (v) puts negative impact on long-term customers, (vi) increases uncertainty which reduces the ratio of possible new customers, (vii) attracting new customers is more expensive than retaining existing and (viii) risk to company's image in the competitive market with loss of customer base.

Churn prediction has been widely studied in the recent decade, particularly in the following domains: Open Social Network [14–19], Banking sector [20–24], Credit Card & a Financial Service provider [7,25,26], Online Gaming industry [27–29], Human Resource department of competitive organizations [12,30,31], Subscription service market [32,33], Question and Answer Q & A forums [34] and Insurance service providers [35]. It is clear from this discussion that customer churn as a problem is crucial for various organizations. Simultaneously, CCP is rapidly being observed in the telecommunication industry around the globe as well. Table 1 provides a brief overview of previous studies on CCP in the telecommunication industry.

### 2.2. Review of CCP approaches

A prediction model can be defined as the process of discovering hidden patterns from data and predicting future events [36]. The marketing strategy within competitive companies has evolved from a product-oriented approach to a customer-centric one, due to the advancements in the field of ML [1]. The database technologies not only provide useful information to the organization's decision makers about their past and current customers' behavior, but also, provide future prediction with the help of prediction modeling techniques.

Churn prediction is an equally alarming issue for the service sector. Keaveney [37] published an early and influential study by conducting a survey to find out reasons as to why customers switch services (i.e. customer churn). According to his study, the critical incidents between the customers and firms can be categorized into two broad cases; firstly, Core Service Failure (CSF) and secondly the Service Encounter Failure (SEF). It was also investigated that occurrence of any of the two cases (e.g. CSF or SEF) can be the reason for customer churn. Another study [38] has identified that service quality and customer response are two important drivers of churn and also established a link between these drivers and user churn.

Bloemer et al. [22] assumed that customer satisfaction and service quality are both directly proportional to each other. When customers are getting a higher degree of satisfaction, it has a positive impact on the overall performance of the company in a competitive market. Various machine learning techniques have been previously used for CCP, including the Support Vector Machine (SVM) [39–43], Neural Networks [2,13,41,44,45,46], Decision trees [4,41,44], Regression analysis [44], Naïve Bayes and Bayesian Network [4] and Neuro-Fuzzy [47]. However, the most important problem, specifically of which classification technique can be used to approach churn prediction in a more appropriate fashion, still remains an open research problem. Although the literature [39] argues that SVM is one of the state-of-the-art classification approaches due to its ability to efficiently model arbitrary nonlinearities, on the other hand, SVM also generates a black box model [48] which is considered its drawback. Similarly, several other studies [13,49] have reported that computationally expensive multi-layered neural network can outperform as compared to other conventional ML algorithms.

A benchmarking empirical approach is proposed in this study with the aim of achieving the best performance accuracy, through efficient

**Table 1**

Overview of previous studies on churn prediction in telecommunication industry.

| Ref. | Dataset | Techniques | Tool | Records | Outcomes |
|---|---|---|---|---|---|
| [51] | Malaysian subscriber in telecom industry database | Data mining by evolutionary learning (DMEL) | Evolutionary Learning process | 100000 subscribers data and 251 variables associated with subscriber | DMEL effectively discovered rules and accurately predicted churn in telecom data. |
| [10] | A proprietary dataset of Taiwan telecom company. | Decision tree (C5.0) and Neural Network back propagation (BPN). | Use LIFT and Hit Ratio to assess model performance, Exploratory data analysis | 160,000 subscribers with 14,000 churners. | Both DT and NN techniques can deliver accurately while the BPN performance is better than DT without segmentation. |
| [18] | Proprietary dataset of largest mobile operator about Call Detail Record | Social Network Analysis, J48 Decision Tree, link-based & collective classifications | Weka | 60 GB data which contains detail about voice call, SMS, value-added call etc. | Proposed a technique which efficiently predicted potential churns by underlying social network analysis. |
| [40] | CRM Dataset from Duke University | Support Vector Machine- Recursive Features Elimination (SVM-RFE) | MATLAB | 65000 customer data with 171 attributes | Introduced SVM-RFE for attributes selection in churn prediction which showed the satisfactory predictive result. |
| [52] | Six Real life proprietary European churn datasets (Bank1, Bank2, Mobile telecom, Pay TV, Newspaper, Supermarket) | Random Forests and Logistic Regression | Weka | Telecom Dataset which includes 100205 customers while 2983 customers are churns | Under-sampling increased the predictive performance while using advanced sampling technique CUBE has not shown better accuracy. |
| [53] | Collected data from in-house customer database, proprietary call record data from company & research survey | Back Propagation Neural Network algorithm. | Data mining using MATLAB | Dataset of 895 customers where 17.67% were Churns and 82.33% were non-churn | Predicted churn at risk which possibly may churn. |
| [2] | Telecom Dataset, UCI Repository, University of California, Irvine | Artificial Neural Network | Clementine Data mining SPSS | 2427 objects | Artificial Neural Network based model obtained 92% accuracy |
| [3] | Telecom Dataset, KDD Library, UCI Repository | C4.5 and RIPPER | AntMiner+ and Weka | Dataset of 5000 instances | ALBA with C4.5 & Ripper Shown highest accuracy while Ant Miner+ High performance. |
| [47] | Telecom Dataset, UCI Repository, University of California Irvine | Adaptive Neuro-fuzzy Inference system (ANFIS) | Fuzzy Logic toolbox MATLAB | Dataset contains 5000 customers | Neuro-Fuzzy obtained better accuracy, specificity, sensitivity than C4.5, RIPPER |
| [41] | The Dataset is obtained from an anonymous mobile service provider | Decision tree, neural network, and SVM. | Weka | Dataset of 5000 instances | Decision Tree accuracy: 77.9%, Neural Network accuracy: 83.7% SVM accuracy: 83.7% |
| [54] | 11 datasets of wireless telecom including KDD Cup2009, UCI, Duke and other private | NN, Li SVM, rbfSVM, linLSSVM, rbfLSSVM, Ripper, Part, C4.5, CART, ADT, RF, LMT, Bag, Boost, RBFN, VP, Logit, KNN10, KNN100, BN, NB | Weka, MATLAB, SAS, and R | Smallest data contains 2180 and largest up to 338874 observations | Oversampling does not improve classifiers' performance as on churn prediction in the telecom sector and a large group of classifiers found yields comparable performance. |
| [4] | Proprietary European telecommunication company dataset. | Naïve Bayes and Bayesian Network, C4.5 decision. | Weka | 106405 with 5.6% Churns | All predictive models performed with improved prediction rate. Decision tree performed better in high accuracy rate while other two achieved a higher true positive rate. |
| [8] | Publicly available Orange large & Cell2Cell datasets for telecom churn prediction. | RotBoost (RB), Random Forest, Rotation Forest and Decorate (DEC) ensemble with mRMR | Weka and MATLAB | 50,000 instances with 260 features where 190 features numerical and 70 nominal features. | mRMR returns more suitable features as compared to Fisher's ratio and f-score for ensemble and Random +rotation forest and RB+DEC with mRMR performed best. |
| [39] | Telecom Dataset, UCI Repository, University of California, Irvine | Support Vector Machine (SVM) Algorithm | IBM SPSS | Dateset for 3333 unique customers | SVM based model obtained 88.56% accuracy |
| [55] | Telecom Dataset includes the customer personal information data and the CDR data | logistic regression and Multilayer perceptron neural networks | MATLAB | Dataset contained 89,412 Instances with 9.7% instances as churners | Developed an efficient approach using SPA method as propagation process. |
| [56] | Asian Mobile telecom operator dataset. | Logistic Regression, Voted perceptron, | WEKA | Contains 2000 subscribers and 23 variables with 534 churns. | Proposed a hybrid learning model to predict churn which shows the most accurate result. |
| [57] | Telecom Dataset, UCI Repository, University of California, Irvine | Rough Set Theory | RSES | Dataset of 3333 Instances where 85.51% non-churn & 14.49% churns. | Rough Set as multi-class classifier can provide more accurate results for binary classification problem. |
| [46] | Telecom Dataset, UCI Repository, University of California, Irvine | SVM, Decision tree, Artificial Neural Network, Naïve Bayes, Regression Analysis, boosting | Package C5.0 in R-language for statistical computing | Dataset of 5000 Instances | The best overall classifier was the SVM-POLY using AdaBoost. |

extraction of appropriate decision rules from hidden existing patterns and attribute level analysis. Based on these rules, the decision maker can easily adopt new retention policies and improve the overall performance of the organization. With attribute level analysis, the decision maker is able to identify the reasons for customer churns and, develop a corresponding retention policy as well. Decision rules which require additional information related to the data e.g. value of possibility in the fuzzy set, grade of membership, basic probability in Dempster-Shafer theory, or probability in statistics, will result in less realistic decisions in comparison to decisions based on decision rules which do not require additional information. RST is unique as it does not require any additional information. Hence, the ultimate decision made, based on RST, is argued to be more realistic [50].

## 3. Rough set theory

RST can be used as a mathematical tool to approach vagueness [9]. RST philosophy is based on the assumption that we can associate information i.e. knowledge, data or information with every object of the universe of discourse (e.g. also known as $U$). An RST has a precise concept of lower & upper approximation, and the boundary region. The boundary region separates the lower approximation from upper approximation (i.e. boundary-line). For example, those objects that cannot be classified with certainty as elements of either the lower or upper bounds. It is clear that borderline instances cannot be characterized in terms of available knowledge about the elements. Therefore, any rough concept is replaced by either lower or upper approximation of vague concept [9,58,59].

Mathematically, the concepts of lower or upper approximation and boundary region have been defined as: suppose set $X \subseteq U$ and $B$ is an equivalence relation and $U$ is the universe of objects (i.e. the partition of the universe set U to create a new subset of interest from U which has the same value of outcome attribute) in information system $IS = (U, A)$ of non-empty, finite set $U$ and $A$, where $U$ is the universe of objects and $A$ is a set which contains attributes. Then $LB = \cup \{Y \in U/A: Y \subseteq X\}$ is a lower approximation and an exact member of $X$ while $UB = \cup \{Y \in U/A: Y \cap X \neq \emptyset\}$ is upper approximation which is possibly a member of $X$. $BR = UB - LB$ is the boundary region. The whole RST based classification system can be visualized as (see Fig. 1.).

### 3.1. Decision table

The special case of Information system (IS) is known as decision table, which is usually known as training sets or samples in ML. An *IS* can be represented by a data table with rows "*R*" and columns "*C*", where $R$ is labeled by objects and $C$ is labeled by attributes. Formally, an information system is $IS = (U, A)$ where $U$ is a non-empty finite set of instances called the universe is and $A$ is a non-empty finite set of attributes or properties i.e $A = \{a_1, a_2, a_3........a_n\}$ such that $a{:}U \rightarrow V_a$



**Fig. 1.** RST based classification model.

for every $a \in A$ where $V_a$ represents value set of attribute $a_i$. A decision system is any information system of the form $S = (U, C \cup \{d\})$, where $C$ is conditional attributes and $d \notin C$ is the decision attribute. The union of $C$ and $\{d\}$ are elements of Set A [60].

### 3.2. Indiscernibility relation

The notion of a concept approximation is precisely based on the similarity between objects, which is referred to as indiscernibility (IND) relation. A decision table may contain a large number of unnecessary or redundant objects or attributes. For example, if objects (customers) are suffering from certain critical behavior i.e. churns; symptoms of the churn behavior from information can be obtained about the customers. These objects (customers) can be characterized by the same information, and are indiscernible (similar) in the pattern of the existing information about them. The RST indiscernibility approach is defined relative to a given set of properties (attributes). Any set of IND objects is called an elementary set which is also known as a basic granule of information about the universe, whilst any union of elementary sets is called a crisp set. If the union set is not a crisp set, then it is referred to as vague (rough) [9,58,59].

Suppose we are given a pair of non-empty $IS = (U, A)$, finite sets $U$ and $A$, where $A$ contains a set of attributes (i.e. $a{:}U \rightarrow V_a$, where $V_a$ is the set of values of attributes $a$) and $U$ is the universe of objects. Any subset $B$ of $A$ determining a binary relation $I(B)$ on Universe $U$ is called an indiscernibility relation which can be defined as: $xI(B)y$ IFF $a(x) = a(y)$ for every $a \in B$,

where $a(x), a(y)$ are values of attribute $a$ for object $x$ and $y$ respectively. Obviously, $I(B)$ is an equivalence relation which is determined by $B$. If $(x, y) \in I(B)$ then we can say that $x$ and $y$ are B-indiscernible where $I(B)$ an equivalence classes or B-elementary set or rough set. In RST the elementary set (granules) is a concept of available knowledge about the reality [61]. For any subset of attributes $B \subseteq A$ indiscernibility relation $IND(B)$ is defined as follows: If $IND(B) = IND(B - \{a\})$ (i.e. recall equivalence relation) then $a \in B$ is dispensable otherwise indispensable in $B$ while set $B$ can be called independent if all attributes are indiscernible. If $(i,j) \in U \times U$ belongs to $IND(B)$, then we can say that $i$ and $j$ are indiscernible by attributes from $B$.

### 3.3. Reduction of attributes and core set

The reducts (reduction) and discernibility relation are the two key notions in the conventional RST [60]. The data table contains some superfluous data. Therefore, a reduction process was used to remove some data from the decision table whilst preserving its basic properties [61]. Let us mathematically express the reduction of attribute idea more precisely by: *If* $C, D \subseteq A$, where *C* is a set of condition attributes and $D$ is a set of decision attribute and $X \subseteq C$ is a D-reduct (reduct with respect to $D$) of $C$, then $X$ is a minimal subset of $C$ such that $\gamma(C, D) = \gamma(X, D)$. Every element of core belongs to a reduct set in such a way that the core contains the most important subset of attributes and none of its element can be removed due to effect on the classification power [61]. So the core is the intersection of all reducts $Core(B) = Red(B)$. Where $Red(B)$ is a set of all reducts of $B$.

### 3.4. Cut and discretization

Discretization is a process that generates partitions of value sets of conditional attributes, into certain groups or intervals. Discretization process is not specific to the RST approach; however, most of the tree induction and rules extraction algorithms which are based on discretization also perform well in data mining and ML [61]. As a result, this process can produce a new consistent decision table in such a manner that all the original values of attributes, which belong to objects, are grouped into intervals where the attribute value lies [60]. It
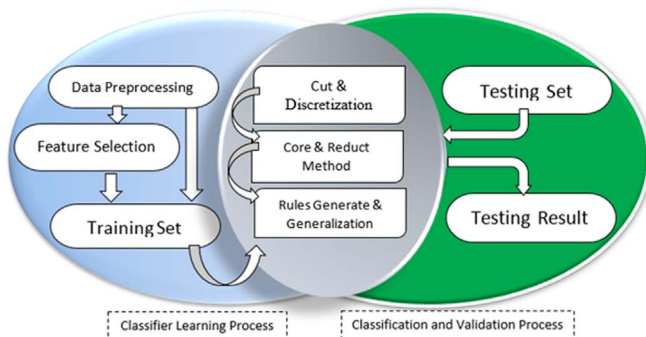
may ultimately reduce the size of value sets of attributes and thus reduces the computational cost. On the other hand, the grouping of attribute's data in discretization process is actually based on the calculated cuts, where the continuous variables are converted into discrete attributes or into intervals i.e. grouping of values [6]. The idea of cut can be incorporated in the discretization process. Actually, cut mostly appears in the discretization process as a pair (a, c), (where *a* is a continuous value and *c* is a cut to split the value into two disjoint sub-intervals [6]). For example, there may exist some unseen objects which cannot match with the rules or can slow down the ML process and increase computational costs. Therefore, cut and discretization methods are used to obtain a high quality of classification [60].

### 3.5. Rules generation

Decision rules are often denoted as "IF C THEN D" where C is the set of condition attributes and *D* represents the decision attribute in the decision table [61]. Given two unary predicate formulae are $\propto(x)\ and\ \beta(x)$, where x executes over a finite set U. Łukasiewicz [62], defined a value i.e. $\frac{card\,(\|\alpha(x)\|)}{card\,(U)}$, assign to $\propto(x)$ where $\|\propto(x)\| = \{x \in U{:}x\ satisfies \propto \}$while the fractional value is assigned to implication $\propto(x){=}>\beta(x)$ is then $\frac{card\,(\|\propto\,(x)^{\beta(x)}\,\|)}{card\,(\|\propto\,(x)\|)}$ with assumption that $\|\propto(x)\| \neq \varnothing$. The decision rules can be constructed by overlaying the reduct sets over the decision table. Mathematically, it can be expressed as; $(a_{i1} = v_1)^\wedge.....^\wedge(a_{ik}= v_k) => d = v_d$, where $1 \le i_1 < ... < i_k \le m$, $v_i \in V_{ai}$: for simplification we can state it represented in IF-Else statement i.e., IF C THEN D; where C is set of conditions and D is decision value. To extract the decision rules, the following four well-known rules generation algorithms are used [60]:

– **EA:** It takes subsets of features incrementally and then returns the reducts set and minimal decision rules. The generated decision rules are those rules which have minimal descriptors in the conditional attributes. It needs more concentration because it may lead to extensive computations in the case of a complex and large decision table. It is based on a Boolean reasoning approach [63].
– **GA:**It is based on order-based GA coupled with a heuristic. It is used to reduce the computational cost in a large and complex decision table [64,65].
– **CA**: It is the customized implementation of the LEM1 algorithm idea and is implemented in the RSES covering method. It was introduced by Jerzy Grzymala [66].
– **LA**: It is a divide and conquer technique paired with lower and upper approximation of RST and is based on local covering determination of each object from the decision class [66,67].

Since RST is employed as a base classifier in the proposed study, we have provided a brief overview of RST to help the reader understand the basic terminologies used in RST (specifically, the decision table, indiscernibility, reduct & core sets, cut & discretization and decision rules). In section 2, the reasons for using RST in the proposed study are outlined. Moreover, the implementation of RST in the CCP problem is illustrated pictorially in Fig. 2. Wherein it depicts the three major parts: (i) data preprocessing (creating a decision table), (ii) training process (or applying rules generation algorithms followed by reduct and core sets concepts of RST), and (iii) classification process to validate the prediction performance of the RST based approach for the CCP problem. Each part is explained in section 5 with step-wise details in the context of the fundamental study of RST. This study is an extension of our previous work [80]. The next section describes the evaluation measures which are used to evaluate the performance of the proposed approach.

## 4. Evaluation measures

It is nearly impossible to build a perfect classifier or a model that could perfectly characterize all the instances of the test set [52]. To assess the classification results, we count the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The FN value actually belongs to Positive P (e.g. TP + FN = P) but is wrongly classified as Negative N (e.g. TN + FP = N); while, FP value is actually part of N but wrongly classified as P. The following measures were used for the evaluation of proposed classifiers.

– **Sensitivity/recall:** it measures the fraction of churn customers who are correctly identified as true churn.

$$Sensitivity\ (Recall)\ =\ \frac{\mathbf{TP}}{\mathbf{P}} \tag{1}$$

– **Specificity:** it measures the fraction of true non-churn customers who are correctly identified.

$$Specificity\ =\ \frac{\mathbf{TN}}{\mathbf{N}}\ =\ > \mathbf{1-Specificity}\ =\ \frac{\mathbf{FP}}{\mathbf{N}} \tag{2}$$

– **Precision:** it is characterized by the number of correctly predicted churns over the total number of churns predicted by the proposed approach. Formally, it can be expressed as:

$$Precision\ =\ \frac{\mathbf{TP}}{\mathbf{TP}\ +\ \mathbf{FP}} \tag{3}$$

– **Accuracy:** overall accuracy of the classifier can be calculated as:

$$Accuracy\ =\ \frac{\mathbf{TP}\ +\ \mathbf{TN}}{\mathbf{P}\ +\ \mathbf{N}} \tag{4}$$

– **Misclassification error:** it is referred to as a classification error where an instance is falsely classified to a class to which the instance does not belong. Different types of misclassification errors can be calculated as:

$$Misclassification = \mathbf{1{-}Accuracy} \tag{5}$$

$$Type - I\ Error = \mathbf{1-Specificity}\ =\ \frac{\mathbf{FP}}{\mathbf{FP + TN}} \tag{6}$$

$$Type - II\ Error = \mathbf{1-Sensitivity}\ =\ \frac{\mathbf{FN}}{\mathbf{TP + FN}} \tag{7}$$

– **F-Measure:** a composite measure of precision and recall to compute the test's accuracy. It can be interpreted as a weighted average of precision and recall.

$$F - Measure = \mathbf{2.}\,\frac{\mathbf{Precision.\ Recall}}{\mathbf{Precision + Recall}} \tag{8}$$

– **Coverage:** the ratio of classified objects that are recognized by a classifier from the class to the total number of objects in the class. Where C is classifier, A is a decision table, Match A(C) is a subset of objects in A that areclassified by classifier C.

$$Coverage\ \text{A(C)} =\ \frac{|Match\ A(C)|}{|A|} \tag{9}$$

## 5. Evaluation setup

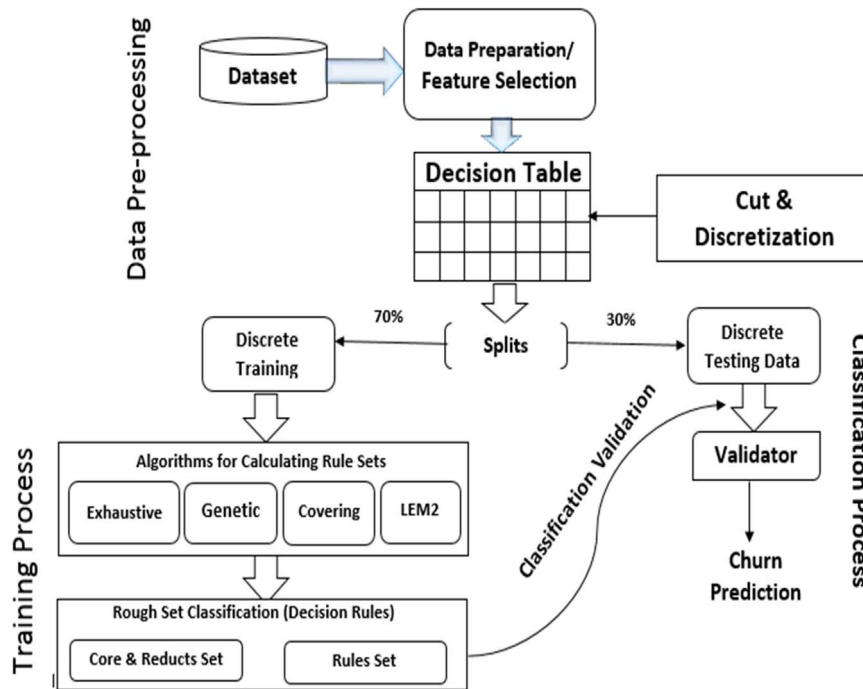In this section, we have evaluated the performance of RST based on

**Fig. 2.** An evaluation setup for customer churn prediction.

four rules generation algorithms for customer churn prediction using the RSES toolkit [60]. An analytical environment was setup to perform the proposed technique as shown in Fig. 2. RSES toolkit is very helpful to: (i) convert dataset into decision table (ii) apply Cut and Discretization (iii) extracting the decision rules set from the training set, and (iv) validate the result. These experiments were carried out to fulfill the objectives of the proposed study as well as to address the following points:

**P1:** Which features are more indicative for churn prediction in the telecom sector?
**P2:** Which algorithm (EA, GA, CA, and LA) is more appropriate for generating rules sets for RST classification approach in the telecommunication sector?
**P3:** What is the predictive power of the proposed approach for churn prediction in the telecom sector?
**P4:** Can the derived rules help the decision makers in strategic decision making and planning process?

### 5.1. Data preparation and feature selection

Evaluation of data mining approaches on a publicly available dataset has many benefits in terms of comparability of results, ranking techniques and evaluation of existing methodologies with new ones [68]. For this study, we have used a publicly available dataset. Description about the used dataset can be obtained from the URL[1]. The dataset consists of 3333 instances, where 85.5% customers are Non-churn (NC) and 14.49% are Churn (C). The number of churns was much smaller than non-churns customers in the selected dataset, which can make it difficult for the churn prediction classifier during the learning process. The training set contains 2333 customers, where 85.16% are NC and 14.8% are C. The test set contains 1000 instances, where 86.3% are NC and 13.7% are C. Table 2 reflects descriptive statistics of the target dataset that are selected after applying "Information Gain Attribute Evaluation" method.

Data preparation and feature selection are important steps in the

knowledge discovery process. In order to identify those variables or attributes, from a large number of attributes in a dataset, that are relevant and will reduce the computational cost [64,69,70], the selection of the most appropriate attributes from the dataset in hand was carried out using a feature ranking method known as "Information Gain Attribute Evaluator" using a Weka toolkit [71]. It evaluates the attributes worth through the information gain and measurement procedure as per the class value. It ranks the attributes and selects the top ranked attributes, which significantly improves the computational efficiency and classification. After feature ranking, it includes most relevant and ranked attributes in the decision table.

### 5.2. Preparation of a decision table, cut and discretization

The preparation of the decision table is an important stage of the proposed study. The decision table which consists of objects, conditional attributes, and decision attributes is organized in Table 3.

Cut and discretization is the plausible approach to reduce the dataset horizontally in order to handle the large data efficiently. It is a common approach used in the RST where the variables containing continuous values are partitioned into a finite number of intervals or groups [72]. The cut and discretization process was carefully performed on the prepared decision table using the RSES toolkit. The cuts are added in the decision table using the toolkit at every iteration and generates fewer number of cuts [60]. For example, in this study, the cuts of attribute "*Day_Mins*" were grouped after the discretization process as listed in Table 4. The first column shows the groups, that are represented by numeric data for simplicity purposes, and are listed in ascending order. The second column represents the intervals that are obtained after the discretization process. The third column is the count of the attribute's values that fall into certain groups; while, the last column is the percentage of a variable's value in each interval. It is clear from Table 4 that the value of *Day_Mins* has been changed from the continuous nature into 14 different intervals or groups, after the cut and discretization process.

### 5.3. Training and test sets

In data mining, validation is extremely important to ensure that the

---

[1] ("Data Source Link http://www.sgi.com/tech/mlc/db/")

**Table 2**
Statistical information about ranked attributes.

| Attributes | Distinct counts | Min value | Max value | Means | StdDev | Ranks | Values | Description |
|---|---|---|---|---|---|---|---|---|
| Intl_Plan | 2 | – | – | – | – | – | Y(323), N(3010) | Whether a customer subscribed international plan or not. |
| VMail_Plan | 2 | – | – | – | – | – | Y(922), N(2411) | Whether a customer subscribed Voice Mail plan or not. |
| VMail_Msg | 46 | 0 | 51 | 8.099 | 13.68 | 0.0096 | – | Indicates number of voice mail messages. |
| Day_Mins | 1667 | 0 | 350.8 | 179.8 | 54.46 | 0.0448 | – | No. of minutes that a customer has used in daytime. |
| Day_Charges | 1667 | 0 | 59.64 | 30.56 | 9.259 | 0.0448 | – | A continuous variable that holds day time call charges. |
| Eve_Mins | 1611 | 0 | 363.7 | 200.9 | 50.74 | 0.0081 | – | No. of minutes that a customer has used at evening time. |
| Eve_ Charges | 1440 | 0 | 30.91 | 17.68 | 4.311 | 0.0081 | – | A continuous variable that holds evening time call charges. |
| Intl_Mins | 162 | 0 | 20 | 10.23 | 2.792 | 0.0115 | – | No. of minutes used during international calls. |
| Intl_Calls | 21 | 0 | 20 | 4.479 | 2.461 | 0.0096 | – | Total No. of calls used as international calls. |
| Intl_Charges | 162 | 0 | 5.4 | 2.765 | 0.754 | 0.0115 | – | A continuous variable that holds international call charges. |
| CustSer_Calls | 10 | 0 | 9 | 1.563 | 1.315 | 0.1244 | – | Total No. of calls made a customer to customer service. |
| Churn? | 2 | – | – | – | – | – | C (483), NC(2850) | Class Label |

**Table 3**
Organization of attributes for decision table.

| Sets | Description |
|---|---|
| Objects | {3333 distinct objects} |
| Conditional attributes | {Intl Phan, VMail_Plan, VMailMsg, Day_Mins, Day Charges, Eve_Mins, Eve_Charges, Intl_Mins, Intl_Calls, Intl_Charges, CustServ_Calls} |
| Decision attribute | {Churn? } |

**Table 4**
Cuts Distribution of attribute *Day_Mins*.

| Group# | Intervals | Count | Percentage |
|---|---|---|---|
| 1 | (263.55, 281.05) | 115 | 3.45% |
| 2 | (151.05, 163.45) | 295 | 8.85% |
| 3 | (237.85, 251.85) | 158 | 4.74% |
| 4 | (281, *) | 103 | 3.09% |
| 5 | (163.45, 178.15) | 339 | 10.17% |
| 6 | (217.65, 237.85) | 337 | 10.11% |
| 7 | (178.15, 189.25) | 235 | 7.05% |
| 8 | (251.85, 263.55) | 95 | 2.85% |
| 9 | (108.8, 151.05) | 687 | 20.61% |
| 10 | (195.45, 208.85) | 302 | 9.06% |
| 11 | (189.25, 195.45) | 168 | 5.04% |
| 12 | (*, 78.65) | 109 | 3.27% |
| 13 | (78.65, 108.8) | 202 | 6.06% |
| 14 | (208.85, 217.65) | 188 | 5.64% |

**Table 5**
Parameter configuration for the two settings of GA used for Feature Reduction.

| GA parameter | Value |
|---|---|
| Mutation probability | 0.033 |
| Crossover probability | 0.6 |
| Population | 50 |
| Maximum generation | 100 |

**Table 6**
Statistics about rules induced using four methods.

| Description | Methods for calculating rules | | | |
|---|---|---|---|---|
| | EA | GA | CA | LEM2 |
| Total no. of rules | 4184 | 9468 | 369 | 625 |
| # of rules induced that classifies customer as churn | 1221 | 2674 | 122 | 160 |
| # of rules induced that classifies customers as non-churn | 2963 | 6715 | 247 | 465 |

**Table 7**
Evaluation of four rules generation methods through RST classification approach.

| | EA | GA | CA | LA |
|---|---|---|---|---|
| **Coverage** | 1 | 1 | 0.64 | 0.668 |
| **Precision** | 0.72 | 0.86 | 0.47 | 0.67 |
| **Recall** | 0.74 | 1.00 | 0.50 | 0.73 |
| **Misclassification** | 0.074 | 0.019 | 0.122 | 0.067 |
| **Accuracy** | 0.926 | 0.981 | 0.878 | 0.993 |
| **Specification** | 0.96 | 0.98 | 0.93 | 0.96 |
| **F-measure** | 0.726 | 0.925 | 0.487 | 0.698 |

prediction model not only remembers the instances it was trained on, but also performs equally well for unseen new instances [73]. One way to overcome this problem is not to use the entire dataset for the classification learning process. It can also exclude some data that could be used to validate the performance of a learned classifier for the new data. This overall idea of model evaluation or classifier evaluation is called cross-validation. Literature reveals several validation methods to evaluate the performance of classifiers.

Hadden [74] divided the dataset into training and validation sets. The training set contained 50% of non-churn and 50% of churn; whereas, the validation set contained 30% churns and 70% non-churn. Abbasimehr [47] has validated his work by splitting the data set into 67% and 33% for training and test sets respectively, and also over-sampled the proportion of churns in the training set by 50% churn and 50% non-churn in order to provide a better predictive model. Shaaban et al. [41] divided the dataset by 80:20% ratio where 80% data was used for training purposes and the remaining 20% data was used for testing purposes. In the proposed study, we divided the dataset into two sets: the training set and the test sets as shown in Fig. 2. After multiple splitting attempts during experiments, we concluded that the best performance is obtained if the split factor parameter is set to 0.7 as we avoid reusing the training set for the test; no over-fitting is involved in this case.

### 5.4. Reduct and decision rules sets generation

The decision rules can be obtained from the training set by selecting either of the rules generation methods (EA, GA, CA and LA). Where EA and GA scan the training set object-by-object and generate rules sets by matching the objects and attributes with reducts. The CA and LA can induce rules sets without matching with reduct sets using the RSES toolkit.

GA encodes the problem domain objects into chromosomes and applies selection, mutation, and a reproduction operator from the theory of natural selection, on the population of chromosomes which are candidate solutions. In each generation, the fittest chromosomes are retained after evaluation through a fitness function, and eventually
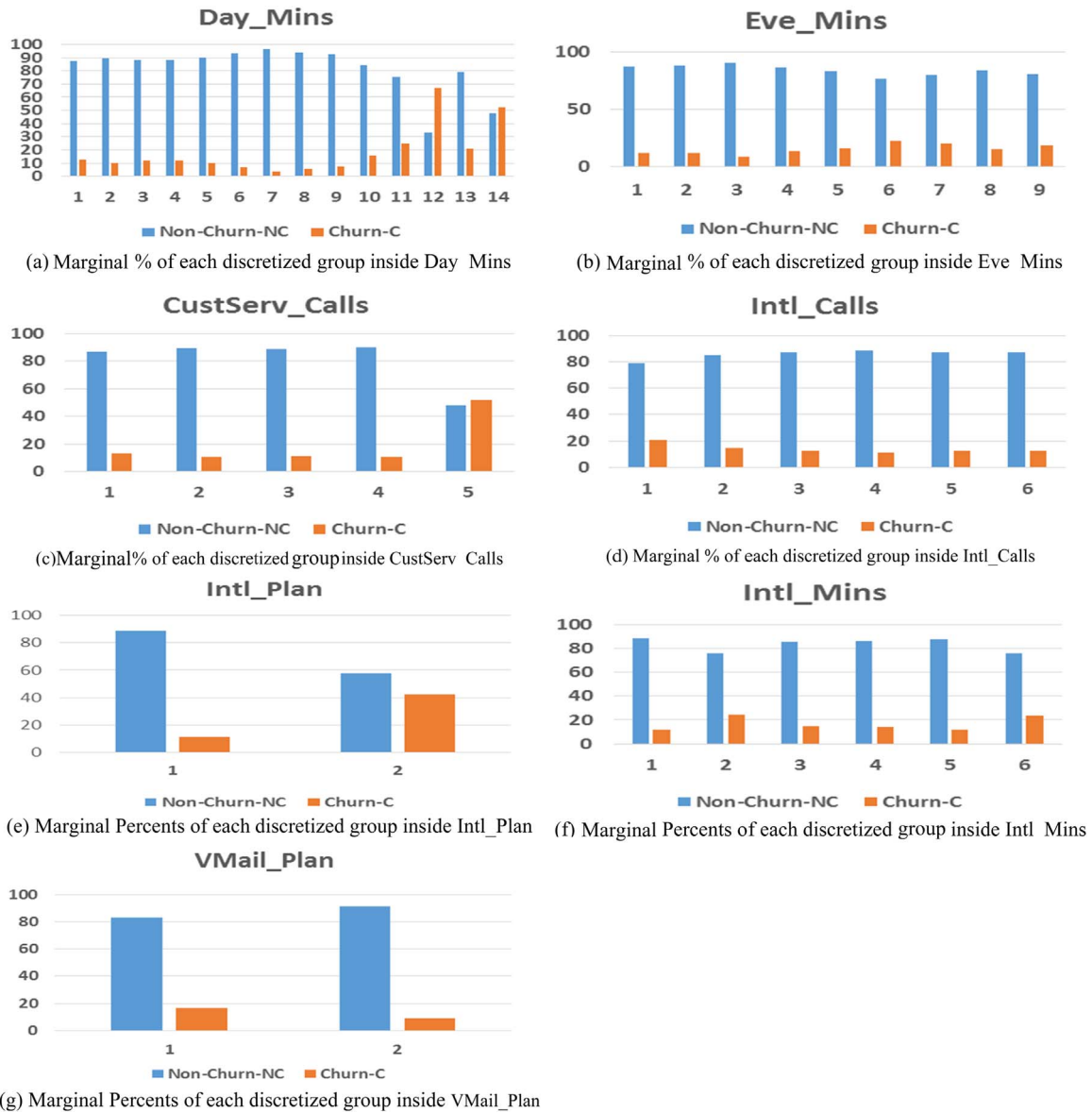
(a) Marginal % of each discretized group inside Day  Mins

(b) Marginal % of each discretized group inside Eve  Mins

(c)Marginal% of each discretized group inside CustServ  Calls

(d) Marginal % of each discretized group inside Intl_Calls

(e) Marginal Percents of each discretized group inside Intl_Plan

(f) Marginal Percents of each discretized group inside Intl  Mins

(g) Marginal Percents of each discretized group inside VMail_Plan

**Fig. 3.** Marginal percent of each discretized group inside all attributes. The discretized groups of each variable are labeled by integer number e.g. 1, 2,.....N. on the x-axis in the graphs while y-axis indicates marginal percent.

the algorithm converges after defining a number of generations. The GA applied to CCP problem using subset evaluation as the fitness function. Each rule generated is passed through the genetic operators and the best amongst the rules, having maximum support and confidence, are selected as final candidate solutions. The rules having lesser support and confidence are simply discarded. The parameters for the GA are described in Table 5.

The GA for computing reducts has a probability of 60%, starting with an initial population of 50 chromosomes. The chromosomes are designed on the basis of the number of features in the dataset with one bit allocated for a class label. The GA converges in the span of 100 generations and produces an exhaustive rule base method containing more than nine thousand rules. In the proposed study, important decision rules were extracted from the training set through these four different rules generations algorithms. The decision rules set, specifies the rules in the form of "if C then D" where C is a condition and D refers to decision attribute. For example:

*IF Day_Mins (108.8, 151.05) &*
    *Eve_Mins (207.35, 227.15) &*

*CustServ_Calls (3.5, \*)*
    *ThenChurn= (True)*

Based on these simple and comfortably interpretable rules, the decision makers can easily understand the flow of customer churn behavior and they can adopt a more suitable strategic plan to retain their customer churn. The total number of generated decision rules induced from the training set is summarized in Table 6.

## 6. Results and discussion

This section represents the performance of classifiers (in subsection 6.1), the analysis of discretized groups (in subsection 6.2), analysis of features (in subsection 6.3) and finally, RST approximation analysis (in subsection 6.4). We have also performed a comparison of the proposed approach with other related techniques that are applied to the same dataset in section 6.5.

### 6.1. Performance evaluation of classifiers

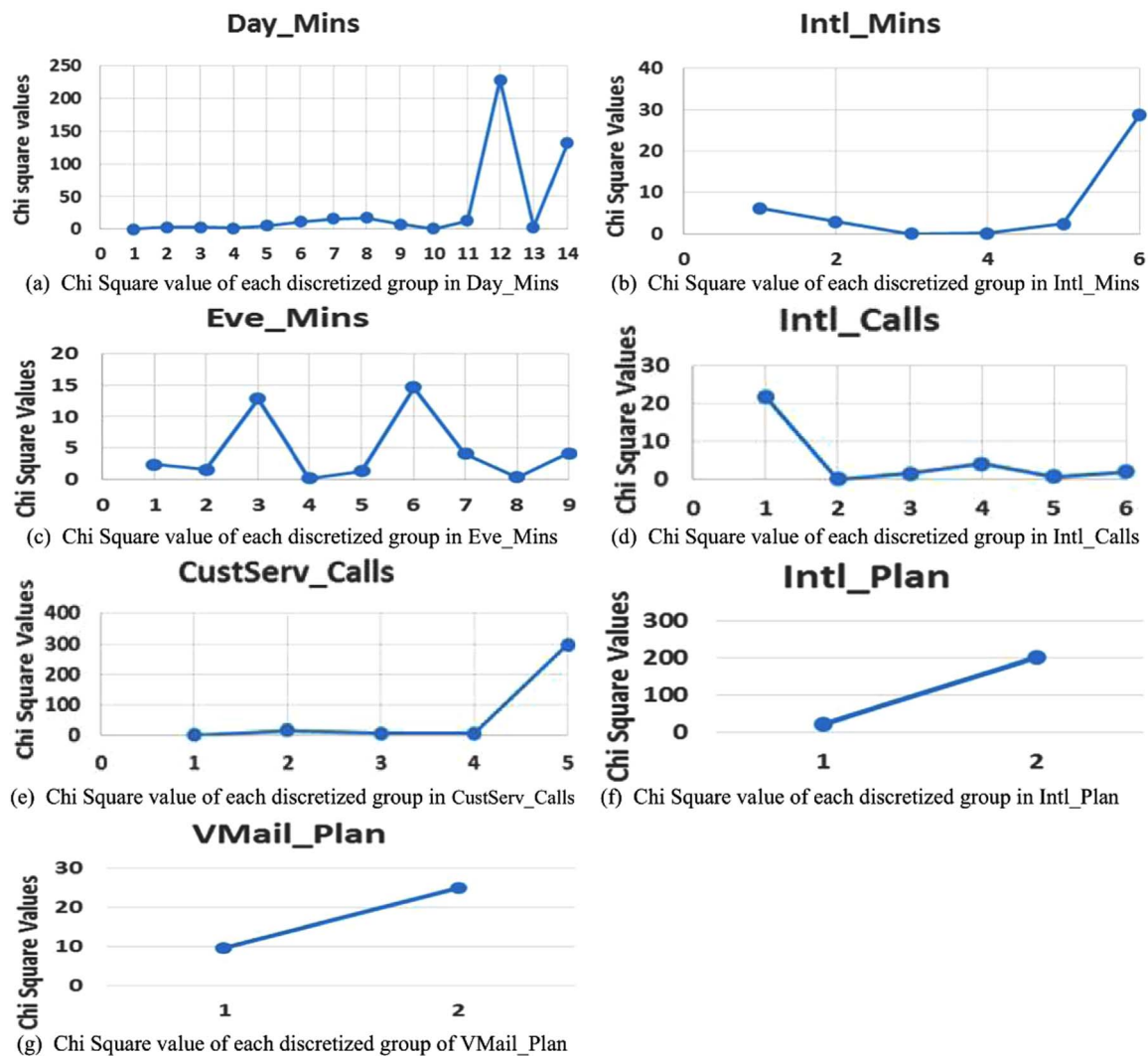We have evaluated four different algorithms for rules generation

**Fig. 4.** Contributes an amount to Chi Square Statistic of all attributes.

**Table 8**
Summary of statistical most significance discretized groups.

| Attributes | Group # | Chi Square |
|---|---|---|
| Day_Mins | 12 | 229.09 |
| Intl_Mins | 6 | 28.38 |
| Eve_Mins | 6 | 14.69 |
| Intl_Calls | 1 | 21.69 |
| CustServ_Calls | 5 | 298.08 |
| Intl_Plan | 2 | 203.24 |
| Vmail_Plan | 2 | 25.16 |



**Fig. 5.** Reflect sensitivity on y-axis and 1-specificity on the x-axis for each individual feature.

with an RST based classification approach. All these four methods are applied to the same telecom dataset. Table 7 reflects the comparison of these four rules generation algorithms (i.e. GA, CA, EA, LA) with RST based classification. It was investigated that the GA with the RST has shown more suitable predictive capacity which is reported to P3 (i.e. Section 5). LA gives maximum accuracy which is about 0.993; however, it has a coverage of 66.8% of customers (which means it has only classified 668 instances while 332 customers were ignored). Similarly, covering algorithm has only classified 64% customers with 0.878 accuracy which is the least of accuracy among four rules generation algorithms. Although, the EA achieved less accuracy (e.g. 92.6%) as compared to LA; nonetheless the EA method performed better than both algorithms (i.e. LA and CA) in terms of coverage, recall, and F-measures. On the other hand, Table 7 reflects that the overall best
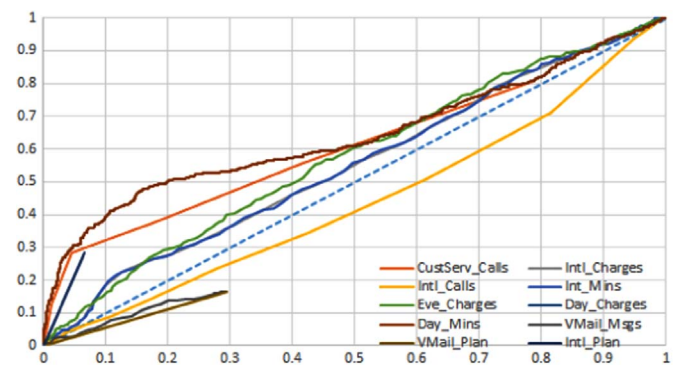
**Table 9**
Approximation Analysis.

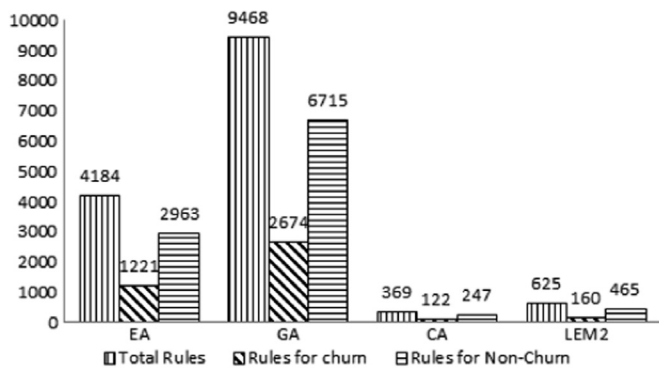| Decision class | No. of objects | Lower approximation | Upper approximation | Accuracy |
|---|---|---|---|---|
| NC | 2850 | 2850 | 2850 | 1.000 |
| C | 483 | 483 | 483 | 1.000 |

Fig. 6. Summary of total rules induced by each rule generation algorithm where vertical-line bars represent the total number of rules induced, diagonals-line bars reflect the total number of rules induced for classifying churns, and finally, the horizontal line bar illustrates the total number of rules induced for classifying non-churns objects.

discretized group towards customer churn in each attribute.

The following graph represents the statistical significance of each discretized group in the corresponding attribute.

It is clear from Fig. 3 that 66.99% customers are churns compared to 33.01% of non-churn customers in group 12 in *Day_Mins*. This finding indicates that those customers who have utilized maximum minutes in daytime have become churn. Similarly, group 6 in *Intl_Mins* and group 5 in *CustServ_Calls* indicate that those customers who used a maximum number of international minutes and maximum calls to the customer service center are converted into churn category respectively. Based on the statistical analysis, we suggest that group 12 in *Day_Mins*, group 6 in *Intl_Mins*, and group 5 in *CustServ_Calls* need more concentration of the decision makers to retain as non-churn customers and make a policy to re-attract the customer who churned. On the other hand, group 6 in *Eve_Mins* has a higher ratio of churns (i.e. 23.08%) which reflects that those customers used evening minutes, between 227 and 244, have become churned customers. In contrast, it

**Table 10**
Comparison of predictive performance of proposed & previous approaches applied to the same dataset.

| Techniques | Recall | Precision | Misclassification | Accuracy | F-Measure |
|---|---|---|---|---|---|
| [2] | 81.75 | 66.27 | 6.76 | 93.2 | 73.2 |
| [41] | 76.47 | 80.60 | 22.10 | 77.9 | 78.5 |
| [41] | 83.90 | 83.40 | 16.30 | 83.7 | 83.7 |
| [41] | 83.37 | 84.20 | 16.30 | 83.7 | 83.8 |
| [39] | 52.12 | 81.90 | 14.37 | 85.6 | 63.7 |
| [39] | 39.15 | 79.05 | 22.14 | 77.9 | 52.4 |
| [39] | 59.44 | 80.95 | 11.44 | 88.6 | 68.5 |
| [46] | 30.49 | 71.43 | 29.47 | 70.5 | 42.7 |
| [46] | 78.45 | 91.72 | 3.15 | 96.85 | 84.6 |
| Proposed | **86.0** | **100** | **1.90** | **98.1** | **92.5** |

performance is achieved by GA for generating the rules set using RST based classification approach for churn prediction in the telecommunication sector. GA correctly predicted 86% true churn, fully classified false churn and with 98.2% overall accuracy achieved with minimum misclassification error, i.e., about 1.9% with 100% coverage of instances as well. Therefore, GA indicating as the best approach among these targeted algorithms (i.e. CA, LA, and EA) with RST based prediction of customer churn which is reported to P2 (i.e. Section 5).

### 6.2. Discretization groups' analysis

We have performed Chi-Square statistical test in order to find the statistical significance in terms of the association between the discretized groups and expected values of customer churns. We have set the significance level to 0.0500, and calculated the *p-Value* for each attribute, which is 0.0000. The *p-value* is less than significance level that this suggests the exclusion of the null hypothesis (e.g. no association exists) and support of the alternative hypothesis (e.g. association exists). While performing the discretization process, we also observed the frequency of customer churn behavior inside every discretized group of each variable that provide some interesting information to strategic decision makers as shown in Fig. 3, Fig. 4 and Table 8.

In Fig. 3, the discretized groups of each variable are labeled by integer number e.g. 1, 2,.....N. on x-axis in the graphs while y-axis indicates marginal percent (%) of minutes used (i.e. for *Day_Mins, Eve_Mins, Intl_Mins*), calls (i.e. for *CustSer_Calls, Intl_Calls*) and package activated (for *VMail_Plan, Intl_Plan*) of customers in each discretized group. To find the most significant discretized group with respect to target class (i.e. churn) in attributes, we conducted the Chi-Square, a statistical test. Table 8 describes the most significant

is investigated that group 1 in *Intl_Calls*, have a high ratio of churn (i.e. 20.84%) which shows that those customers who have utilized minimum international calls have become churn. Similarly, group 1 in *VMail_Plan* has also a maximum number of churn (i.e. 16.72) which shows that those customers who have not activated the voice mail plan have more tendency of churn. Finally, those customers who have activated the international call plan have a high ratio of customer churn (i.e. 42.41). The decision makers can use this information while designing their customer retention policy which is reported to P4.

### 6.3. Analysis of attributes

We have also examined the features' sensitivity to determine which features are more indicative for churn prediction in telecom sector in order to address P1. Fig. 5 shows the point of inflection of each variable, such as *Intl_Plan, Day_Mins, Day_Charges, Eve_Mins, Eve_Charges, Intl_Charges* and *CustServ_Call*, after these inflection points of each variable, the churn rate increases at lesser rate and the curve reflects the churn behavior; while, those features which are below the curve line e.g. *Intl_Calls, VMail_Msg* and *VMail_Plan* show the churn rate constantly decreasing. The curve is increasing at increasing rate up to point of inflections and beyond these points, churns behavior increases at a lesser rate. It is clear from the following ROC graph (see Fig. 5) that the "*Day _Mins*" variable alone is the best performing one as its curve dominates all the others while "*Intl_calls*" performs pretty badly alone. This ultimately implies that the churn rate is higher on those features which are above the curve except *Intl_Calls, VMail_Messages*, and *VMail_Plan*. On the basis of above discussion, it is observed that attributes (i.e. *Intl_Plan, Day_Mins, Day_Charges, Eve_Mins, Eve_Charges, Intl_Charges*, and *CustServ_Call*) are more important for CCP in the telecommunication industry.

### 6.4. Approximation analysis

We have also analyzed the vagueness in the dataset by using upper approximation and lower approximation of RST described in Table 9. Based on the results of approximation analysis, it is observed that there is no vagueness in the samples.

### 6.5. Comparison and discussion of simulation results

In this study, the proposed approach evaluated four well-known rules-generation algorithms (discussed in subsection 3.5) based on RST, to mine concise rules for efficiently predicting customer churns. The total number of rules induced from the telecom sector by applying rules-generation algorithms are summarized in Table 6. The LEM2 algorithm is seen to perform well on symbolic attributes only [75]. Fig. 6 reflects the graphical depiction of the induced concise rules.

It is observed that by applying both CA and LA in the proposed approach for rule generation, tend to produce a lesser number of rules than GA and EA. Specifically, the CA and LA both produced fewer rules (i.e., 369, and 625 respectively), as opposed to the total number of decision rules (4184 and 9468) generated by EA and GA respectively.

In this empirical study, the GA produced the maximum number of decision rules as compared to the other rules generation algorithms. The CA and LA are also faster on average, than EA and GA due to searching in few rules for predicting the customer's churns. Although, out of these four rules-generation algorithms, CCP using the CA yields maximum accuracy, however, both the CA and LA return very less valuable and meaningful rules that have direct effects on the performances of the both algorithms. Further, they also produced very low coverage as compared to the EA and GA (see Table 7). On the other hand, Nandita & Jaya [76] concluded from their empirical results that the CA yields the best accuracy compared to both the GA and LA. However, they were only focused on accuracy, without considering the coverage evaluation measure. In our case, the CA and LA covered 64% and 66% of customers respectively and the CA ignored 36% customers with 87% accuracy, whereas the LA ignored 34% customers with 99.3% accuracy. Therefore, we cannot recommend the CA and LA to use in such a scenario, where the coverage can be compromised in the favor of accuracy. On the other hand, EA checks all the possible combinations in search for the absolute optimum solution. Therefore, EA may turn out to be unacceptably slow since the required time is proportional to the maximum number of all possible solutions [77]. It was also observed that the GA and EA performed best as compared to the CA and LA, because both the GA and EA covered all the objects, whilst offering reasonable performance, as shown in Table 7. However, although the GA and EA demonstrated the best performance, in overall terms, the GA performed much better than EA, CA, and LA in recognition of true churns (precision) and true non-churns (recall). To show the balance between the two measures (i.e., precision and recall), we observed the performance of these rules-generation algorithms based on f-measure which combines the precision and recall. The GA obtained the best score i.e., 92.5% f-measure as compared to all other rules-generation algorithms.

In the proposed study, EA could not achieve optimal results as compared to GA, since the EA with RST was found to be relatively ineffective in reducing large data in the decision table (information system) for producing effective decision rules. On the other hand, GA begins by evaluating the problem as a population of solution candidates and generates new offspring through a cross-over and mutation process, with the aim of having the best-fit candidate when the next step for evaluation starts and so on [78]. It is concluded that a classifier trained on a large decision table using RST classification, coupled with the GA for inducing efficient decision rules, can produce the optimal solution for CCPin the telecommunication sector, subject to optimizing parameter settings (see Table 5). More importantly, the proposed approach is also flexible in evolving itself to a new situation following a

new generation, since background knowledge is not really required about the dataset [79].

Finally, it is clear from the comparative results, that the proposed approach performed very well as compared to previously applied techniques on the same benchmark dataset from the telecommunication industry (see Table 10).

### 7. Conclusion and future work

In this study, the application of RST is explored to predict customer churn in the telecommunication sector by constructing a predictive classifier that can forecast churn behavior based on accumulated knowledge. To evaluate the results of the proposed approach, a benchmarking study is applied and the performance of four different rules-generation algorithms (EA, GA, CA and LA) is investigated. It is found that RST classification based on GA outperforms other rules-generation algorithms in terms of precision, recall, the rate of misclassification, lift, coverage, accuracy, and F-measure. The discretization process applied to different attributes revealed some important insights regarding reasons for customer churn that can ultimately help the decision makers to develop retention policies accordingly. The proposed approach also outperforms other techniques applied on the same dataset in terms of precision, recall, accuracy, F-measure and rate of misclassification. It is important to note that the study conducted is pertaining to the specific dataset used for this study; results may vary with other datasets.

This study has shed some light on the performance of popular ML techniques for the CCP issues, and supported the advantage of RST application in the proposed approach. In future, we intend to further theoretically and experimentally investigate the proposed approach while considering several other pertinent issues. Firstly, churn datasets exhibit the class imbalance problem; whereby, the churn class (minority class or class of interest) contains fewer number of samples as compared to the non-churn class (majority class). This makes it difficult to recognize the minority class for some ML techniques; although they may achieve high overall accuracy. Secondly, eliminating and detecting of outliers would greatly contribute to providing better results. Finally, in this study the profiles of predicted customer churns were not considered; while these might be of interest to organizations in decision making related to retention of specific churn customers or letting them go. Thus worthy churn customers, at focus, could possibly have greater lifetime value. Hence, we wish to address these challenging issues in future research.

### References

[1] J. Hadden, A. Tiwari, R. Roy, D. Ruta, Computer assisted customer churn management: state-of-the-art and future trends, Comput. Oper. Res. 34 (10) (2007) 2902–2917.

[2] A. Sharma, P. Prabin Kumar, A neural network based approach for Predicting Customer churn in cellular network services, Int. J. Comput. Appl. 27 (11) (2011) 26–31.

[3] W. Verbeke, D. Martens, C. Mues, B. Baesens, Building comprehensible customer churn prediction models with advanced rule induction techniques, Expert Syst. Appl. 38 (3) (2011) 2354–2364.

[4] C. Kirui, L. Hong, W. Cheruiyot, H. Kirui, Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining, IJCSI Int. J. Comput. Sci. Issues 10 (2) (2013) 165–172.

[5] B. Huang, M.T. Kechadi, B. Buckley, Customer churn prediction in telecommunications, Expert Syst. Appl. 39 (1) (2012) 1414–1425.

[6] C.-S. Lin, G.-H. Tzeng, Y.-C. Chin, Combined rough set theory and flow network graph to predict customer churn in credit card accounts, Expert Syst. Appl. 38 (1)

(2011) 8–15.

[7] R.H. Wolniewicz, R. Dodier, Predicting customer behavior in telecommunications, IEEE Intell. Syst. 19 (2) (2004) 50–58.

[8] A. Idris, A. Khan, Y. Soo, Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification, Springer Science Business Media, New York, 2013, pp. 659–672.

[9] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (5) (1982) 341–356.

[10] S.-Y. Hung, D.C. Yen, H.-Y. Wang, Applying data mining to telecom churn management, Expert Syst. Appl. 31 (3) (2006) 515–524.

[11] C.-P. Wei, I.-T. Chiu, Turning telecommunications call details to churn prediction: a data mining approach, Expert Syst. Appl. 23 (2) (2002) 103–112.

[12] V.V. Saradhi, G.K. Palshikar, Employee churn prediction (Mar.)Expert Syst. Appl. 38 (3) (2011) 1999–2006 (Mar.).

[13] V. Lazarov, M. Capota, churn prediction, Bus. Anal. Course TUM Comput. Sci. (2007).

[14] L.Backstrom, D.Huttenlocher, J.Kleinberg, X.Lan, Group formation in large social networks, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06pp. 44–54, 2006.

[15] L.Xi, Y.Wenjing, L.An, N.Haiying, H.Lixian, Q.Luo, C.Yan, Churn Analysis of Online Social Network Users Using Data Mining Techniques, Preced. International multi Conference Eng. Comput. Sci., no. 1, pp. 14–16, 2012.

[16] W. Verbeke, D. Martens, B. Baesens, Social network analysis for customer churn prediction, Appl. Soft Comput. 14 (2014) 431–446.

[17] D.Archambault, N.Hurley, C.T.Tu, ChurnVis: Visualizing mobile telecommunications churn on a social network with attributes,Adv. Soc. Networks Anal. Min. (ASONAM), 2013 IEEE/ACM, pp. 894–901, 2013.

[18] K.Dasgupta, R.Singh, B.Viswanathan, D.Chakraborty, S.Mukherjea, A.A.Nanavati, A.Joshi, Social ties and their relevance to churn in mobile telecom networks, in: Proceedings of the 11th international conference on Extending database technology Advances in database technology - EDBT '08, pp. 668–677, 2008.

[19] J. David Nunez-Gonzalez, M. Grana, B. Apolloni, Reputation features for trust prediction in social networks, Neurocomputing 166 (2014) 1–7.

[20] U. Prasad Devi, S. Madhavi, Prediction Of churn behavior Of Bank customers, Bus. Intell. J. 5 (1) (2012) 96–101.

[21] K.Chitra, B.Subashini, Customer Retention in Banking Sector using Predictive Data Mining Technique, ICIT 2011 5th International Conference Inf. Technol., 2011.

[22] J. Bloemer, K. de Ruyter, P. Peeters, Investigating drivers of bank loyalty: the complex relationship between image, service quality and satisfaction, Int. J. Bank Mark. no. 16 (1998) 276–286.

[23] N. Nguyen, G. LeBlanc, The mediating role of corporate image on customers' retention decisions: an investigation in financial services, Int. J. Bank Mark. 16 (2) (1998) 52–65.

[24] A. Zakaryazad, E. Duman, A profit-driven artificial neural network (ANN) with applications to fraud detection and direct marketing, Neurocomputing 175 (2016) 121–131.

[25] K.C. Lee, N. Chung, K. Shin, An artificial intelligence-based data mining approach to extracting strategies for reducing the churning rate in credit card industry, J. Intell. Inf. Syst. 8 (2) (2002) 15–35.

[26] D. Van den Poel, B. Larivière, Customer attrition analysis for financial services using proportional hazard models, Eur. J. Oper. Res. 157 (1) (2004) 196–217.

[27] J.Kawale, A.Pal, J.Srivastava, Churn Prediction in MMORPGs: A Social Influence Based Approach, in: Proceedings 2009 International Conference Comput. Sci. Eng., vol. 4, pp. 423–428, 2009.

[28] M.Suznjevic, I.Stupar, M.Matijasevic, MMORPG Player Behavior Model based on Player Action Categories, in: Proceedings 10th Annu. Work. Netw. Syst. Support Games. IEEE Press, 2011.

[29] K.Chen, C.Lei, Network game design: Hints and implications of player interaction, in: Proceedings 5th ACM SIGCOMM Work. Netw. Syst. Support games., pp. 1–9, 2006.

[30] S. Meaghan, B. Nick, Voluntary turnover: knowledge management, J. Intellect. Cap. 3 (3) (2002) 303–322.

[31] M.Kane, Laura, Predictive Models of Employee Voluntary Turnover, 2007.

[32] J. Burez, D. Van den Poel, CRM at a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services, Expert Syst. Appl. 32 (2) (2007) 277–288.

[33] K. Coussement, D. Van den Poel, Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques, Expert Syst. Appl. 34 (1) (2008) 313–327.

[34] G.Dror, D.Pelleg, O.Rokhlenko, I.Szpektor, Churn prediction in new users of Yahoo! answers, in: Proceedings of the 21st International Conference companion World Wide Web - WWW '12 Companion, pp. 829–834, 2012.

[35] R.A. Soeini, K.V. Rodpysh, Applying data mining to insurance Customer churn management, Int. Proc. Comput. Sci. Inf. Technol. 30 (2012) 82–92.

[36] C. Rygielski, J.-C. Wang, C.D. Yen, Data mining techniques for customer relationship management, Technol. Soc. 24 (4) (2002) 483–502.

[37] S.M. Keaveney, Customer switching behavior in service industries: an exploratory study, J. Mark. 59 (2) (1995) 71–82.

[38] B. Padmanabhan, A. Hevner, C. Michael, S. Crystal, From information to operations: service quality and customer retention, ACM Trans. Manag. Inf. Syst. 2 (4) (2011).

[39] I. Brandusoiu, G. Toderean, Churn prediction in the telecommunications sector using support vector machines, Ann. ORADEA Univ. Fascicle Manag. Technol. Eng. (1) (2013).

[40] C.Kang, S.Pei-ji, Customer Churn Prediction Based on SVM-RFE, in 2008 International Seminar on Business and Information Managementvol. 1, pp. 306–309, 2008.

[41] E. Shaaban, Y. Helmy, A. Khedr, M. Nasr, A proposed churn prediction model, Int. J. Eng. Res. Appl 2 (4) (2012) 693–697.

[42] H. Kaizhu, Y. Haiqin, M.R. Lyu, Machine Learning: modeling data locally and globallyAdvanced Topics in Science and Technology in China, Springer-Verlag Berlin Heidelberg, 2008.

[43] H. Kaizhu, D. Zheng, J. Sun, Y. Hotta, K. Fujimoto, S. Naoi, Sparse learning for support vector classification, Pattern Recognit. Lett. 31 (13) (2010) 1944–1951.

[44] S.A.Qureshi, A.S.Rehman, A.M.Qamar, A.Kamal, A.Rehman, Telecommunication subscribers' churn prediction model using machine learning, in: Proceedings of the Eighth International Conference on Digital Information Management (ICDIM 2013)pp. 131–136, 2013.

[45] Z. Khawar Malik, A. Hussain, Q.M.J. Wu, Multi-layered echo state machine: a novel architecture and algorithm, IEEE Trans. Cybern. (2016).

[46] T. Vafeiadis, K.I. Diamantaras, G. Sarigiannidis, K.C. Chatzisavvas, A comparison of machine learning techniques for customer churn prediction, Simul. Model. Pract. Theory 55 (2015) 1–9.

[47] H. Abbasimehr, A neuro-fuzzy classifier for Customer churn prediction, Int. J. Comput. Appl 19 (8) (2011) 35–41.

[48] M.A.H. Farquad, V. Ravi, S.B. Raju, Churn prediction using comprehensible support vector machine: an analytical CRM application, Appl. Soft Comput. 19 (2014) 31–40.

[49] M.C. Mozer, R. Wolniewicz, D.B. Grimes, E. Johnson, H. Kaushansky, Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry, IEEE Trans. Neural Netw. 11 (3) (2000) 690–696.

[50] Z. Pawlak, Rough Sets And Data Mining, Kluwer Acad. Publ., 1997, p. 6.

[51] W. Au, K.C.C. Chan, X. Yao, A novel Evolutionary data mining algorithm With applications to churn prediction, IEEE Trans. Evol. Comput. 7 (6) (2003) 532–545.

[52] J. Burez, D. Van den Poel, Handling class imbalance in customer churn prediction, Expert Syst. Appl. 36 (3) (2009) 4626–4636.

[53] R.J. Jadhav, U.T. Pawar, Churn prediction in Telecommunication using data mining technology, Int. J. Adv. Comput. Sci. Appl. 2 (2) (2011) 17–19.

[54] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens, New insights into churn prediction in the telecommunication sector: a profit driven data mining approach, Eur. J. Oper. Res. 218 (1) (2012) 211–229.

[55] K. Kim, C.-H. Jun, J. Lee, Improved churn prediction in telecommunication industry by analyzing a large network, Expert Syst. Appl. 41 (15) (2014) 6575–6584.

[56] G. Olle, A hybrid churn prediction model in mobile Telecommunication industry, Int. J. e-Educ. e-Bus. e-Manag. e-Learn. 4 (1) (2014) 55–62.

[57] A.Amin, C.Khan, I.Ali, S.Anwar, Customer Churn Prediction in Telecommunication Industry: With and without Counter-Example, in: Proceedings of the 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Springerpp. 206–218, 2014.

[58] Z. Pawlak, rough set, rough relations and rough functions, Fundam. Inform. IOS Press 27 (2–3) (1996) 103–108.

[59] Z. Pawlak, Theoretical Aspects of Reasoning About Data, Kluwer Acad. Publ. Dordr., 1991.

[60] J.G. Bazan, S. Marcin, The rough set exploration system, Trans. Rough. Sets III, Springer Berlin Heidelb. (2005) 37–56.

[61] Z. Pawlak, A. Skowron, Rough Sets and ConflictAnalysis 37, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[62] J. Łukasiewicz, Die logischen Grundlagen der Wahrscheinlichkeitsrechnung, Akademie der Wissenschaften, Krakau, 1913.

[63] S.H.Nguyen, H.S.Nguyen, Analysis of STULONG Data by Rough Set Exploration System ( RSES ),Proceedings ECML/PKDD Work., pp. 71–82, 2003.

[64] J.G. Bazan, H.S. Nguyen, S.H. Nguyen, P. Synak, Jakub Wroblews, "Rough set algorithms in classification problem, Rough. Set. Methods Appl. Phys. HD (2000) 49–88.

[65] J. Wróblewski, Genetic algorithms in decomposition and classification problems, Rough. Sets Knowl. Discov. 2 19 (1998) 471–487.

[66] J.W. Grzymala-Busse, A. New Version, of the Rule induction System LERS,", Fundam. Inform. IOS Press 31 (1) (1997) 27–39.

[67] J.Grzymala-Busse, A system for learning from examples based on rough sets. In Intelligent Decision Support. In Intelligent Decision Support. Handbook of Applications and Advances of theRough Sets Theory, ed. byR. Slowinski, 1992.

[68] O. Vandecruys, D. Martens, B. Baesens, C. Mues, M. De Backer, R. Haesen, Mining software repositories for comprehensible software fault prediction models, J. Syst. Softw. 81 (5) (2008) 823–839.

[69] Z. Khawar Malik, A. Hussain, W. Jonathan, An online generalized eigenvalue version of Laplacian Eigenmaps for visual big data, Neurocomputing 173 (2016) 127–136.

[70] M.B. Stojanović, M.M. Božić, M.M. Stanković, Z.P. Stajić, A methodology for training set instance selection using mutual information in time series prediction, Neurocomputing 141 (2014) 236–245.

[71] G.Holmes, A.Donkin, I.H.Witten, WEKA: a machine learning workbench, in: Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conferencepp. 357–361, 1994.

[72] F.He, X.Wang, B.Liu, Attack Detection by Rough Set Theory in Recommendation System, in 2010 IEEE International Conference on Granular Computing, pp. 692–695, 2010.

[73] R. Bellazzi, B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines, Int. J. Med. Inform. 77 (2) (2008) 81–97.

[74] J. Hadden, A Customer Profiling Methodology for Churn Prediction (P.hD thesis), Cranf. Univ., 2008.

[75] J.W.Grzymala-busse, A Comparison of Three Strategies to Rule Induction from Data with Numerical Attributes.pdf, in: Proceedings of the International Workshop on Rough Sets in Knowledge Discovery (RSKD 2003)pp. 132–140, 2003.

[76] N. Sengupta, J. Sil, A.I. System, Comparison of different Rule calculation method for rough set theory, Int. J. Inf. Electron. Eng. 2 (3) (2012) 464–466.

[77] J. Nievergelt, Exhaustive search, combinatorial optimization and enumeration: exploring the potential of raw computing power, Sofsem 2000 Theory Pract. Inform. 1963 (2000) 18–35.

[78] N. Ariffin, M. Zin, S. Norul, H. Sheikh, N. Faridatul, A. Zainal, A Comparison of Exhaustive, heuristic and genetic algorithm for travelling Salesman problem in PROLOG, Int. J. Adv. Sci. Eng. Inf. Technol. 2 (2012) (2012) 49–53.

[79] C.Hor, P.A.Crossley, D.L.Millar, Application of Genetic Algorithm and Rough Set

Theory for Knowledge Extraction, in: Power Tech, 2007 IEEE Lausannepp. 1117–1122, 2007.

[80] A. Amin, S. Shehzad, C. Khan, I. Ali, S. Anwar, Churn prediction in telecommunication industry using rough set approach. New Trends in Computational Collective Intelligence, Springer, 2015, pp. 83–95.

**Adnan Amin** received the MSc degree in Computer Science from University of Peshawar, MS degree (with Distinction) in Computer Science with major Databases from Institute of Management Sciences Peshawar, Pakistan in 2008 and 2015 respectively. He is currently a Ph.D Scholar and Lecturer at Department of Computing Sciences at Institute of Management Sciences Peshawar. His research interests include Data mining, Databases, Big Data and Machine Learning.

**Sajid Anwar** obtained his BSc (Comp. Sc) and MSc (Comp. Sc) degrees from University of Peshawar in 1997 and 1999 respectively. He obtained his MS (Comp. Sc) and PhD (in Software Architecture) from the University of NUCES-FAST, Pakistan, in 2007 and 2011 respectively. He is currently Assistant Professor of Computing Science, and coordinator of the BS-Software Engineering at the Institute of Management Sciences Peshawar, Pakistan. His research interests are concerned with Software Architecture, Software Requirement Engineering, Searched Based Software Engineering and Mining Software Repository.

**Awais adnan** is Assistant Professor and Cooerdinator of Master Program, Department of Comptuer Science in Institute of Management Sciences Peshawar. He has done his Ph.D. from IMSciences|Peshawar and MS from NUST Islamabad. He is manager of ORIC in IMSciences| Peshawar to promote and facilitate the research students in commercialization of their research. His major areas of interest are Multimedia and Machine Learning.

**Muhammad Nawaz** received his MSc (Computer Science) and MS in Information Technology from University of Peshawar-Pakistan. He worked as a lecturer at the University of Peshawar; followed by working as a Computer Programmer at Khyber Teaching Hospital, Peshawar; and then appointed as Assistant Professor in Multimedia at the Institute of Management Sciences, Peshawar - a position he still holds. Currently he is the Head of PhD and MS-Computer Sciences at the IMSciences | Peshawar.

**Khalid Alawfi** received the B.Sc degree in computer engineering in 1999 from King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, and MSc and PhD degrees in Informatics from Bradford University, UK, in 2002 and in 2006 respectively. During 2002–2006, he worked as part of the Networks and Performance Engineering Research Group at Bradford University. Currently he is an Associate Professor in Computer Science, and Dean of the College of Computer Science and Engineering at Taibah University in Saudi Arabia. He is also a Senior Honorary Fellow at the Cognitive Big Data Informatics (CogBID) Research Laboratory at the University of Stirling, Scotland, UK.

**Kaizhu Huang** is currently an Associate Professor in Xi'an Jiaotong-Liverpool University, China. Before that, he was an Associate Professor at National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA). He was a student of the Special Class for Gifted Youth at Xi'an Jiaotong University and received the B.Sc. degree in Engineering in 1997. He received the M.Sc. degree in Engineering from CASIA in July 2000 and the Ph.D. degree from The Chinese Univ. of Hong Kong (CUHK) in 2004. He worked as a research scientist in Fujitsu R & D Centre from 2004 to 2007. During 2008 and 2009, he was a research fellow in CUHK and a researcher at University of Bristol, UK. He is the recipient of 2011 Asian Pacific Neural Network Assembly (APNNA) Distinguished Younger Researcher Award. He also received Best Book Award in National "Three 100"Competition 2009. He has published 6 books in Springer and over 110 international research papers (40 SCI-indexed international journals and 60+ EI conference papers) e.g., in journals (JMLR, Neural Computation, IEEE T-PAMI, IEEE T-NN, IEEE T-BME, IEEE T-SMC, NN) and conferences (NIPS, IJCAI, SIGIR, UAI,CIKM, ICDM, ICML,ECML, CVPR). He serves as Advisory Board Member in Springer Book Series Bio- Neuroinformatics. He is the member of CCF Technical Committee of Artificial Intelligence and Pattern Recognition. He served on the programme committees in many international conferences such as ICONIP, IJCNN, IWACI, EANN, KDIR. Especially, he serves as chairs in several major conferences or workshops, e.g., AAAI 2016 (Area Chair), ACML 2016 (Publication co-Chair), ICONIP 2014 (Program co-Chair), DMC 2012–2016 (Organizing co-Chair), ICDAR 2011 (Publication Chair), ACPR 2011 (Publicity Chair), ICONIP2006, 2009–2011 (Session Chair).

**Amir Hussain** received the BEng (with highest first class Hons.) and the Ph.D. degree in novel neural network architectures and algorithms from the University of Strathclyde, Glasgow, U.K., in 1992 and 1997, respectively. He is currently full Professor of Computing Science, and founding Director of the Cognitive Big Data Informatics (CogBID) Research Laboratory at the University of Stirling in Scotland, UK. He has conducted and led collaborative research with industry; partnered in major European and international research programs, and supervised over 30 Ph.D. students. He has (co)authored over 300 papers, including over a dozen books and 100+ journal papers. He is founding Editor-in-Chief of the journals: Cognitive Computation (Springer Nature), and Big Data Analytics (BioMed Central), and Chief-Editor of the Springer Book Series on Socio-Affective Computing, and Cognitive Computation Trends. He is Associate Editor of the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Transactions on Systems, Man, and Cybernetics: Systems, and the IEEE Computational Intelligence Magazine. He is Chapter Chair of the IEEE UK & RI Industry Applications Society, Vice-Chair of the Emerging Technologies Technical Committee of the IEEE Computational Intelligence Soceity (CIS), and founding General Chair for the IEEE CIS sponsored, flagship IEEE SSCI (CICARE Symposium) series. He is a senior Fellow of the Brain Science Foundation (USA).