CHAPTER 3

METHODOLOGY

The dramatic rise in the use of mobile phones and the improvement in communications in general, in the recent years have transformed telecommunication industry as one of the strongest sector. The 21st century is envisaging intense competition in telecommunication industry and as a consequence, majority of the companies are paying more attention on improving customer relationship.

The successful application of data mining in highly visible fields like e-business (Astudillo et al., 2014), marketing (Ramanathan et al., 2012) and retail (Doherty and Fiona, 2006) have led to the popularity of its use in Knowledge Discovery in Databases (KDD) in other industries and sectors. Among these sectors that are just discovering data mining are the fields of telecommunication. Advances in communication technology have enabled telecom industries to automatically collect huge amount of customer data.

Recent advances in software and technological breakthroughs in hardware had made the storage and accessing of huge amount of data economical. It is now possible to operate on large datasets in a reasonable time to perform exhaustive searches and find brute force solutions. In spite of these advanced techniques, the task of analyzing these huge databases is still a big challenge, especially in the field of mobile communication. Knowledge discovery using data mining techniques is a perfect solution to these situations.

The main challenge in the area related to the research work is to find techniques that bridge the two fields, data mining and telecommunication, for an efficient and successful knowledge discovery. The eventual goal of this data mining effort is to identify factors that will improve the quality and cost effectiveness of customer care (Yu and Ying, 2009).
Usage of data mining techniques for customer relationship management is becoming increasingly popular due to several reasons. One important factor is that the huge amount of data generated by customer transactions are too complex and voluminous to be processed and analyzed by traditional methods. The data consists of customer details, billing details, services (offered and used) and equipment details. This large amount of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-saving and improve decision-making by discovering patterns and trends.

Yet another factor motivating the use of data mining applications in CLA-AKD is the realization that data mining can generate information that is very useful to all parties involved in the telecommunication industry. For example, data mining applications can help the business management to identify dissatisfied customers. Service providers can gain assistance in making decisions, for example, in customer relationship management (Koh and Tan, 2010).

Integration of decision support with computer-based CRM can reduce errors, enhance customer care, decrease customer churning and improve customer loyalty (Wu et al., 2002). This integration is promising as data modeling and analysis tools, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of customer relationship.

The main goal of using data mining techniques in CLA-AKD is to identify those customers who share common attributes, which can be interpreted to group them as either loyal or churners. This output, if accurately discovered, can be utilized to improve customer relationship with the telecommunication industry and in long run the revenue of the company. In this research work, two types of operations are performed on customer database. They are,
(i) Customer Loyalty Assessment: To recognize and classify customers with multivariate attributes as loyal customers (non-churners) and churners.

(ii) Actionable Knowledge Discovery – To predict a set of actions that can be used to convert churners to valuable loyal customers

Both operations can be efficiently performed using cluster analysis and classification. Cluster analysis aims at organizing a given dataset into groups (clusters) of similar objects or characteristics and classification aim at predicting the class of objects whose class label is unknown.

In order to meet the objectives formulated in Chapter 1, Introduction, this research work uses data mining techniques extensively. The steps of the proposed CLA-AKD model are shown in Figure 3.1 and the research methodology is presented in Figure 3.2.

Figure 3.1: Architecture of CLA-AKD Model
The proposed CLA-AKD model performs customer loyalty assessment and actionable knowledge discovery in three steps, namely,

(i) Preprocessing

(ii) Customer Loyalty Assessment

(iii) Actionable Knowledge Discovery.
The result of these three steps can be used by companies as a profit-based objective function. Each of the above three steps are treated as a separate phase in the study and are interrelated to each other and are discussed in the following sub-sections.

3.1. PHASE I: DATA CLEANING OPERATIONS

The dataset used for knowledge discovery, in general, are collected or created from existing warehouses, databases and data marts. Data cleaning operations describe the different processing operations performed on raw data collected to prepare it for another processing procedure. Regardless of how powerful data mining method used is, the resulting CRM model will not be valid if the data is not preprocessed correctly and efficiently. Data cleaning describes any type of processing performed on raw data to prepare it for another processing procedure. It transforms the data into a format that will be more easily and effectively processed.

Data cleaning operations has become important in CLA-AKD because of the low quality of the data acquired. In most of the cases, the data collection is performed by third parties and may not be in a format that is required by the application. The collected data might have data errors or outliers, data with gaps or missing values and duplicates, all of which degrade the performance of data mining process. Data cleaning routines attempt to smooth out noise while identifying outliers, fill-in missing values and correct inconsistencies in data. The amount of time and cost spend on preprocessing has a direct impact on the accuracy (Figure 3.3).
Data preprocessing is a task that should be given utmost attention for the following two reasons:

- Quality decisions are based on quality data - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
- Data preparation, cleaning, and transformation comprise the majority of the work in a data mining application (90%).

It is a well-known fact that “improving data quality can improve the quality of any analysis on it”. Analyzing data that has not been carefully screened often produce misleading results. Therefore, using data cleaning routines that improve the representation and quality of data is an important task that should be performed before running an analysis (Kotsiantis et al., 2006). If the amount of irrelevant and redundant information or amount of noisy and unreliable data is high, then knowledge discovery becomes more difficult. Based on this, many CRM applications perform data cleaning operations as a mandatory step.

The data cleaning operations use routines that handle incomplete, noisy and inconsistent data, which may slow down the system. Noisy data contain errors or outliers. Outliers are data objects that deviate heavily from the normal characteristics of a dataset (Example: Bill Amount = -1003). Incomplete data can be missing attribute values, lack of certain attributes of interest, or can be data containing only aggregate data (Example: Customer Address="").
Inconsistent data those which contain discrepancies in attributes like codes or names (Example : Age="42" Birthday = “03/07/2011”).

To solve the problems of noisy data and incomplete customer dataset, the first phase of the study focuses on two data cleaning operations. The first operation removes data records that are not important for analysis, while the second operation implements a missing handling procedure to fill in missing data items or records in the customer dataset.

3.1.1. Outlier Detection and Removal

Outlier detection and removal is a process that is considered very challenging and important by all data mining applications. The first task of the first phase concentrates on detecting noise or outliers in the telecom dataset and removes them to obtain a clean dataset. For this purpose, an enhanced density based outlier detection algorithm using Local Outlier Factor (LOF) is proposed.

The conventional algorithm using Local Outlier Factor (LOF) (Breunig et al., 2000) compares the local density of a point’s neighborhood with the local density of its neighbors. This operation results with high computations and high memory requirements. The study proposes techniques, which incorporates optimizers that reduce high computations and high memory requirements. The high computation problem is solved by reducing the number of distance calculations, while the memory requirement is reduced by using a data partition method.

The enhancement operation on LOF algorithm to reduce computations is based on the fact that most of the data are impossible outliers. It is local in that the degree depends on how isolated the object is with respect to the surrounding neighborhood. The speed optimization is brought by eliminating the check made to determine whether there are core objects in the neighbourhood. All points of a core object along with its reachable neighbours are considered as inliers and distance calculation and comparision is avoided.
for all these points. Thus, inclusion of this simple check procedure reduces the computational cost considerably.

Data partitioning is an effective technology to solve the problem of memory and huge computations in LOF. The proposed algorithm considers the entire cleaned dataset, from which the statistical data distribution characteristics projected on each dimensionality respectively is obtained. The next step determines and chooses partition points in such a manner that it enhances the uniformity of the data density distribution. Statistical analysis of data distribution is performed using histogram method for this purpose. Next, the data is divided into several subsets using the partition points obtained from the previous step. Finally, the enhanced LOF algorithm is applied on each partition to identify and remove outliers.

3.1.2. Missing Value Handling

A missing data is defined as an attribute or feature in a dataset which has no associated data value. Correct treatment of these data is crucial, as they have a negative impact on the interpretation and result of data mining processes. Incomplete data is an unavoidable problem in dealing with most of the real world data sources (Karmaker and Rahman, 2009). In the present research work, while considering the state attribute of customer address, a method based on recursive matching is proposed.

While considering filling missing values in attributes related to customer behavioural and billing, initially, a check is made to ascertain whether all the values are NULL for a particular record. In this scenario, the entire record is considered as irrelevant information and is deleted. Otherwise, an enhanced KNN (K-Nearest Neighbour) imputation method is used.
3.2. PHASE II: CUSTOMER LOYALTY ASSESSMENT MODELS

Frameworks for predicting customer’s future behaviour has captured the attention of several researchers and academicians in recent years, as early detection of future customer churns is one of the most wanted CRM strategies. Future of company depends on the stable income provided by loyal customers. However, attracting new customers is a difficult and costly task involved with market research, advertisement and promotional expenses.

According to Keaveney (1995) and Sharma and Panigrahi (2011), these expenses are several times greater than the cost of efforts that might enable the firm to retain a customer. It has been established by Kim and Yoon (2004) that the best core marketing strategy for future is to retain loyal existing customers by avoiding customer churning.

However, the telecom services industry in developing countries is yet to standardize a set of customer profitability measurements and the second phase of the study focuses on developing a customer loyalty assessment model using data mining techniques. Most of the existing customer loyal assessment models focus on using any one of the data mining technique, namely, classification or clustering, to extract knowledge on churners and non-churners. Some proposals have also used a combination of clustering and classification algorithms (Yao et al., 2010).

This study, in continuation with combination analysis, proposes CLA system that uses both clustering and classification techniques for churn management. In CLA, instead of using clustering to improve classification process, the clustering is used to reduce the dataset to consist only churners, which are further analyzed using classification algorithm. Thus, the proposed model consists of two steps, where the first steps identifies churners and non-churners using clustering algorithm and the second step further prioritizes the churners according to three loyalty levels, namely, low churners, medium churners and high churners using classification algorithm.
3.2.1. Step 1 : Clustering

Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. An example of clustering is depicted in Figure 3.4. The input patterns are shown in Figure 3.4a and the desired clusters are shown in Figure 3.4b. Here, points belonging to the same cluster are given the same label. The variety of techniques for representing data, measuring proximity (similarity) between data elements and grouping data elements has produced a rich and often confusing assortment of clustering methods.

Figure 3.4 : Data Clustering

Clustering is useful in several exploratory pattern-analysis, grouping, decision-making and machine-learning situations, including data mining, document retrieval, image segmentation and pattern classification. However, in many such problems, there is little prior information (e.g., statistical models) available about the data and the decision-maker must make as few assumptions about the data as possible. It is under these restrictions that clustering methodology is particularly appropriate for the exploration of interrelationships.

57
among the data points to make an assessment (perhaps preliminary) of their structure.

Typical pattern clustering activity involves the following steps (Jain and Dubes 1988):

- Pattern representation (optionally including feature extraction and/or selection),
- Definition of a pattern proximity measure appropriate to the data domain,
- Clustering or grouping,
- Data abstraction (if needed), and
- Assessment of output (if needed).

Figure 3.5 depicts a typical sequencing of the first three of these steps, including a feedback path where the grouping process output could affect subsequent feature extraction and similarity computations.

![Figure 3.5: Stages in Clustering](image)
**Pattern representation** refers to the number of classes, the number of available patterns and the number, type and scale of the features available for the clustering algorithm. Some of this information may not be controllable by the practitioner. **Feature selection** is the process of identifying the most effective subset of the original features to use in clustering. **Feature extraction** is the use of one or more transformations of the input features to produce new salient features. Either or both of these techniques can be used to obtain an appropriate set of features to use in clustering.

**Pattern proximity** is usually measured by a distance function defined on pairs of patterns. A variety of distance measures are in use in the various communities. A simple distance measure like Euclidean distance can often be used to reflect dissimilarity between two patterns, whereas other similarity measures can be used to characterize the conceptual similarity between patterns (Serira et al., 2012).

The **grouping step** can be performed in a number of ways. The output clustering (or clustering) can be hard (a partition of the data into groups) or fuzzy (where each pattern has a variable degree of membership in each of the output clusters). Hierarchical clustering algorithms produce a nested series of partitions based on a criterion for merging or splitting clusters based on similarity. Partitional clustering algorithms identify the partition that optimizes (usually locally) a clustering criterion.

Additional techniques for the grouping operation include probabilistic (Brailovsky, 1991) and graph-theoretic (Pavan and Pelillo, 2003) clustering methods. In general, the grouping algorithms can be broadly classified into (Paquet, 2004), Partition clustering, Association based clustering, Hierarchical clustering, Spectral clustering, Density-based clustering and Grid-based clustering. All the algorithms depend on the distance measure between two objects and have the common goal to minimize the distance of every object from the center of the cluster to which the object belongs.
Phase II of the research work, to group customers into churners and non-churners, proposes a hybrid model that combines the advantages of SOM (Self Organizing Maps), KMeans and DBScan clustering algorithms.

3.2.2. Step 2 : Classification

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae or neural networks. Out of these the decision trees representation is more popular as they can be easily converted to classification rules.

Classification can be used for predicting the class label of data objects. However, in many applications, users may wish to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data. Although prediction may refer to both data value prediction and class label prediction, it is usually confined to data value prediction and thus is distinct from classification. Prediction also includes the identification of distribution trends based on the available data.

A classification technique, or a classifier, is a systematic approach to building classification models from an input data set. Examples include, Decision Tree Classifiers, Rule-Based Classifiers, Neural Networks, Support Vector Machines and Naïve Bayes Classifiers. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records it has never seen before. Therefore, a key objective of
the learning algorithm is to build models with good generalization capability, that is, models that accurately predict the class labels of previously unknown records. Figure 3.6 shows a general approach for solving classification problems.

![Figure 3.6: General Process of Classification](image)

First, a training set consisting of records whose class labels are known must be provided. The training set is used to build a classification model, which is subsequently applied to the test set, which consists of records with unknown class labels.

The research work, to determine the churning level (loyalty level) of the churners, proposes the use of three classification algorithms, namely, Support Vector Machine (SVM), Back Propagation Neural Networks (BPNN) and Decision Tree (DT) classifiers.

### 3.3. PHASE III: ACTIONABLE KNOWLEDGE DISCOVERY MODELS

Phase II of the study focuses on developing techniques that discovered patterns of data for predicting and prioritizing churners and non-churners. Phase II, on the other hand, focuses on Actionable Knowledge Discovery (AKD), which is a paradigm shift towards mining more usable and more applicable knowledge in their corresponding domains (Cao et al., 2010).
order to improve customer relationship, the enterprise must know what actions to take to change customers from an undesired status (such as churn) to a desired one (such as loyal customers). This can be done in the telecommunications industry, by reducing the monthly rates or increasing the service level for a valuable customer.

Actions, such as direct mailing and sales promotion, will cost money to the enterprise (Yang et al., 2003). At the same time, enterprises are increasingly constrained by cost cutting. There is thus a strong limitation on the number of customer segments that the company can take on, or in the number of actions the company can exploit. To make a decision, one must take into account the cost as well as the benefit of actions to the enterprise. However, for each customer, there may be a large number of possible actions or action sets that can be applied to the customer and which of the actions to take depends not only on the particular customers’ situation, but also on other customers who might benefit from the same action.

In the past, several approaches have been designed to extract knowledge from the CRM data. When the actions are initially unknown, however, few approaches have been designed to invent new actions that can minimize the total cost and bring about profitable changes. In this research, the AKD is performed by improving the model proposed by Yang et al. (2007), who implemented a post-processing procedure to obtain actions that are associated with attribute-value changes that are directed to maximize the profit-based objective functions.

The proposed AKD model consists of four steps after importing preprocessed customer data from Phase I of the study. The first step implements two dimensionality reduction algorithms, namely, ACO (Ant Colony Optimization) or UDA (Uncorrelated Discriminant Analysis) to reduce the dataset size. The second step builds customer profile from the training data, using two classifiers, a decision tree learning algorithm and
Bayesian Network classifier. These two algorithms were selected because of their efficiency provided during churn prediction. The third step searches for optimal actions for each customer. This is a key component of the system’s proactive solution. The fourth step produces reports for domain experts to review the actions and selectively deploy the actions.

3.4. EXPERIMENTAL RESULTS

Experiments were conducted to evaluate the performance of the proposed solutions to each data mining problem using a telecommunication customer dataset. The dataset consists of 1,00,000 records with 20 attributes. The attributes store data related to each customer, including their personal, behavioural and accounting details.

The objective of the experiments conducted was to study the effect of the proposed algorithms on loyalty assessment and actionable knowledge discovery. The method using evaluation of the proposed methods in each phase is given in Figure 3.7.

![Evaluation Method Used](image)

**Figure 3.7 : Evaluation Method Used**
The experiments were conducted in four stages. The first and second stage evaluated the performance of the missing value handling and outlier removal algorithms using Normalized Root Mean Square Error (NRMSE) and outlier detection rate as performance metrics. The customer loyalty assessment models were analyzed using accuracy and error rate performance metrics, while AKD was evaluated using the net profit gain obtained due to action discovery. The experimental results are discussed in Chapter 7, Results and Discussion.

3.5. CHAPTER SUMMARY

This chapter presented the research methodology and introduced the various enhanced techniques are used to design the proposed CLA-AKD system. The next chapter, Design of Preprocessing Algorithms, presents detailed description of the proposed missing value handling algorithm and outlier detection algorithm.