ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

DEVELOMPENT OF UNSUPERVISED TELECOM CUSTOMERS CLUSTERING MODEL USING CUSTOMER DETAIL RECORDS (CDR) THE CASE OF ETHIOTELECOM

Banchalem Abebaw

A Thesis Submitted to the Department of Computer Science in

Partial Fulfillment for the Degree of Master of Science in Computer Science

Addis Ababa, Ethiopia

April, 2021

Addis Ababa University

College of Natural Science

Banchalem Abebaw

Advisor: **Mesfin Kifle (PhD)**

This is to certify that the thesis prepared by Banchalem Abebaw, titled: ***Development of Unsupervised Telecom Customer Clustering Model Using Customer Detail records (CDR) The Case of Ethiotelecom*** and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the university and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

	Name	Signature	Date
Advisor:	Mesfin Kifle (PhD)		
Examiner:			
Examiner:			

Dedication

To My Mother

Abstract

Almost in every discipline, people are using phones that generate detail records containing information's about phone usage, such as, connected time of the call, the identities of sources, the identities of destinations, the duration of each call, the amount billed etc. named as (customer detail records) CDR. It's difficult for someone to scan through all the data and establish the relative decision for the business because it is time. As a result, there is a growing interest towards better solutions for finding, organizing and analyzing these CDR data in telecommunication companies. The effective ways of organizing telecom data form later decision making and business management efficient, less complicated, friendly and low-cost. Customer detail record (CDR) clustering is one of the common methods of managing customers in business using their usage behavior in network.

This study proposes a model that cluster telecommunication customers using detail records (CDR) of the user. During the clustering process, all the CDR records pass through pre-processing stages to prepare data for processing. Then transformation of preprocessed data was done by through scaling algorithm. The scaled features were extracted from the sampled CDR record. Finally, customers are clustered based on the usage behavior using the most popular K-means algorithm that is based on the cosine similarity of the weighted features. Ethio-telecom customer detail records were used for experimentations. The clustering results are evaluated to find optimal cluster size using elbow and silhouette methods. The result shows the value of silhouette coefficient is greater when cluster size is four, with this we clustered the sampled data into four classes. When we see the number of customer's distribution in each class from hundred thousand sampled data, in the first class 19.959 % of the sample was grouped. In the second class 0.603% of the sample was grouped together. In the 3rd class there are 4.383% of the sample with related usage characteristics. In the 4th class there are 75.054% customers with related usage characteristics. The four groups are considered as customers with low usage history of the service, common usage history of the service, customers with very good usage history and customers with very midst usage history. The ultimate goal of any business would be to have as many customers up there in the service category. This can be further used in decision making to have many numbers of customers to our needed group i.e., a group with very good usage behavior of services.

Keywords: ***CDR Clustering, Customer Data Analysis, Feature Scaling, K-means Clustering***

Acknowledgments

First of all, I would like to thank God and gratefully acknowledge the help, guidance and support of God in my whole life for giving me the wisdom, strength, support and knowledge in exploring things. Next, I would like to express my gratitude and deeply thankful to my advisor Mesfin Kifle (PhD), who was always there during the process of this thesis work for giving me support, encouragement and continuous advice.

Finally, I would like to thank all of my family members, friends for giving me support and encouragements and also ethiotelecom staffs.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

A	Accuracy
CDMA	Code-division multiple access
CDR	Call detail record
EM	Expectation Maximization
ETC	Ethiopian Telecommunication Corporation
DF	Inverted Term Frequency
IMSI	International mobile subscriber identity
IR	Information Retrieval
P	Precision
R	Recall
SQL	Structured Query Language

Chapter 1: Introduction

1.1 overview

Recent advancement in storage, networking, data processing, and related technologies have significantly eased the process of generating and collecting large data. Today the mobile phone especially a smart phone is a mobile phone with an operating system capable of doing task which are not feasible in simple mobile phone. It has provided us convenience to do task we normally would do on our computer like email, shopping and banking as well as providing the various way for entertaining like gaming, watching videos and movie and the social media which has brought people near to each other and increased the communication among the people [1].

Thus, data processing and analysis become important to assist and perceive the business. information analysis in telecommunication trade helps in characteristic the telecommunication patterns that embrace, analysis of telecommunication information, deceitful pattern analysis, identification of bizarre patterns, successive patterns analysis, create higher use of resource, and improve quality of mobile telecommunication services [2]. Understanding customer behavior better companies can provide customized services and products. As a consequence, segmentation has proven to be a valuable source of information for marketing, sales and business planning. Traditionally, companies conduct segmentation based on different market research and customer surveys. However, with the introduction of data mining methods, there are now ways to perform segmentation based on the actual behavior of the customers. Currently the corporation has a lot of daily stored data about customers. The stored data can be used for analysis that can be further used for decision support, planning and customer management system.

Because of increased number of industries, competition in the mobile telecommunications industry is becoming more increased. Mobile operators' profits are facing tremendous challenges. Customers' demands become diversified, differentiated and requirements of service quality become more rationalized and stricter. In order to improve mobile operators' competitiveness and customer values, several data analysis technologies and methods are used to help decision makings [4]. In supervised methods, a classifier is trained on a manually tagged corpus of data and classes are specified, but doing this in CDR data is so difficult. Unsupervised learning methods like clustering automatically split the dataset into groups based on their similarities. With a good categorization method, telecom data can be

organized into a meaningful category, which facilitates an efficient management of customer or efficient decision making by focusing on relevant categories rather than whole data.

Our main effort in unsupervised clustering model of telecom industry data was focused on the development of model that allows customer management and business intelligence better by categorizing the users based on network usage. This research will be conducted to explore the advantage of using data analysis for new service delivery in telecom customers. Unsupervised clustering method using mobile network usage behavior represented by attributes (voice duration, internet, SMS) was employed. Finally, an application of a mobile customer clustering analysis is discussed in this paper.

1.2 Motivation

Recently, mobile network users and customer detail records has been dramatically increasing. For example, in 2019, the number of Internet users in world was 5.11 billion [13]. 100 million people have started using smartphones in the past year. After one year, in 2020, the number of global smartphone users is projected to total 3.5 billion, marking a 9.3 percent increase from 2019. The current global population of 7.7 billion people means the smartphone penetration rate is at 45.4 percent.

Thus nowadays large collections of user data are found in service provider. Using a good clustering method, these user data can be organized into meaningful clusters (groups), which is used for easily understanding user needs or assuring quality of service. The motivation to work on unsupervised clustering model of telecom user's data arises from the need of accessing and processing the huge collections user data to satisfy the need of the customer, to build long term and profitable relationships with chosen customers and get closer to those customers at every point of contact and customer service delivery better by targeting a group of customers

1.3 Statement of the Problem

With the globalization of world economy and market internationalization, the market competition of the domestic mobile communications industry has become more reasonable and increasingly fierce. The fast-growing user group, diversified operations and competition environment have put forward higher requirements for the service quality of the mobile communications industry [5]. The competition to acquire and retain customers

among mobile service providers is fierce. Different customers have different preferences in using mobile telecommunication services. The key to survival in this competitive industry lies in knowing the customers better.

The mobile telecommunication marketplace is highly competitive. Increasing the number of customers is the main challenge in modern telecommunication industry [4]. In this paper, the author has showed that through the use of customer segmentation, a telecommunication company can easily attract its customers with right products and services. This also helps in offering packages, offers and bundles for customers.

A work on two-layer customer clustering model [5] provides ways to assist mobile Customer Relationship Management. Marketers can use pre-analysis and data mining to target their customers and sell the company's products and services. In this work customer clustering is only included in cluster modeling through mobile voice, data usage behavior, customer contributions, and customer base data and it doesn't include customer-feature selection function that could make clustering efficient and valuable. In practice, the definition of customer value varies by industry. Even the same industry can have different features, such as the amount of consumption (call duration, SMS usage, number of callings), and so on. Works on Searching customer patterns of mobile service using clustering and quantitative association rule [6] presents the patterns of use for added services to mobile telecommunication subscribers in chosen, correlational analysis, bunch and quantitative association rules were wont to notice the service adoption patterns of divided teams.

The aim of clustering is to categorize prospective customers into distinct groups for distinctive contact strategies and proximal offerings. This targeting practice has been proved manageable and effective for mobile telecommunications industry [4], [5], This research will be conducted to explore the advantage of using data analysis for new service delivery in telecom customers. Therefore, this thesis work was an initial attempt to explore the use of both clustering and quantitative association rules to extract useful group from the information provided by customer's usage behavior from CDR of telecommunication company. Moreover, based on all inferred feature associations, it is to investigate a category of customers that would adopt new services and functions. This study could also contribute telecom Companies with an approach based on customers' operational needs (including

voice, data, SMS, Data usage and other business) in ac-accordance with the unsupervised clustering model for business management.

1.4 Objectives

General objective

The goal of this research is to design unsupervised clustering model for telecom network users using call detail record (CDR) of customers.

Specific objective

To accomplish the above-mentioned general objective, the following are specific objectives:

- Identifying the process of categorizing a data.
- Collecting telecom CDR dataset.
- Identifying and defining the feature identification techniques.
- Identifying and defining features.
- Adopting the appropriate feature scaling algorithm
- Adopting the appropriate categorization algorithm.
- Designing a generic clustering model for telecom CDR data.
- Implementation and testing the performance the model.

1. 5 Methodology

In order to accomplish the general objective, the following system of principles, practices, and procedures were applied.

Literature Review: review of literature was conducted to understand various component of data categorization. Specifically, we reviewed literature in the area of data analysis for business intelligence.

Data collection: telecom network usage data was collected from the company. The data was organized and structured in a way that they are easy for experimentation and testing.

Experimentation: in order to accomplish the objectives of the research, we experimented using available tools (PyCHARM editor, Python) and programming was engaged in the process.

Testing and Evaluation: to evaluate the performance of the proposed solution, the system was tested using collected data from the company. The result was evaluated by applying selected evaluation techniques.

1.6 Scope and Limitation

This research was conducted to explore the advantage of using customer data record (CDR) for clustering customers. The scope of the study was to propose and develop an unsupervised clustering model. In this research work available CDR data from ethiotelecom was used for experimentation. In this study, we considered only openly available information's that contain usage characteristics of customer.

1.7 Application of Results

The result of this study could be used as an input for other researches, and possible application with the following area.

Supervised clustering model of telecom mobile network usage can be used:

- in customer data searching to improve efficiency and search results;
- for data filtering, pointing to specific processing mechanisms such as a group of customer data extraction and new activities on network identification;
- as an input for other data management tasks like organizing, structuring, processing, controlling, evaluation and reporting of information;
- for search optimization: Clustering helps a lot in improving the quality and efficiency of system as the user query can be first compared to the clusters instead of comparing it directly to the whole data in the store.
- for telecom organization and application developers which has a large CDR data to automatically cluster customers for better management and for recommendation of new services;
- For decision support, better business management and future planning.

In addition, this study will open a way for further researches in the area of using customer data stored daily on different data stores.

1.8 Thesis Organization

The remaining part of the thesis is organized as follows. Chapter Two covers literature review in which different concepts and approaches related to our thesis are presented. Moreover, CDR segmentation, clustering and classification, clustering methods, similarity measurement methods are described. Chapter Three is about works related to our study that are previously done by other researchers in different place using customer detail record for better decision making. Chapter Four deals with the design of our model i.e. unsupervised clustering using customer detail record. It presents the general architecture of the system with its basic components; the discussion of the components and their interaction within the system. The algorithms we developed for achieving the goal was presented. Chapter Five focuses on the detail testing and evaluation of the system. It discusses the details about the testing and the results obtained together with their explanations. Conclusions from the thesis result, the contributions of our thesis research work and recommendations on possible future works related to this work are described in the last chapter of the thesis

Chapter 2: Literature Review

2.1 Introduction

The organization of data based on the information into homogeneous groups plays a major role in many fields of research; and clustering is a widely studied grouping method in the different domain. The method finds numerous applications [13] in customer segmentation or division, classification, collaborative filtering, document organization and etc. In this chapter, extensive reviews of general concepts on customer data record in telecom Company, challenges in CDR data collection, clustering techniques, similarity measurement approaches and cluster evaluation techniques are presented. The customer data record features of the telecom company were also reviewed.

2.2 Mobile Cellular Network and Services

Cellular network or mobile network is a communication platform wherever the last link is wireless. The network is spread over land areas known as cells, each served by at least one transceiver at a fixed location, but more typically by three cell sites or base transceiver stations [14]. These base stations offer network coverage to the cell that may be used for voice,data, and alternative styles of content transmission. A cell generally uses a special set of frequencies from neighboring cells, to avoid interference and supply secure service quality among every cell. once joined along, these cells offer radio coverage over a good geographical region. this allows an oversized range of moveable transceivers (e.g., mobile phones, tablets and laptops equipped with mobile broadband modems, pagers, etc.) to communicate with one another and with fastened transceivers and telephones anyplace within the network, via base stations, albeit a number of the transceivers square measure moving through quite one cell throughout transmission.

Cellular networks offer a number of desirable features [14]:

- More limit than a solitary enormous transmitter, since a similar recurrence can be utilized for numerous connections as long as they are in various cells
- Mobile gadgets utilize less force than with a solitary transmitter or satellite since the cell towers are nearer
- Larger inclusion territory than a solitary earthly transmitter, since extra cell pinnacles can be added uncertainly and are not restricted by the horizon

For more than phone calls, cellular networks are constantly being used. The creation of a number of advanced cell phones or 'smartphones' and portable PCs, tablets and other mobile devices has resulted in better handsets and the increased data transmission speeds of the networks. Smartphones and other mobile devices have access to the internet for staff, mobile business software, wireless shopping, billing and online purchases, email accounts and other services. It allows them much greater flexibility - enabling them to continue to function remotely. 3G, 4G, or 5G network organizations offer quicker information move and permit the advancement of more handset-based ongoing business applications [15]. Despite the fact that GSM and GPRS information administrations could in any case be accessible for a long time, the quicker information rates accessible with 4G and 5G imply that they offer a more alluring assistance, and over the long run other will be replaced.

2.3 Call Detail Record

Call detail record (CDR) is an information record delivered by a phone trade or different media communications gear that record the subtleties of a phone call or different broadcast communications exchange (e.g., instant message) that goes through that office or gadget. CDR offers key metadata on how and when to utilize your business telephone framework [17]. Also, a call detail record (CDR) gives data about calls made over a telephone administration. A CDR can offer selling precise answers about where, when, furthermore, how calls are made for revealing and charging purposes. A CDR provides metadata (data about data) typically includes:

- the telephone number of the client beginning the call
- the telephone number accepting the call
- the call span and charging telephone number that is charged for the call
- a special number recognizing the record
- the course by which the call entered the trade
- the course by which the call left the trade
- Call type (voice, SMS, web, and so on)
- How long the call kept going and how much the call cost
- Date/time the call began and finished.
- The beginning and ending towers and so on the phone number of the customer originating the call etc.

Customer data record (CDR) can also include SMS messaging metadata and any other official communications transmission transactions. However, the actual contents of the messages/calls are not exposed through the CDR. The call detail record simply shows that the calls or messages took place, and measures basic properties.

2.3.1 CDR Data Specification

A CDR for a phone number provides you with more of a bird's eye view of how your phone system is used at a company-wide stage [17]. The operator commonly to provide the following standardized information:

- International mobile equipment identity (IMEI) of both calling and called party is unique for each device.
- International mobile subscriber identity (IMSI) of both calling and called party is unique for SIM (subscriber identification module) card.
- Time stamp of call-start and call-end format is YYYY-MM-DD HH.
- Base station calling party identity and coordinates, cell identity, longitude, latitude.
- Base station called party identity and coordinates: cell identity, longitude, latitude.
- Mobile phone number of calling party.
- Activity type, voice, SMS, or data.

It is useful to produce CDR reports for individual workers to find out who is on the phone the most, what the length of the call is, and how much money each employee spends per call. This is helpful if your organization has several phone numbers for various places or divisions. This kind of report allows you to take an aggregate look at company's usage activities or behaviors.

2.3.2 Subscriber Information Vs CDR

Although the CDRs can provide a lot information, what they cannot tell us is who actually made the activity [18]. Therefore, it's vital to understand the distinction between the CDR and therefore the subscriber info. Subscriber information and CDRs aren't an equivalent. Basically, subscriber information would include things such as the name, address, and telephone number.

For billing and other reasons, carriers typically maintain careful records of customers and their activities. During an investigation, this data stockpile can be incredibly helpful. These transporter records can reveal to us the endorser's name, address, extra telephone numbers,

Social Security number, etc. These ought's not be mistaken for cost records. Cost records allude to landline data as opposed to cell phones. When requesting the call detail records, you should determine a date range. It's an astute practice to cushion your solicitation with a little while on the two closures. A full list of inbound and outbound calls is not represented by the billing records. The call logs contain data that has not yet been included in the billing system.

It is likely that information maintained by the carriers have a short, predetermined shelf life. Each carrier has some control as to how such information is processed and how long it is stored. This is generally defined in the retention policies of the company. This helps to ensure that your evidence won't get purged before it can be preserved and collected. We all love to chat, email, and browse the internet on our cell phones all over the world. Whenever we use the phones in our pockets to send SMS, make phone calls, send money or use mobile data, data is created by operators that provide our handsets with network service. This metadata is a Call Detail Records or CDR. A CDR contains information such as the time of day and a rough location where the activity took place, but never the content which was sent, be it voice message or an SMS. Network operators then use this data for billing purposes, understanding the efficiency of the network or business and other business management operations.

2.3.3 Uses of Call Detail Record

Wide collections of anonymized CDRs may extract surprisingly useful knowledge about human actions. They are mainly used to assist in call monitoring and billing by corporations. CDRs are used by billing agencies to settle conflicts, maintain track of how money is allocated, and document the telephone system's use. IT agencies may also use CDRs to assess if telephone service failures have occurred [18].

Mobility: The movement patterns of a group can be reconstructed as mobile users send and receive calls and messages from distinct cell towers.

Social interaction: For both constructing demographic profiles of aggregated call traffic and recognizing changes in behavior, the geographic distribution of one's social interactions can be useful.

Economic activity: tracking airtime spending, i.e. the amount of money expended during a network link, for patterns and abrupt shifts, may be useful both for detecting the early

impact of an economic recession and for assessing the impact of livelihood improvement programs.

CDRs are also valuable to examiners.

- **Evaluating patterns of telephone use**: It is possible to extract several use patterns from CDR data. The calling activity of individuals in various demographic groups is one example. This includes both spatial and temporal data such as number of calls, call average, call duration for established user groups (such as male, female, student, worker, housewife), position (where individuals are in the same user groups), minutes of usage per user on average, local call percentage, long distance call percentage, roaming percentage, idle period local call percentage, idle period long distance call percentage, idle period roam call percentage.
- **Build customer profiles**: Operators may create customer profiles from the trends and then build a pricing system to optimize performance, whether it is based on mobile usage data or customer service calls, the information available in call detail reports can help enhance services and encourage opportunities or ways to shorten call time.
- **Sales forecasting**: Customers make calls or use services, and by forecasting service demand, operators can enhance services and network performance. Analysis of usage may assess a scheduled obsolescence plan or classify complementary services. Forecasting also looks at the number of clients who can forecast consumption and estimate revenue from consumer categories, market share.
- **Identifying fraud or overuse on mobile phone:** Through the use of historical information of CDRs, operators can spot unusual calling behavior and over-use behavior, and can help operators to limit or stop related services, and alert customers etc.

2.3.4 CDR data Limitations and Difficulties

Although the use of CDR data can provide many advantages in different sectors and applications, as mentioned above, there are some barriers and difficulties that hinder the use of CDR data.

- **Privacy concerns**: It gather and record other information as cell phones transmit voice, not only intrinsic to CDRs but also user-specific and personal data, and this

poses new problems in terms of the tension between technological advancement and privacy rights.

- **Accuracy**: accuracy of their geographic location or approximate position is one major factor that could affect the continued progress of local research on these mobile devices. Nearly all telecommunication network CDR data use the position of the base tower to infer the geographical location of the devices. To mitigate the accuracy issue, the report applies an estimation method of people stay locations and movement routes.
- **Availability of data:** the research and analysis of cell phone data in most countries is limited to the availability of operators' data. While data sets have become available in recent years and have opened up the possibility for researchers to carry out large-scale study of urban and social impacts, both mobile industry support and data access are still very small.
- **Data discontinuity**: When individuals use their mobile phone, call information records are created, creating a lack of consistency in consecutive access points in CDRs, creating a key user location issue when analyzing the data. The discontinuity issue is tackled by the estimation method of people stay locations and movement routes, described above. In addition, and despite industrywide formatting standards, the variation of data format requires much data pre-processing and cleaning before analysis.

2.4 Customer Segmentation and Customer Profiling

Customer segmentation is the process of separating clients into groups based on shared characteristics so that businesses can efficiently and accurately market to each group. A company could segment customers according to a wide range of factors in business-to-business marketing. Segmentation is a way for clients to provide more targeted contact. The segmentation method defines within the data the characteristics of the customer groups (called segments or clusters). Consumer profiling is achieved by creating the behavior model of a customer and calculating its parameters. Customer profiling is a way for a population of potential customers to submit external data. It can be used to recognize potential customers or to recognize current bad customers, depending on the data available. We can have segmented customers based on demographic, psychographic, geographic, or usage behavior data of customers [19]. Depending on their location, buying habits, gender,

language, local traditions, education, past purchases or many other items, we can segregate our clients. We need to decide on which variables we want our customers to be separated and why you want to do so.

Geographic segmentation targets customers based on a predefined geographic border. Such as by city, country, state, region or country.

Demographic segmentation is based on age, ethnicity, religion, gender, family size, income from race, educational background, occupation, date of birth and education. In order to help a company, reach customers more effectively, demographic data may be segmented into different markets.

Psychographic segmentation based on common psychological traits, clients are divided into subgroups, including subconscious or conscious views, motivations and goals to describe and forecast consumer behavior, purchasing patterns, spending habits, and other values.

Behavioral segmentation. Process of dividing customer based on usage frequency, purchasing behavior, benefit sought, buying pattern of customer., customer journey stage, occasion or time based, customer satisfaction, customer loyalty and interest .in this paper we use behavioral one

For example, by getting more extra features built into the product and another segment that wants the product to be inexpensive, with minimal services, one segment of the client able to pay a much higher price. Based on this segregation you might want to develop two separate products for these two segments.

2.4.1 Advantages of Customer Segmentation

Therefore, any marketing activity aimed at impacting the consumers in the long term involves their segmentation, first and foremost, as it is important to consider in advance how any specific marketing action is likely to impact them. One of the key market developments as telecommunications offerings grow is that operators are seeking to find ways through their customer service organization to engage in a dialogue with their customer base. A segmentation-based approach to service helps to deliver on both customers and management's expectations by providing the right level of service for each customer or customer group based on their value and needs.

In basic terms, the benefit of consumers is not equal; their treatment plans should also not be equal. Currently, [19] new platforms such as mobile, social media and the Web create opportunities for telecom operators to follow a segment approach that

enable them to improve their quality of service. It helps to recognize the least and most successful clients, thereby allowing the company to focus marketing efforts on those that are most likely to purchase the goods or services; helps create loyal customer relationships by cultivating and supplying them with the products and services they want and also helps to enhance customer service; to optimize the use of your resources; to change or tweak goods to meet customer requirements; to raise profit by reducing costs.

Base your attention on marketing: understanding your most valuable segments encourage you to change your marketing efforts to more specifically target those segments. To cater more directly to them, you may edit your advertisement language and message. We can also change our advertisements for each product line to appeal to the corresponding segments that buy that product line.

Prioritize multiple segments: segmentation will help you see which of the most valuable consumer groups are for you. You can see more quickly which groups are going to do the most business with you and purchase the most costly goods or service packages. We can then compare this to the money you spend advertising to them or serving them to assess which are the most profitable. This information will allow us to reassess our customer prioritization and, by doing so, we can increase our profit.

Improve your offerings: we may make improvements to them to better cater to these groups until you know exactly who is purchasing and of your goods or services. These modifications would make the product more useful to the community in question and more desirable. This, will give us an advantage over competitors and may increase customer satisfaction.

2.4.2 Challenges in Making Segmentation

To manage with today's dynamically fragmented customer landscape, segmentation is important. Marketers are more successful at channeling capital and finding opportunities by the use of segmentation. It is not an easy task to create user segmentations.

Difficulties in making good segmentation are [11]:

- **Relevance and quality of data are essential to develop meaningful segments:** If the business has inadequate consumer data, the importance of an inaccurate and almost useless customer segmentation. Instead, too much data will contribute to complicated and time-consuming research. Bad data organization (different formats, different source systems) often makes it difficult to extract interesting data. Furthermore, the resulting segmentation can be too complicated for the organization to implement effectively. Many of these problems are due to an inadequate customer database.
- **Intuition:** while data can be highly insightful, in order to determine the 'right' data for analysis, data analysts need to constantly establish segmentation hypotheses.
- **Over-segmentation:** A segment can become too small and/or insufficiently distinct to justify treatment as separate segments. Data mining techniques that belong to the clustering algorithms group provide one solution for constructing segments.

2.5 Data Clustering

Clustering and classification are the two types of machine learning methods which characterize objects into groups by using different features. In today's information age, the role of grouping is highly important, as the vast increase in data makes it easy to process. The clustering and classification processes tend to be partially identical, although there is a gap in the sense of data mining between them. The fundamental qualification among arrangement and group is that characterization is utilized in directed learning method where predefined names are allotted to cases by properties in supervised learning technique where predefined labels are assigned to instances by properties wherever as cluster is used in unattended learning that similar instances classified, supported their features or properties. The class label of the training tuple is known and then evaluated when the training is given to the method, which is called supervised learning. On the other hand, unsupervised learning does not require preparation or learning, and no prior knowledge of the training sample is available.

Clustering is that the method of dividing or grouping the information into variety of teams specified data points within the same team’s square measure additional almost like different information points within the same cluster than those in other teams [20]. instead of shaping teams before gazing the info, clustering allows USA to search out and analyze the teams

that have grouped naturally. For instance, we could be interested in finding representatives for homogeneous groups, in finding natural clusters and describe their unknown properties, in finding useful and suitable groupings (*searching from useful classes*), in finding unusual data objects (*outlier detection*), grouping of search results, suggestion of related information, recommendation of contents and products etc. machine to learn.

2.5.1 Similarity Measure Techniques

For certain machine learning techniques, identifying measurements of similarity is a prerequisite. Similarity is necessary in a machine learning context in order to compute the closeness between elements in a dataset. This allows the structure [21] to be understood within the input data. Furthermore, it is widely used in clustering and classification tasks in order to find new instances and the best match that is already known.

Cosine Similarity

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. When data are represented as term vectors, the similarity of two data items corresponds to the correlation between the vectors. The cosine similarity of two data on the vector space is a measure that calculates the cosine of the angle between them [22].

For two data d_i and d_j the similarity between them is defined as [23]:

$$\mathbf{Cos}(\mathbf{d_i}, \mathbf{d_j}) = \frac{\mathbf{d_i}, \mathbf{d_j}}{||\mathbf{d_i}|| \, ||\mathbf{d_j}||} \quad (1)$$

The two data are identical when the cosine value is 1, and 0 if there is nothing in common between them.

Jaccard Coefficient

For text documents, Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two document but are not the shared terms [23]. If we have two documents 'a' and 'b', let, terms on a 't_a', terms on b 't_b' then the formal definition of Jaccard similarity is:

$$\boldsymbol{SIM\ Jaccard}(t_a, t_b) = \left(\frac{t_a . t_b}{|t_a|^2 + |t_b|^2 - t_a . t_b}\right) \quad (2)$$

The Jaccard coefficient ranges between 0 and 1. It is 1 when $t_a = t_b$ and 0 when t_a and t_b are disjoint, where 1 means the two things are the same and 0 means they are completely different. The corresponding distance measure '$\boldsymbol{D_J}$' is defined as:

$$\mathbf{D_J = 1 - Sim\ Jaccard} \tag{3}$$

Euclidean Distance

In two or three-dimensional space, it is the ordinary distance between two. Euclidean distance, including clustering text, is commonly used in clustering issues. Measuring distance between text documents, given two documents d_a and d_b represented by their term vectors t_a and t_b respectively with their weight values $w_{t,a}$, $w_{t,b}$, the Euclidean distance of the two documents is defined as [22, 23]:

$$\boldsymbol{D_E}\left(\overrightarrow{\boldsymbol{t}}_{\boldsymbol{a}}, \overrightarrow{\boldsymbol{t}}_{\boldsymbol{b}}\right) = \left(\sum_{\boldsymbol{t=1}}^{\boldsymbol{m}} |\boldsymbol{w_{t,a}} - \boldsymbol{w_{t,b}}|^{\mathbf{2}}\right)^{\frac{\mathbf{1}}{\mathbf{2}}} \tag{4}$$

Where the term set is $T = \{t_1, t_2, t_3 \ldots, tm\}$.

Pearson Correlation Distance

This distance is based on the Pearson correlation coefficient, which is the calculation of the degree to which the sample values and their standard deviations are connected and determined by two vectors [22]. The correlation coefficient (c) takes values from -1 (large, negative correlation) to +1 (large, positive correlation). The pearson distance (*pd*) is computed as $pd = 1 - c$ and lies between 0 (when correlation coefficient is +1, i.e. the two samples are most similar) and 2 (when correlation coefficient is -1). Furthermore, more similar two vectors are, the shorter their distance will be. The distance will approach 0 as the correlation goes to 1.

2.5.2 Clustering Approaches

Most of important and commonly used clustering algorithms fits into either hierarchical clustering or non-hierarchical clustering approach [24, 25]. Where the aim is to generate clusters that do not fit into a given information hierarchy, non-hierarchical text clustering is applied. The hierarchical method will group, for instance, two similar clusters within a major cluster, such as taxonomy, if a hierarchy is required to organize the texts.

Hierarchical clustering methods do not create a single clustering result, but the whole hierarchy of clustering.

Hierarchical Clustering Approaches

Hierarchical clustering involves creating clusters that build a tree of the data that successively merges similar groups of point data. Algorithms for hierarchical clustering are either top-down or bottom-up [25]. At the beginning, bottom-up algorithms treat and data as a single cluster and then merge pairs of clusters successively until all clusters have been merged into a single cluster containing all records. There are different methods for doing bottom-up (*agglomerative*) clustering. **Single linkage** method defines the gap between 2 clusters to be the minimum distance between any single datum within the initial cluster and any single datum within the second cluster. **Complete linkage** method defines the space between 2 clusters to be the utmost distance between any single information within the initial cluster and any single information within the second cluster.

Non-hierarchical Clustering Approaches

A non-hierarchical approach generates some categories by partitioning a dataset [24, 25]; giving a set of non-overlapping groups having no hierarchical relationships between clusters. The data is partitioned into a group of K clusters in a non-hierarchical system and this can be a random partition or it can be a partition based on a first guess at seed points forming the initial cluster centers. Data points are then transferred iteratively through various clusters until no reassignment is feasible. There are a number of techniques for non-hierarchical clustering, but we have described K-means which is widely used in text document clustering.

K-means Clustering

K-means is a popular unsupervised clustering algorithm used when we have unlabeled data (data without defined categories or groups) [24, 26] to organize the data. There are many methods of estimating K but there is no method for determining exact value of K. One simple rule for deciding the optimum number of clusters (K) to have is $K = \sqrt{N/2}$.

Steps for basic k-means clustering algorithm is given below:

Let $X = \{x_1, x_2, x_3 \ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots, v_c\}$ be the set of centers.

1) Randomly select *c* cluster centers.
2) Calculate the distance between each data point and cluster centers [26].
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers [24].
4) Recalculate the new cluster center using:

$$\boldsymbol{Vi} = \left(\frac{\mathbf{1}}{\mathbf{Ci}}\right)\sum_{\boldsymbol{j=1}}^{\boldsymbol{Ci}} \boldsymbol{Xi} \qquad (\mathbf{5})$$

Where, c_i represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3.

The results of the K-means clustering algorithm are [26]:

- The centroids of the *K* clusters, which can be used to label new data.
- Labels for the training data (each data point is assigned to a single cluster).

Spherical k-means is the most popular method of clustering text data in which algorithm takes cosine similarity between data [27]. In grouping (clustering) process, each cluster mean vector is updated, only after all document vectors being assigned, as the (normalized) average of all the document vectors assigned to that cluster. Spherical k-means Algorithm is given as:

1) Normalize each data point
2) Clustering by finding center with minimum cosine angle to cluster points
3) Similar iterative algorithm to basic k-means

K-means algorithm does not depend on the order, where, n is count of points, k is the count of clusters and i is the number of iterations.

Density based clustering

Density based clustering is based on the idea that a cluster in a data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. The data points are usually called noise/outliers in the dividing regions of low point density [28]. Clusters are known to be data space regions in which the objects are dense and which are separated by low object density regions (noise). The basic clustering algorithm based on density that can be used to find flat clusters is DBSCAN. The density based clustering results is influenced by parameter, because in most cases it needs to define the neighborhood density threshold and radius.

Steps for basic density based clustering algorithm is given below:

1) Choose a random point 'r' radius.
2) Calculate all data points which satisfy density from 'r' with respect to radius and density i.e. minimum points.
3) If 'r' is a core point, then it forms a cluster.
4) If 'r' is a border point, no data points reach the density from 'r', then the algorithm goes to the next data point in the space.
5) Repeat the process until all the points in the space are covered.

Density based clustering algorithm can create nonlinear set of clusters and It is not sensitive to the noise. Density based clustering is the second best clustering method after k-means and the complexity is low.

Expectation Maximization Based Clustering

Expectation maximization (EM) is a well-known iterative clustering method for learning probabilistic categorization model from unsupervised data. The expectation maximization clustering method initially assume random assignment of examples to categories. It uses the following two steps until convergence: Expectation (E-step) where each object is assigned to the centroid such that it is assigned to the most likely cluster: Compute probability for each example given the current model, and re-label the examples based on these posterior probability estimates. Maximization (M-step): Re-estimate the model parameters from the probabilistically re-labeled data. Where the model (centroids) are recomputed.

2.5.3 Clustering Evaluation Techniques

The final goal of clustering is attaining high intra-cluster similarity (similarity of data within a cluster) and low inter-cluster similarity (similarity of data from different clusters). When comparing a cluster solution, we can consider internal and external quality of clustering, the standard measures of Purity, Entropy, F-measure and recall, precision are often commonly used to determine the quality of clusters [29].

Choosing Appropriate Number of Clusters

In the literature for measuring clustering outcomes, a number of methods have been suggested. In order to design the method of testing the results of a clustering algorithm, the term clustering validation is used. There are more than thirty indices and methods to classify the optimum number of clusters [30], so we only concentrate on a widely used clusters here.

Elbow Method

The most notable strategy, in which the number of squares at each number of groups is determined and envisioned, and afterward we search for a difference in slant from steep to shallow (an elbow) to decide the ideal number of bunches. The Elbow Curve strategy is helpful in light of the fact that it shows how expanding the quantity of the bunches contribute isolating the groups in a mean.

The optimal number of clusters can be defined as follow:

1. Compute grouping calculation (e.g., k-implies clustering) for various estimations
2. 2For every k, ascertain the absolute inside group amount of square.
3. Plot the bend of amount of square as per the quantity of bunches k.
4. The area of a twist (knee) in the plot is for the most part considered as a pointer of the fitting number of grouping.

Silhouette Method

Another visualization that can help determine the optimal number of clusters is named as the silhouette method. Normal outline strategy registers the normal outline of perceptions for various estimations of k. The ideal number of bunches k is the one that expand the normal outline over a scope of potential qualities for k.

The algorithm is similar to the elbow method and can be computed as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k.

2. For every k, compute the normal outline of perceptions.
3. Plot the bend of normal outline as per the quantity of bunches k.
4. The area of the most extreme is considered as the proper number of bunches.

Precision and recall

Basically, when we consider **precision** and **recall** from information retrieval concept, each cluster is considered as if it were the result of unsupervised clustering and each class as if it were the desired set of data for the category. Formally defined as follows [31]. Which are widely used to evaluate the performance of unsupervised learning algorithms.

$$\boldsymbol{Recall}\ (\boldsymbol{i},\boldsymbol{j}) = \frac{\boldsymbol{n_{ij}}}{\boldsymbol{n_j}}; \quad \boldsymbol{Precision}(\boldsymbol{i},\boldsymbol{j}) = \frac{\boldsymbol{n_{ij}}}{\boldsymbol{n_i}} \tag{6}$$

Where $\boldsymbol{n_{ij}}$ the number of data with class is label *i* in cluster *j*, $\boldsymbol{n_i}$is the number of data with class label *i* and $\boldsymbol{n_j}$ is the number of data in cluster *j*.

Accuracy is defined as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0. The higher f-measure is the higher accuracy of cluster.

Other measurement method related to the internal quality of clustering is ***entropy*** measurement and it is defined as [34]:

$$\boldsymbol{E_j} = -\sum_{\boldsymbol{i}} \boldsymbol{P}(\boldsymbol{i},\boldsymbol{j}).\boldsymbol{LogP}(\boldsymbol{i},\boldsymbol{j}) \tag{7}$$

where, P (i, j) is probability that a data has class label i and is assigned to cluster j.

Thus, the total entropy of clusters is obtained by summing the entropies of individual cluster weighted by the size of each cluster. The lower value of entropy, the higher quality of cluster.

Purity is an external evaluation technique of cluster quality. The *purity* measure evaluates the coherence of a cluster, that is, the degree to which a cluster contains data from a single category [31]. Purity can be interpreted as the classification rate under the assumption that all samples of the cluster are predicted to be members of the actual dominant class for the cluster. High purity is can be easily achieved when the number of clusters is large; purity is 1 if each data gets its own cluster. External measures are related to how representative

are the current clusters to true classes. The purity and entropy measure the ability of a clustering method, to recover known classes.

2.6 Summary

This chapter explained about clustering which is the process of grouping similar data into different groups, moreover the partitioning of a data set into subsets, so that the data in each subset according to some defined similarity measure. Mostly, clustering deals with unsupervised data; thus, unlabeled whereas classification works with supervised data; thus, which are labeled. This is one of the major reasons why clustering does not need training sets while classification does. Clustering algorithm are useful for exploring data. K-means is especially useful and commonly used. The most common clustering approaches are discussed in this chapter and also concept-based and bag-of-words-based clustering techniques. Similarity measures play an increasingly important role in text related research and applications in tasks. The similarity measure is the measure of how much similar two data items are. Similarity measure in a data mining context is a distance with dimensions representing features of the data item. If this distance is small, it will be the high degree of similarity where large distance will be the low degree of similarity. This chapter also reviewed literatures on commonly used text similarity measurement techniques. We have also reviewed and discussed the most common clustering evaluation techniques. Evaluating the clustering result shows how well the clustering is performed.

Chapter Three: Related Work

3.1 Introduction

This chapter discusses problems and solutions proposed for mobile users from customer detail records using unsupervised and supervised techniques. We concentrate on reviewing research work carried out on mobile users to increase the competitiveness of the mobile operator, addressing consumer value, marketing strategy and service provider quality of service using many data mining technologies. Customer clustering and segmentation is one of the most significant technologies for business management. In this chapter also describes how and where the data set is collected from and to what application domain the research is done. While many methods have been suggested to determine

the best method, it is still very difficult to identify the method of grouping customers based on their actions, which is the best method to use, and many works on mobile customer segmentation are reviewed in this part. In addition, to what extent the authors achieved their work and what basic data analysis technologies was used during the implementation and experimentation of the research works is assessed.

There are many works done in the area of segmenting mobile customers in different researchers in different place [4, 5] using different telecom company CDR data based on different attributes like call detail records, customer profiling. customer behavior and others with purpose of marketing analysis, customer management, for improving profitability of one company and used different data mining techniques for different mobile operators like china, Pakistan, Korea, India which are discussed in the following sections.

3.2 Dimensionality Reduction in Customer Segmentation

Maha Alkhayrat, Mohamad Aljnidi and Kadan Aljoumaa [37] work on dimensionality reduction to reduce number of attributes in data analysis. In Telecom companies there are logs customer’s actions which generate a huge amount of data that can bring important findings related to customer’s behavior and needs. The large number of features and high sparsity of such data are the key characteristics that make analytics difficult. In order to obtain better quality clustering performance, this paper will investigate dimensionality reduction on a real telecom dataset and test customers' clustering in reduced and latent space compared to original space. There was dataset contains 220 features that belonging

to customers. However, with the existence of the dimensionality curse, dimensionality reduction is an important data preprocessing phase in the data mining process. Then K Means Clustering was applied on both-original and reduced data set. Different internal measures were performed to evaluate clustering for different numbers of dimensions and then evaluated how the reduction method impacts the clustering task. This paper was done to perform feature selection and high dimensional data reduction with different techniques to examine their performance in telecom customer segmentation as case study.

3.3 Customer Segmentation to increase Loyalty

Shohin Aheleroff and M. chandini [5] work of clustering mobile customer research was performed on the basis of their call information records, event detail records and consumer behavior analysis. From CDR (call detail records) analysis of a mobile operator, located in the Middle East that has about 35 million mobile subscribers, the key purposes of customer segmentation in this paper were customer retention to increase the loyalty and avoid churning. This paper concentrated on proposing a structure for consumer segmentation based on real customer actions. Instead of their personal details such as gender, address and revenue, the author used a collection of useful information to classify subscribers' core needs based on their call detail record. A Call Detail/Data Record includes at least the following: the number making the call, the number receiving the call, when the call began, how long the call was (duration), the type of call (voice call, SMS, data (content), GPRS/MMS, balance before and after, mobile generator and terminator location, incoming and outgoing voice, incoming and outgoing SMS and various content types. In addition to CDR the work also considers subscribers interest to active or change any service by capturing their action via analyzing Event Detail Records. They enabled to ensure that consumer behavior is captured by collecting CDRs from multiple sources and that the author evaluates what they are more and less interested in. Many segmentation algorithms and software applications such as SAS and SPSS are already developed but the most important step that the researcher followed is first CDRs& EDRs collected (by push or pull mechanism into a data warehouse) next to that the researcher select Different services selected like GPRS / MMS, SMS, Content e.g. (. RBT, Wallpaper, Java application, Push mail, Music, Clip) and Voice. The k-means as a well-known segmentation algorithm was used. Next to that Customer segmentation as per each recognized factor to generate a matrix of segments. At the end using segmentation output for loyalty and customer churn

application. At the end the work recommends based on the degree of customer loyalty over their life cycle, the service provider really needs to keep the plain loyalty motivated and also provide sufficient incentive and package to increase their loyalty.

3.3.1 Customer Segmentation Based on Common Attributes

S.M.H. Jansen and Yoonseong Kim [6] conduct work on Vodafone which is world's leading mobile telecommunications company, with approximately 4.1 million customers in The Netherlands. The customer data contains information of the telecommunication customers. This research addressed the question how to perform customer segmentation and customer profiling with data mining techniques. In this work context, customer segmentation is a term used to describe the process of dividing customers into similar groups on the basis of shared or common attributes (habits, tastes, etc.). Customer profiling describes clients by their characteristics, such as age, gender, revenue, and lifestyles. Managers can decide which marketing actions to take for each segment by having these two components. In this study, the segmentation of customers was based on the behavior of the usage call. A number of advanced and state-of-the-art clustering algorithms are modified and applied for creating customer segments which include k-means, k medoid, Fuzzy C-means, was used. An optimality criterion was constructed in order to measure their performance. The best i.e., most optimal in the sense of the optimality criterion clustering technique will be used to perform customer segmentation. The researcher segment customers into different group using different clustering algorithm. Each segment will be described and analyzed. With information from the data warehouse, such as age, gender and residential area information, customer profiling can be achieved. Finally, a new data mining method, called support vector machines, calculates the segment of a customer based on the profile of the customer. Different kernel functions with different parameters examined and analyzed.

3.3.2 Customer Segmentation based on Revenue Attributes

Masood et al. [32] conduct research on customer segmentation using clustering algorithms on real data of a telecommunication company in Pakistan with around the number of mobile subscribers has reached around 123 million. In the first step, the data was cleaned and transformed into categorical variable types. Next, correlation analysis was performed so as to select the attributes for clustering. The author used the two-step clustering algorithm to create different customer segments after choosing suitable attributes for clustering. This

was followed by clustering on customer call usage information, sales data and data from recharge analysis, resulting in five revenue segments. The researcher segmented customers based on their revenue attributes such as Final Revenue, Calls Revenue, SMS Revenue and VAS Revenue. The next step was to further segment each revenue section into five sub-segments based on the call and SMS use of the customer. Finally, before suggesting marketing strategies for upselling and better targeted campaigns, the insights gained from each segment were analyzed.

3.4 Segmentation for Improving Profitability

Arora1 et al [5] conducted a work that helps in identifying the activity segmentation of mobile customers of different groups will be done based on some rules. In this work, clustering of mobile customers is done based on the call detail records and using Self Organizing Maps (SOM). SOM transforms high-dimensional data to a lower-dimensional representation scheme that can be easily visualized and interpreted (a two-dimensional map). In this article, under the categorization of network providers, consumers are segmented into categories. Clustering technique of data mining is used to solve the problems of this work. K-Means [10] is one of the simplest unsupervised nonhierarchical learning methods among all partitioning based clustering methods. It classifies a given set of n data objects in k clusters, where k is the number of desired clusters and it is required in advance. For each of the derived measures the work uses those frame work to solve the problem are first attribute selection the second one is database framing next to that the author calculate the score finally customer classification was done.

Total score was calculated by summing the STD calls, local calls, local SMS, STD SMS, roaming, tariff plan, and data plan. Then the researcher gets total score for every entry package for mobile customer. Then total score is computed simply getting sum of all packages. Revenue maximization profiling-use of clustering for identifying maximal cluster. Revenue has been calculated manually. At the end customers are segmented into groups under the categorization of network provider, was done using different datasets. From these datasets the authors calculated score then according to the score they classified the customer as Premium and Non premium. A, B, C & D are different –different type of network provider. There are two categories of customers in different-different networks Premium and Non premium. According to revenue maximization A and D are Non premium type customers and B and C are premium type customers. In the premium and

non-premium Type-B and D both are dominant respectively. According to usage pattern total of all Customer-B is the dominant. At the end the author recommended marketing managements can use the facts above to design distinguishable marketing strategy in order to get better marketing results. Following are some examples: design suitable short message price policy for customers of cluster which has highest short messaging index value. Marketing managements need to encourage customers having low consuming ability to use more mobile service. It can be known which group of mobile customers often travels out and suitable roaming policy could be designed for customers belonging to cluster.

3.5 Customer Data Analysis Model

Y. Gopi, [10] work on multilayer clustering model for mobile telecommunication client analysis enhances customer relationship management and focuses on a dynamical marketplace. This model provides a way for companies to plan for long-term CRM and retain high-quality customers. To increase customer retention and customer manage easier this study will provide telecom companies with an approach based on customers' operational needs (including voice, data, and other business) in ac-accordance with the two-layer clustering model for customer segmentation. This research starts by collecting and segmenting the contributions of individual customers, personal preferences, overall customer profile, and other variables in this proposed model, examining customer value in the first layer, and using consumer behavior features for further grouping in the second layer. The behavior of special, data-oriented billed users (customers whose main need is for mobile Internet) becomes an independent group. Mobile customers in the first layer are divided into a total of seven large groups, S1–S7. Clustering algorithms such as k –means two-step is used to establish the cluster model according to customer attributes, behavior preferences, and contract status. Each L1 cluster is split into five-to-seven different L2 clusters. The optimal clustering model is then selected as the basis of the final L2 clustering, based on conditions such as the maximum and minimum cluster ratio, the silhouette coefficient, and readability for marketer. In this study, customers form appropriate segments, which help the company focus on its target customers and then develop CRM, marketing strategies, and promotional activities.

Su-Yeon et al. [33] proposed a framework for analyzing customer value and segmenting customers based on their value. The work first describes a conceptual prototype based on LTV for customer segmentation and then proposes a calculation model applicable to a

wireless telecommunications company for measuring customer value. Next, they apply a wireless company's real data to the model and, with the result of customer value derived, perform customer segmentation. In order to understand the consumer defection, the author utilizes three dimensions, current value, future value and customer loyalty. Background of this article. Present value becomes a measure of the past profitability of consumers, future value becomes a measure of additional revenue possibilities, and customer satisfaction may be a measure of customer retention.

Current value = (Average amount asked to pay for a customer - Cumulative amount in areas for the customer/total period of user)

After segmenting the customer base with three viewpoints, a segment analysis is performed according to the segmentation results. At the end the results and analysis of the papers is depending on the result of Decision Tree are the nodes remaining until the final one. But cases with too small population or insignificant classification were excluded. And then the Characteristics of customers in the low customer loyalty, high current value, high potential value is present. Also, by using the outcome of the decision tree, the researcher develops refined strategies for a company to plan these five possible marketing strategies. Strategy to charge membership fee, strategy to upgrade a phone device, Strategy to provide better services for loyal customers, strategy to strengthen a brand image

Qining Lin and Yan WAN [4] discussed based on the call information records, how to cluster mobile customers and evaluate their user behaviors using various data mining techniques in which the data was obtained from a mobile operator in northern China has approximately 600,000 mobile customers. This research was focusing on cluster by call detail record they have more information to describe customer behavior than billing system data. The authors used one of most important data mining technologies which is customer clustering analysis. The clustering model in this paper was to categorize prospective customers into distinct groups for distinctive contact strategies and proximal offerings. The work explains many clustering methods, for example, fuzzy clustering method, system clustering method, dynamic clustering method and K-means clustering method. But the K-means method of cluster detection was used in this work. The author first prepares the data for clustering get the data of call detail records from data warehouse then clean dirty or outlier data after that Define and calculate some index and finally form the data table for clustering. After preparing the data of call detail records, K means clustering method was

applied. Then all mobile customers are clustered into 15 distinguishable groups using the mining software tools. Then, an application of a mobile customer clustering analysis is given in this paper. At the end the author recommended marketing managements can use the facts above to design distinguishable marketing strategy in order to get better marketing results. For example, from their utilization behavior, the author sees that the mobile customers of group1 often travel out, and marketing managements need take special care for their travel so as to improve their satisfaction.

3.6 Identifying User Habits Using Call Detail Records

Filippo et al [34], proposed a data mining process to automatically identify the common features of mobile network users' telephone activity in order to understand and describe their habits, with a service provider located in the Orange mobile phone users with approximately 50,000 mobile subscribers. The author proposes a framework in this paper to classify patterns and regularities in the pseudo anonymized Call Data Records (CDR) pertaining to a generic mobile operator subscriber. They present an adaptive visualization process, which encodes information and data into visual objects. The main goal of the work was to communicate information clearly and effectively through graphical tools, in order to express and to quantify the results, through visual human interfaces. The approach instead is based on PROCLUS, the well-known sub clustering algorithm. The analyzed dataset contains records that are distinguished by only a few features and therefore illustrate how to generate additional fields that define implicit information concealed in data. Finally, they test the system's output by contrasting it with a well-established subspace clustering algorithm, proposing an efficient graphical representation of the data-mining process results, which can be easily interpreted and used for constrained environments by analysts.

3.7 Segmentation Based on Smartphone Measurement

Fadly Hamka [35] work on customer segmentation was based on Smartphone measurement. Customer segmentation for mobile services was based on demographics and reported use. In this paper the author used smartphone measurement software enables to add directly observed user behavior. This paper reports on an exploratory study focused on data collected through smartphone measurement tools that aims to elicit customer segmentation. I.e. a measurement program runs on the cell phone context, periodically transmits log files to the server on user activities, and then a massive amount of log data was generated in the analysis. They analyze the results of a smartphone measurement

project among 129 participants launched mobile applications 130,000 times over the 28-day period. Aggregate metrics were computed by calculating the average number of voice calls, MS messages, MBs sent and received, applications installed, applications used and URLs browsed per day then the researcher use latent class analysis which assumes that each observation is a member of one and only one of K latent, i.e., unobservable classes, with K being a finite, natural number. After selecting the best fit model, they analyze the characteristic of each cluster based on the variance of observed variables, the analysis leads to the following clusters like Basic service users, Basic service users with cellular data, WLAN-only users, Medium overall user, Data-only users, a High overall user Next, the author describe the clusters from an application developer perspective. Six clusters result from only three observed variables, Application ignorant users, Basic application users, Average app users, Information seekers, App savvy users, High utility users.

3.8 Finding Customer Patterns Using Clustering

The work [36] presents the patterns of use for additional services that are currently provided to mobile telecommunication subscribers in Korea. The authors applied three empirical techniques to classify the service adoption trends of segmented classes, which are factor analysis, clustering and quantitative association rules. 115 input variables, including both service use details and personal information such as personal ID and the first day of subscription, are used in the mobile telecommunication service data used in this analysis. Since the service use pattern is represented by too many variables, the authors first apply factor analysis. Principal component analysis (PCA) method is employed for factor extraction of the underlying patterns from the large number of variables. The authors segment the 17,000 customers using a clustering algorithm. Several clustering methods are applied. The author used K-means clustering algorithm for eight clusters using the newton method, other clustering methods segmented the customers into too many clusters or caused imbalance in the number of customers. These eight were therefore determined by the authors as the best clusters that reveal the most important differences in demographic information and customer experiences. In order to discover important association rules for additional services for each cluster, association rule analysis is subsequently applied. Then the work allows prediction of the patterns of use for the additional services for each cluster. As a result, it was possible to divide the customers into 8 clusters. Three types of user groups will be identified from the analysis. This work utilizes the rules of association

contained in each cluster to provide strategic guidance to develop the corresponding group's mobile service sector.

3.9 Summary

In order to increase customer satisfaction service and develop strategy, the major areas and works associated with customer clustering are briefly discussed in this chapter. To get insight for our work, the different approaches, techniques, algorithms, models, and APIs are reviewed. Many related research works are reviewed in this chapter including the [33, 34] work that have conducted on customer segmentation based on CDR for marketing purpose and the researcher clusters mobile users into three groups for reason of marketing campaign with some characteristics. In this work there was recommended work and other groups which have some features can be used for marketing, but their features are not as distinguished as the above three groups. The work on [35] clustering based on customer profiling was recommended that using the application of more specialized methods than the elbow criterion could be applied and may perform better result with relatively less time, for instance. The customer profile used in this work is not sufficient detailed enough to describe the wide spectrum of customers. As the researcher recommend one reason for this is the missing data in the Vodafone data warehouse. Consequently, an enhanced and more precise analysis of the data ware house will lead to improved features and, thus, to an improved classification but in our case, this type of segmentation is impractical to our country Ethiopian because of the profile of customers which is stored in the CRM of our service provider is not updated with time except age and location of user.

Thus, different authors [34, 36] recommended to cluster customers by using different feature or attribute values for the customer segmentation. This can lead to different clusters and different segments. The next recommended work was to improve on determining the number of clusters (K) to group the data set, the application of more specialized number of cluster selection methods. Finally, we note that the study would improve noticeably by involving multiple Criteria to evaluate the user behavior, rather than mere phone usage as employed here. Similarly, it is challenging to classify the profile of the customer based on the corresponding segment alone. In future work, they intend to increase the grouping of the customer-variables selection function. For different marketing or business needs, a customer clustering model will be established to increase the flexibility of customer-

clustering applications However, this is a complex course and it essentially requires the availability of high-quality features.

classification (supervised learning) of customers to predefined categories is difficult during defining categories by seeing the usage characteristics (CDR). It also excludes thus different types of customers which are unrelated to predefined category. In clustering the idea is not to predict the target class as like classification, it's moreover trying to group the similar kind of customers by considering the most satisfied condition. Although many related research works have been performed earlier, we know no any work is done in the telecommunication market in case of Ethiopia due to, the definition of customer value and customer behavior are varying by industry. Even the same industry can have different attributes, such as the amount of consumption, the number of consumers, the number of stores, and so on. Thus, it needs different customer segmentation model for each service providers. In addition to that the design of call detail record of one service provider differ from other because that the works done for others service provider cannot be applied directly for others service provider users. This research will be conducted to explore the advantage of using data analysis for new service delivery in telecom customers using basic consumption attributes (SMS frequency, data usage, number of callings in or out and call duration). Unsupervised clustering method using mobile network usage behavior represented by attributes (voice duration, internet, SMS, number of incoming and outgoing calls) in ethiotelecom was employed in case of Ethiopia.

Chapter 4: Design of Unsupervised Clustering Model for telecom Network Customers

Introduction

In this chapter, we discuss the proposed design of ethiotelecom mobile customer clustering using Call Detail Record (CDR). Furthermore, we focus on the activities of data collection, normalization, data aggregation, feature extraction, and grouping of related telecom network customers (clustering). The clusters of network users are also discussed in this chapter.

4.1 Design Considerations

When we are designing the model, the questions how dimension of large-scale telecom data is reduced, selecting appropriate samples using sample selection methods and selecting the appropriate computational aggregate measures are technically considered.

Dimensionality reduction is defined as a basis of representation within a data which we can use for clustering but not all of the variance within our data store, thereby holding the relevant information. Dimension reduction for large-scale data is attracting much attention nowadays because of high dimensionality causes serious problem for the efficiency of most of the algorithms. Feature extraction and data aggregation is used to reduce computational overhead, making it easy for clustering.

Data understanding. CDR data reflects the actual usage behavior understanding of the customer. Attributes used as an in input to the clustering algorithm were related to subscribers' service usage amount, usage frequency, usage day and time, spending amount, week to week usage plan, ...etc. In the study, attribute values were aggregated or summarized in a record to indicate the usage behavior of subscribers. Hence information about each subscriber was summarized using sum, average, frequency, and ratio. In this study, Ethiotelecom CDR Data was used to build usage-based customer segments. The main tasks in this subsection are discussion on the data acquisition, data set description and data quality verification

4.2 Proposed System Architecture

We propose the potential use of telecom CDR data as a way to group customers for better decision making. In order to perform clustering based on CDR data, we considered that

the model would have the following main components: Data collection, data preprocessing (containing data cleaning, transformation, aggregation and attribute selection); feature aggregation is used to select essential attributes and construct new attributes with aggregate values. Feature selection, scaling and finally clustering algorithm is applied to create a group of related customers.

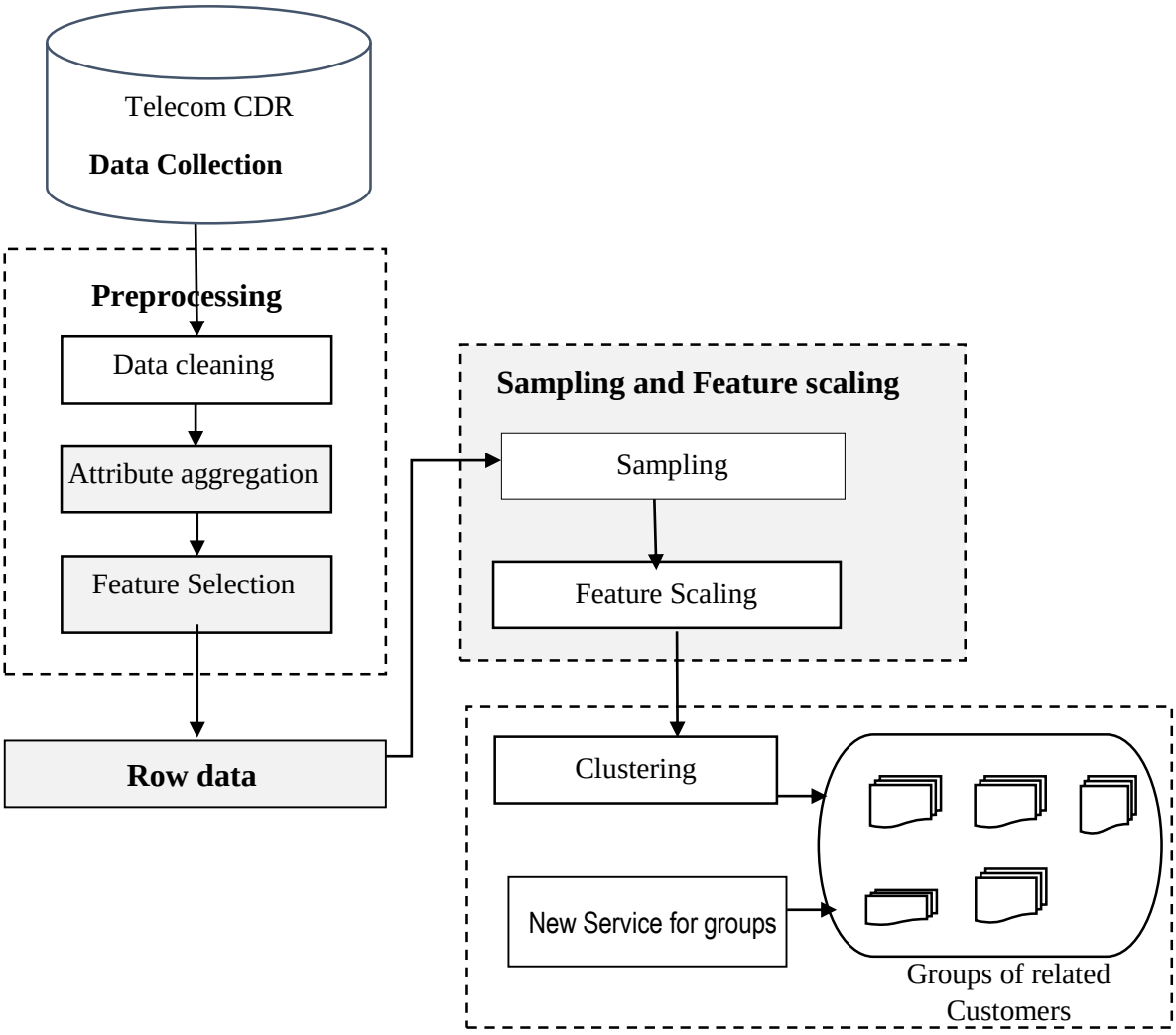


Figure 4.1: Architecture of the proposed system

In next subsections, we describe each of the components of proposed model in detail using a simple example to elaborate clustering ability of our proposed solution showed on Figure 4.1.

4.2.1 Call Detail Record (CDR) Data Acquisition

Whenever a subscriber makes a call over the operator's network atoll ticket is prepared which contains complete information of the call, including subscriber id, called number, duration of the call, call start and end time, destination and so on. This detail information about the call is named as **CDR**. For example, ethiotelecom CDR Database contain all the information to describe the important characteristics of telephone call and other telecommunication transaction of a subscriber in the network. It contains the fields necessary for billing systems. CDRs give the details of each voice, SMS and Internet data usage transaction, originating from and terminating on subscriber's device. It includes, telephone numbers involved in the call, date and time of the call, duration of the call, identification of the cell transmitting the call to the subscriber's telephone and more fields. For data preprocessing to be successful, it is essential to review overall picture of the ethiotelecom database dump.

Figure 4.2 Ethiotelecom CDR Database attributes

We have collected CDR data having more than 100,000 records of ethio telecom CDR data with the following attributes. Those attribute names are clear and easy for understanding.

Understanding CDR Data

This thesis cluster analysis was based on prepaid cell CDR data from telecom industry. So, CDR data was collected and sampled. All attributes found from the sampled CDR data cannot be used for our study. In addition, verification of the data's usefulness, completeness, redundancy, missing values, and reasonableness of attribute values with regard to the clustering objectives are considered. From one month's telecom CDR data collected we used 100,000 sampled records. After CDR data was sampled, to understand the data, we saw the statistical detail measures in the data. Irrelevant attributes to our work are removed from the collected CDR data in order to enhance clustering algorithm. In our study, customer usage behaviors were used as an input to create a group of customers. In attribute selection task advice from staff members from telecom company was used as input that will improve the time, algorithms performance and reduce complexity.

Telecom CDR Data Limitations and Difficulties

There are some difficulties that limit the use of telecom CDR Data.

Privacy concerns: Most telecom companies have their own security polices about data usage. For example, Ethiotelecom data consists of user specific and personal information, and this creates challenges in terms of the conflict between our task and rights to privacy. Thus the data we are collected and tried to use is very limited.

Availability of data: In most African countries including Ethiopia, the research and study using CDR data is limited to the availability for researchers from operators or it depends on the will of the network service companies. While data sets have become available in recent years and have opened up the possibility for researchers to perform large-scale urban, market and social impact analysis. For example, Ethiopia has one telecom service provider company (Ethio-telecom) with very limited in data availability.

4.2.2 Data Preprocessing

A CDR record is tagged with the phone subscriber's reference number, which will then be joined to the subscriber database to bill the user and other tasks. Therefore, in order to extract meaningful information from these large store of customer data and cluster data for interpretations, it needs preprocessing of those data. This activity includes data cleaning (fixing incomplete, inconsistent and noisy data) and data transformation using various methods. Data preprocessing component handles different attribute value issues that are

imposed by the nature of the stored data to make it ready for processing. The preprocessing steps have a huge effect on the success of clustering customers.

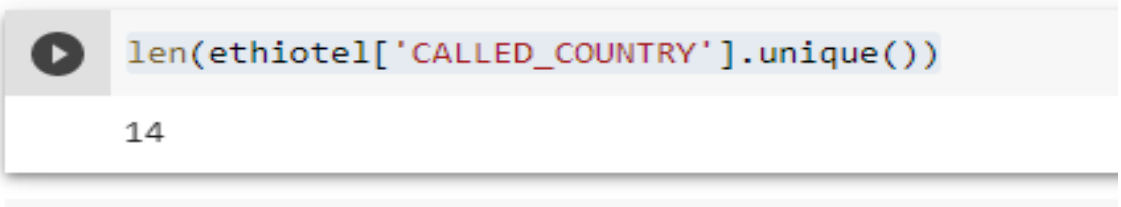
Data Cleaning

There might be missed values in the entire database dump. Before going further for data clustering activities, these missed values should be filled. There are techniques to fill missing values like ignoring the tuple, filling in the missing values manually, using a global constant to fill missed values, using attribute mean or using the most probable value. The telecom database dump has many irrelevant and missing parts. To handle this, data cleaning is done which involves handling of missing data, noisy data etc. There are tuples with multiple missed values in telecom database dump. Before going further for data analysis activities, these missed values should be filled so as to keep the result of clustering achieves the business objective. There are techniques to fill missing values like ignoring the tuple, filling in the missing values manually, using a global constant to fill missed values, using attribute mean or using the most probable value.

Among the techniques listed above, we used the most probable value and attribute mean to fill the missed values in our collected dataset. We used the attribute mean to fill the missed value for call duration. This is done using unsupervised attribute Replace missing values filters. Noisy data are, values containing errors, or outlier values that deviate from the expected. In our dataset, there are some records that deviate from the average call records in respect to CALLDURATION. But these might not be noises instead outliers. Outliers are values that deviate from the normal circumstances. From our dataset we identified and removed the outliers for the attribute CALLDURATION column. It is a two-step process; first we detect outliers using unsupervised attribute InterquartileRange filter, then we remove them using unsupervised instance RemoveWithValues filter.

Examples of cleaning and seeing details from sampled Ethiotelecom data

When we tried to see the number of called countries from the sampled data, we performed counting unique values of called country column from the sampled data. A piece of code we have used: `len(ethiotel['CALLED_COUNTRY'].unique()).`



The result shows 14 different countries are called from the sampled data.

Then when we continue to see the details on CDR about the number of records of each called country, more than 90 % of called records are from Ethiopia (+251) and United Kingdom (+44) follows as shown in the figure below.

```
ethiotel.CALLED_COUNTRY.value_counts()

251.0    106020
44.0       1475
252.0        34
291.0        16
961.0         8
965.0         7
253.0         3
237.0         3
234.0         3
256.0         2
254.0         1
91.0          1
212.0         1
Name: CALLED_COUNTRY, dtype: int64
```

Figure 4.3 number of records of each called country

Data aggregation

Data aggregation (process of gathering data and presenting it in a summarized format) is needed to experiment the daily data gathered from multiple devices and sources. In the attribute aggregation and construction task, new attributes are created from an existing set of attributes and helps in creating new attributes with values that are used for analysis. Deriving new attributes such as total_duration, total SMS, and others from the given time interval. For example, Find the total, which is the sum of SMS in and out activity, call in and out activity, and data usage activity.

- Find the total SMS, which is the sum of SMS in and out activity.
- Find the total call duration, which is the sum of call in and out activity.
- Find total_distinct_out_call which is from called_number count (MSISDN) within given date interval.

Table 4.1 CDR attributes with descriptions

No.	Attributes		Value	Descriptions
1	msisdn		Number	Mobile Station International Subscriber Directory Number

2	total_out		Number	Total number of Outgoing calls
3	total_duration		Numeric	Data of call initiation
4	total_data_usage		Numeric	Time of call initiation (calling time).
5	total_disinct_out		Number	Total number distinct outgoing calls
6	age		Number	The age of Customer
7	total _in		Number	Total number incoming calls
8	total_disinct_in		Number	Total number distinct incoming calls
9	total_SMS		Number	Total number of SMS received and sent
10	forwarding_status		{Yes, No}	Forwarded call or not

Sample SQl Query for aggregation on CDR database dump

- ✓ `select msisdn,sum(call_fee)/10000 from voice_cdr_sample where TO_DATE(call_start_time, 'yyyymmddhh24miss') > '20190625000000'and TO_DATE(call_start_time, 'yyyymmddhh24miss') < '20190725235959';`
- ✓ `SELECT msisdn, COUNT(msisdn) Visit_Count, round((SUM(upload_traffic +download_traffic))Data_Usage,to_date(call_start_time, 'yyyymmddhh24miss') Con_Date FROM CDR.data_source_table;`

4.2.3 Feature Selection

The main advantage of the call detail records is to bill customers for their usage. However, they could also be used in decision making to group the cell phone users based on usage behavior and to manage send or receive a phone call, text message, or Internet data connection. It is a process where we automatically select those features from our preprocessed data that contribute most to the result in which we are interested because attribute features that we use to clustering models have a major impact on the result we can do or in the analysis result.

Feature selection started by removing attributes having zeros or ones and other constant values, redundant value containing attributes. In addition, non-relevant attributes are also removed. The 10 numerical variables are about total number of SMS in and out activity, number of call in and out activity, Internet usage activity, msisdn of the user, and forwarding information about the call status either forwarded or not. For example, the

following table shows selected attributes and their descriptions from our data prepared for experimentation.

Table 4.2 Preprocessed row attributes with description

Preprocessed row data attributes	Description
• Msisdn	▪ Column name: MSISDN ▪ Unique for each customer ▪ Encrypted for security using MD5 Hash function
• Duration	▪ Column name: duration ▪ Duration of call in seconds
• Received _SMS_num	▪ Column name: Received _SMS_num ▪ Number of received SMS
• Data_usage	▪ Column name: DATA_USAGE ▪ Data usage record
• Call start time	▪ Column name: START_TIME ▪ Call start time records
• Call end time	▪ Column name: END_TIME ▪ Call end time records
• Called country code	▪ Column name: CALLED_COUNTRY ▪ Called country code records
• Total number of callings	▪ Column name: TOTAL_OUT ▪ Sum of number of outgoing calls
• Call forwarding status	▪ Column name: F_STATUS ▪ Status of call (forwarded or not)
• Age	▪ Column name: AGE ▪ Age of the user

4.2.4 Sampling and Feature Scaling

Sampling is a method that allows us to get information about the CDR data based on the statistics from a subset of the sample, without having to investigate every customer records. The purpose of sampling is to learn from statistical data by sampling in order to find the characteristics of customer usage behavior. In this technique, each instance in the database has an equal chance of being selected as a subject.

Sampling process is made in a single stage, with each subject selected independently of the other database instances. We follow these methods to select a simple random samples

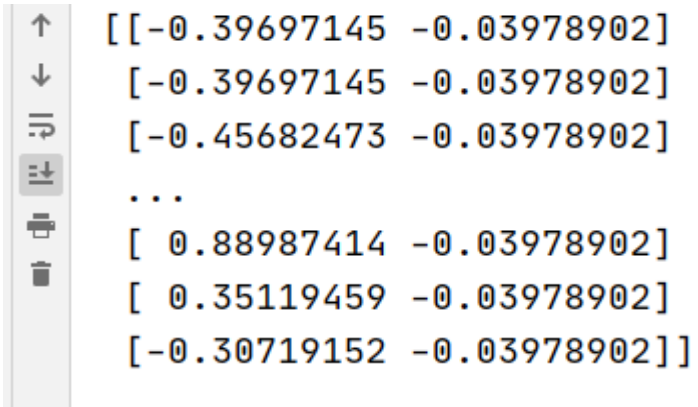
- We prepare a list of all the data records initially, and then each member is marked with a specific number (for example, if there are n^{th} members, then they will be numbered from 1 to N).
- Then we choose random samples using two ways: random number tables and random number generator software. Most works prefer a random number generator software, as no human intervention is necessary to generate samples.

We had used simple random sampling technique to take samples from Ethio-telecom CDR data. We used simple random sampling because it is a meant to be an unbiased representation for clustering. The sampled CDR data contains information of 100,000 observations after feature selection process.

Feature Scaling or Transformation

Transformed and normalized data ensures that it is less sensitive to the size of the function when training the data, ensuring that the value of coefficients can be found easily and efficiently. Many clustering algorithms perform better when numerical input variables are scaled to a standard range. This step is used to transform the data in appropriate forms suitable for clustering process. Normalization is done in order to scale the data values in a specified range or converting all data variable into a given range.

In this study we had used python module pandas data frame with mixed type columns, and applied sklearn's object StandardScalar to columns. We defined to fit on the whole data and then used to create a transformed version of the CDR data.



```
[[-0.39697145 -0.03978902]
 [-0.39697145 -0.03978902]
 [-0.45682473 -0.03978902]
 ...
 [ 0.88987414 -0.03978902]
 [ 0.35119459 -0.03978902]
 [-0.30719152 -0.03978902]]
```

Standardized print sample values to fit clustering algorithms.

On the other hand we also normalized call duration and call fee from database dump as following. From collected database dump transformation of values like,

- ROUND (SUM (CALL_DURATION)*60, 2) sum of call duration converted to minutes by multiplying 60 , and rounded to 2 decimal fraction.
- ROUND (SUM (CALL_FEE) * 100,2) sum of call fee in cents converted to birr by multiplying 100 , and rounded to 2 decimal fraction.

4.2.5 Clustering

K-means clustering is the most commonly used unsupervised machine learning algorithm for segmenting or grouping a given data. It involves assigning examples to groups in an effort to minimize the variance within each group. In this study, we used k-means algorithm for clustering CDR data. Although we were working clustering (unsupervised machine learning), we didn't expect that an algorithm magically clusters the input CDR data into some number of groups. We need to state how many clusters we want. Based on domain knowledge one can easily specify the appropriate number of clusters. Unfortunately, there is no universal theoretical solution to find the optimal number of clusters for any given data. A simple and common approach is to compare the results of multiple executions with different k classes and choose the best one according to a given characteristics and type of the data. The cost function that we used to determine how effective the clustering is in K-Means is distortion cost function that is the average distance of a data-point to the cluster-centroid to which it is allocated from a data point.

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K) = \frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

We used python and run K-means several times with different value of 'K' (i.e. first with only two cluster centroid, then three, etc.). For each experiment, we collect the output of the cost function and discussed in next chapter.

Algorithm 4.4: *Generic steps involved in Clustering Ethiotelecom customer data usage records*

Here *CDR, ColInt, TColInti,* Tdata, clust_labels stands for Customer data record, column of interest (selected attribute), Transformed column of interest and cluster labels assigned for each record in the sample.

```
Start
     Read CDR data
     Preprocess all Columns ()
            Result, preprocessed CDRsample data
     Column of interest (ColInt) = null
    repeat
     input ← Columns of Interest Col_i ,  ColInti ← null
     ColInti ← DataFrame input()
     CDRsample = CDRsample - ColInt
     TColInti ← fit_transform(ColInti)
     ColInt ← TColInti + ColInt
  until (ColInt in CDRsample (ColInt)!= null)
     Tdata = ColInt
     model = KMeans(kcluster)
     model.fit(Tdata)
     clust_labels = model.predict(Tdata)
Stop
output
   ClabeledCDR (cluster labels of CDR data)
```

4.3 Summary

In this chapter design of Design of Unsupervised Clustering model for telecom customers was presented. We described the basic design criteria and elements of the mode for customer data record usage clustering. The model consists of CDR data preprocessing module (cleaning, attribute selection and Aggregation) that is essential for making data ready for clustering. Sampling and Feature scaling components are basic components to take sample from preprocessed data and transform it to the compatible form of algorithm. During this study, out of different alternatives, the selected methodology has been discussed. The approaches that we described illustrate how the data is ready for clustering and how it is used to cluster customers based on their usage behavior. We have described the generic algorithm steps for our designed model. Sample examples in some of the sections have also been discussed.

Chapter 5: Experimentation and Evaluation

5.1 Introduction

This chapter describes the experimentation activities, procedures and evaluation of the results using evaluation methods. Mainly the descriptions of customer data record sample collected, the attribute features selected and the unsupervised clustering results of the experiment are discussed in this chapter. The comparative analysis of the clustering results using different cluster size are discussed. The graphical visualizations, snapshots of the experimentation procedures and evaluations of the results are explained.

5.2 Experimental Procedures

The following subsections discuss the activities and steps to evaluate unsupervised telecom customer data record clustering using usage behavior of the customers for managing and good service deliver.

5.2.1 Data Collection and Preprocessing

The experimentation of unsupervised customer usage data record clustering begins with collection of CDR from data sources. As we have mentioned ethiotelecom data in different examples of chapter four, in this work ethiotelecom CDR data is used for experimentation. The CDR data is produced by telecommunication equipment's on call basis and contain all the information to describe the important usage activities of telephone and other transaction be it from or to a subscriber in the network. It contains the fields necessary for billing systems to describe particular call and bill information's of the subscriber. CDRs give the details of information of each subscriber's usage in terms of voice, SMS and Internet data usage, originating from and terminating on subscriber's device. On the bases of this we have collected CDR data from ethiotelecom and sampled 100,000 records.

Ethiotelecom CDR Data

A complete copy of all ethiotelecom CDR data is stored in ethiotelecom company database, with a number database tables having many transactions. These are not provided for users or researchers freely. It can be provided by considering security issues for researchers but it had not been provided. We have collected ethiotelecom database dump available by passing all biro-cracy there in the will of officers. It consists of tables with a number of attributes that have values and some with null values. These data are cleaned by preprocessing steps and aggregated. We used PostgreSQL database that is object-relational

database management system to store and aggregate the backup dump we got from ethiotelecom. For the CDR cluster analysis, one month of CDR data were collected in ethiotelecom to experiment how customers can be grouped for further decision making. Data was taken from ethiotelecom customers in which privacy issues were concerned. The collected CDR was generated for billing purpose. Sampling is an important step in data analysis tasks, and is often used in handling problems with large data. Considering all the collected row data is challenging in data analysis.

5.2.2 Sampling Ethiotelecom CDR Data

This technique is the most straightforward of all the sampling methods, since it only involves a random selection or randomization and any work performed on this sample should have high internal and external validity. The steps followed to perform simple random sampling are:

1) We start by deciding the daily CDR records that we want to experiment. We ensured that we have access to every CDR data record, so that we can collect sample data from all those.
2) Next, we decided the sample size. Although larger samples provide more statistical certainty, they also cost more and require far more work.
3) Finally we applied lottery or random number method and collected 100,000 CDR daily record samples for experimentation. In the lottery method, we choose the sample at random by using a python program that will simulate the same action.

Sampled data frame information

Visualizing the aggregate summery is needed when doing evaluation of data analysis. To get a quick overview of the sampled dataset we used function that show the non-null counts. When we look the summery of sampled CDR data using pandas module function the following results.

```
Data columns (total 10 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   START_TIME           99999 non-null  object
 1   END_TIME             99999 non-null  object
 2   DURATION             95329 non-null  float64
 3   CALLED_COUNTRY       98172 non-null  float64
 4   RECIVING_FREQUENCY   99999 non-null  int64
 5   TOTAL_OUT            99999 non-null  int64
 6   TOTAL_DIS_OUT        99999 non-null  int64
 7   TOTAL_DURATION       99999 non-null  int64
 8   TOTAL_IN             99999 non-null  int64
 9   TOTAL_DIS_IN         99999 non-null  int64
dtypes: float64(2), int64(6), object(2)
memory usage: 7.6+ MB
```

Figure 5.1 summery of sampled CDR data values

As shown in the image above there are 10 attributes in sampled CDR data and showing the number of non-null value counts in each column. The data type of each column are also shown that the data type for date is object type and for others it is numeric type.

Null values or missing data can occur when no information is provided for one or more items as we see from the above sampled data. Missing Data is a very big problem in real life scenario. In order to fill null values in a sampled dataset, we used the function that replace null values with rounded mean values.

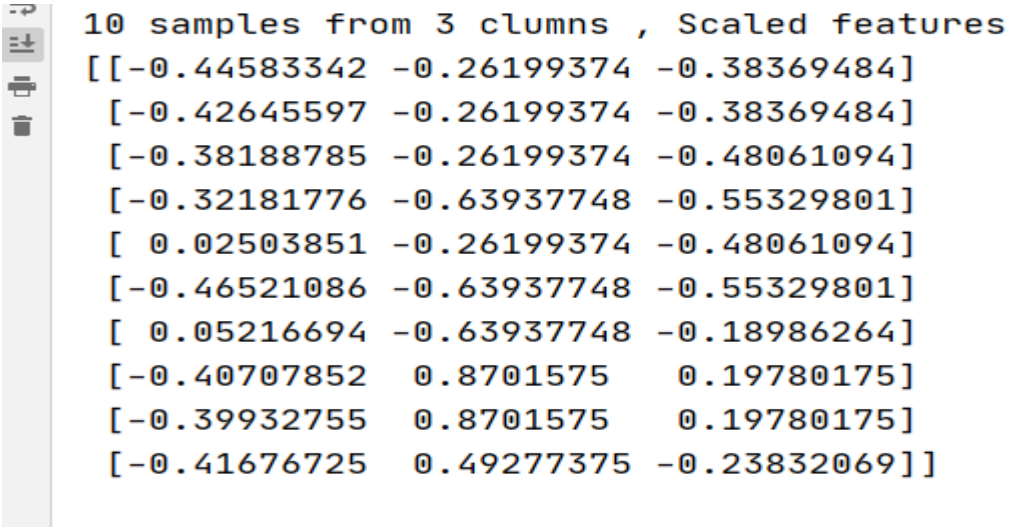
```
sampledDataframe (df) = round(df.fillna(df.mean()),2)
```

5.2.3 Feature Selection and Scaling

For the experimentation, 10 numerical features which are about total number of SMS in and out activity, number of call in and out activity, Internet usage activity, msisdn of the user, and forwarding information about the call status either forwarded or not are selected. Basically, in this work we used voice duration, data usage, SMS frequency, number of callings in to our out of the phone attributes for experimentation.

Most machine learning algorithm consider weight more significant than height only because the values for weight are larger and have higher variability from data to data. The data analysis algorithms need to consider all features on an even operating interval.

We have used the StandardScaler class to transform our sampled CDR data in a way that the clustering algorithm can fit. This class implements a type of feature scaling called standardization. Standardization scales, or shifts, the values for each numerical feature in your dataset so that the features have a mean of 0 and standard deviation of 1. The image below shows how the values have been scaled in scaled features.



```
10 samples from 3 clumns , Scaled features
[[-0.44583342 -0.26199374 -0.38369484]
 [-0.42645597 -0.26199374 -0.38369484]
 [-0.38188785 -0.26199374 -0.48061094]
 [-0.32181776 -0.63937748 -0.55329801]
 [ 0.02503851 -0.26199374 -0.48061094]
 [-0.46521086 -0.63937748 -0.55329801]
 [ 0.05216694 -0.63937748 -0.18986264]
 [-0.40707852  0.8701575   0.19780175]
 [-0.39932755  0.8701575   0.19780175]
 [-0.41676725  0.49277375 -0.23832069]]
```

Figure 5.2 CDR sample values scaled in scaled features

After feature scaling CDR data was ready to be clustered. The K-Means estimator class in scikit-learn was used where we set the algorithm parameters before fitting the estimator to the data. The scikit-learn implementation is flexible, providing several parameters that can be modified during experimentation process.

5.2.4 Feature Value Distributions

When we see the sampled dataset attribute characteristics based on the value distributions by using pie chart by taking one sample attribute duration of customers we have noticed the following.

In order to visualize the value distributions in sampled data we tried to see it in two portions. We used the mean value to divide the duration values in to two and check how much percent of data value of the duration is less than the mean and how much is greater than the mean value. We collected those values in two variables as the values with higher value than the mean (high usage values) and values with low value than the mean (low usage values). We visualized graphically using python pie chart as the following.

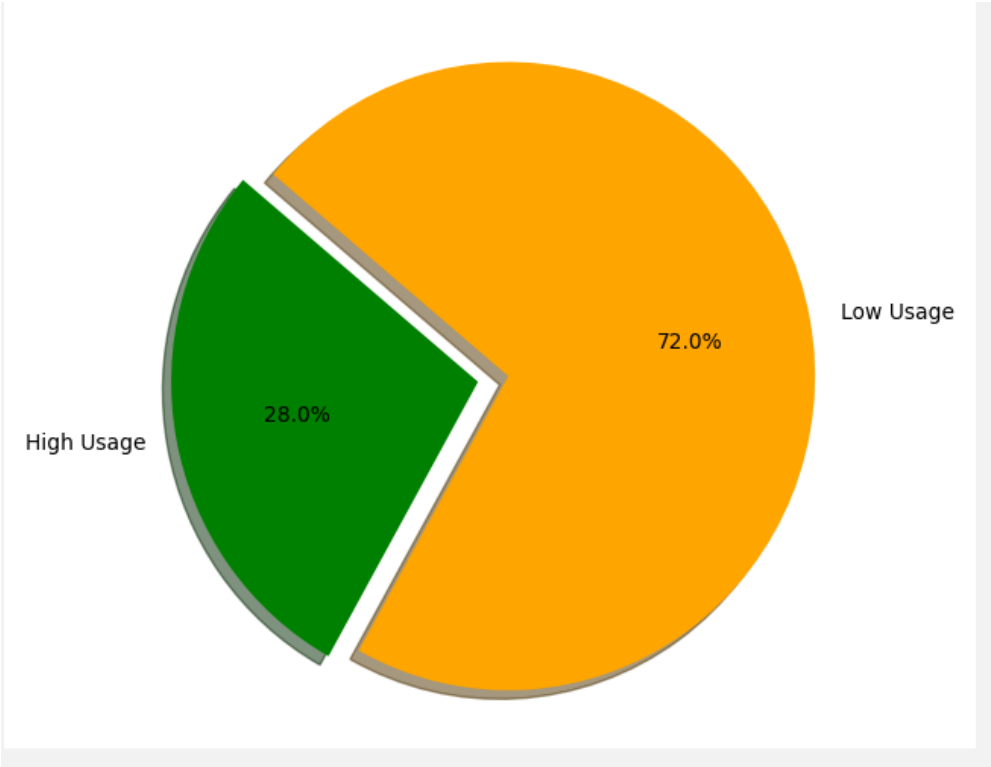


Figure 5.0.3 value distributions in sampled data

As we see from the figure the duration attribute value distribution in the sample is with the 72% values less than the mean of the attribute values and with the 28% values greater than the mean value of the attribute. With this we checked that the sampled data is inclusive that consider most Ethiopian users which have low usage characteristics.

5.2.5 Prototype Development

Tools and **Programming Languages**

In this experimentation, we have used Python and PYcharm IDE programming tools. PYcharm IDE is a free and open source distribution of the Python programming language for data science and machine learning related applications that aims to simplify package management and deployment. Python is a high-level language that is easy to learn, read, maintain and experiment because of:

- Short development time in comparison to other programming languages like Java, C++.
- There are plenty of libraries in python that makes our task easier, for example, NumPy is a library for python that can solve scientific computation easily.

Pandas: an python open-source library is used to read CSV files and perform different operations on the CSV files.

Scikit-Learn: Scikit-learn is a library for python machine learning library. Scikit-learn contains simple and efficient tools for data mining and data analysis algorithms for both supervised and unsupervised problems. Scikit-learn is used to evaluate deep neural classifiers by calculating a confusion matrix which is a table that is often used to describe the performance of model.

We have preprocessed CDR data using Pandas which is open-source csv data processing toolkit implemented in Python.

Experimental Setting

In this study we used a computer with a memory capacity of 8 GB RAM, 2.2 GHZ processor and 64 bit windows 10 operating system.

5.2.6 Applying K-means Clustering

The k-means clustering method is an unsupervised machine learning technique used to identify group of data in a dataset. There are many different types of clustering methods, but k-means is one of the o most approachable. These reasons make applying k-means clustering in Python is needed. This activity is grouping of related users using the weighted features of CDR data. Ethiotelecom user are clustered using k-means clustering algorithm in which all usage behaviors are normalized and cosine similarity measure was applied. Conventional k-means requires only a few steps.

During our experimentation process there were parameters the following parmeters were used:

- **init** controls the initialization system. We used the normal version of the k-means algorithm that was implemented by setting init to "random".
- **n_clusters** sets k for the clustering step. This is the most important parameter to define the number of groups for k-means.
- **n_init** sets the quantity of instatements to perform. This is significant in light of the fact that two runs can join on various group tasks. The default conduct for the scikit-learn calculation is to perform ten k-implies runs and return the aftereffects of the one with the most reduced mistake number of squares (SSE).

- **max_iter** sets the number of maximum iterations for each initialization of the k-means algorithm.

For example we initialized the algorithm with the following arguments:

```
kmeans = KMeans(
        init="random",
        n_clusters=4,
        n_init=10,
        max_iter=100,
        random_state=42
    )
```

The parameter depend on the experimentation task, data, module and language that we were used to apply the *k*-means algorithm. In this step we applied the algorithm on the `scaled CDR feature`.

5.2.7 Clustering Results

The final activity is grouping of related customers using the scaled features of the CDR data. Customers are clustered using k-means clustering algorithm in which all features are normalized and cosine similarity measure is applied. The following Figure shows the simple snapshot obtained during experimentation indicates the customer id with assigned cluster id.

```
prediction ID - class values
 0      3
1       3
2       3
3       3
4       3
      ..
95      3
96      3
97      3
98      2
99      0
Name: cluster_pred, Length: 100, dtype: int32
```

```
prediction ID - class values
 99899    2
99900     3
99901     3
99902     3
99903     3
         ..
99994     3
99995     3
99996     3
99997     0
99998     3
```

Figure 5.4 Experimentation result customer id with assigned cluster id

The result of clustering was recorded, and the distributions of each customer in a cluster can be visualized using clustering id assigned or by plotting the clustered data. The above distribution of customers into different category by assigned cluster id shows that different classes have different usage characteristics.

Table 5.1 number of Customers in each class

Cluster ID	Number of customers in the class
0	19959
1	603
2	4383
3	75054

With two features here ‘Total number calls and ‘SMS frequency we plotted in x y plane to see four different classes. We can easily see the four clusters with different colors as shown on the following figure.

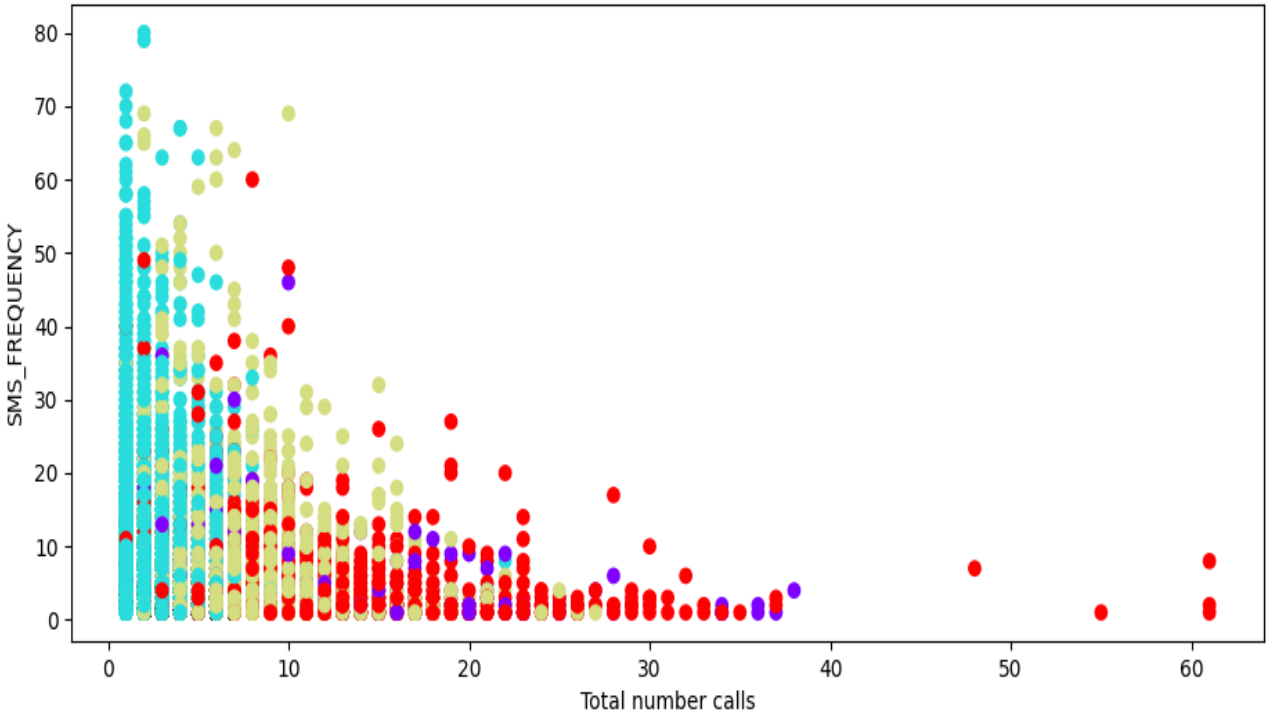


Figure 5.5 Total number calls vs SMS frequency clsuter distributions

Clustering is a set of techniques used to partition data into groups, or clusters. Thus useful clusters, serve as an intermediate step in ethiotelecom business management.

Through the figure 5.5 we can provide the following interpretation:

- The ***DeepSkyBlue*** dots are the customers who are less call history and high SMS usage character therefore can be termed as frequent SMS users.
- The ***red dots*** are people with high calling and less SMS usage behavior therefore can be termed as frequent calling service users.
- The ***Olive dots*** are the customers with average Call and SMS usage behavior can be termed as good users of both services.
- The ***sky-blue*** dots are the customers with small in number who are in the midst of things.

The final goal of any business would be to have as many customers up there in the service category. As we are ready with a new decision and we can target the group of customers as per our analysis.

5.3 Evaluating of Optimal K Number of Clusters

There are two methods that are commonly used to evaluate the appropriate number of clusters:

1. The elbow method
2. The silhouette coefficient

5.3.1 The elbow method

The elbow method is commonly used to determine the optimal number of grouping in k-means clustering. The elbow technique plots the worth of the value perform created by totally different values of k. If k will increase, average distortion can decrease, every cluster can have fewer constituent instances, and therefore the instances are going to be nearer to their various centroids. the worth of k at that improvement in distortion declines the foremost is known as because the elbow, at which we could use it for further clusters.

To apply elbow method, we have done several k-means, by incrementing k with each iteration, and recorded the error sum of squares (SSE).

Figure 5.6 The Elbow method to evaualte number of clusters

With elbow method for selection of optimal k cluster we visualized and we had taken to divide or group our data based on usage behavior in to four groups by considering that the elbow is on that cluster.

5.3.2 Silhouette Coefficient

The silhouette coefficient is a measure of cluster cohesion and separation. It quantifies how well a data point fits into its assigned cluster based on factors how close the customer data is to other customer data in the group and how far away the customer data is from customer data in other clusters.

Silhouette coefficient values range between - 1 and 1. Bigger numbers demonstrate that examples are closer to their groups than they are to another bunch. The Silhouette Coefficient for an example is:

```
(b - a) / max (a, b)
```

To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of. Based on this computation it returns the mean Silhouette Coefficient over all samples. To obtain the values for each sample by using scikit-learn module in python.

Table 5.2 Silhouette Coefficient values for K 1- 10

Number of Clusters	Silhouette Coefficient
2	0.4251035963805581
3	0.4246684377100701
4	0.49507900615748847
5	0.47060420009293136
6	0.4821944044673164
7	0.4848224794301356
8	0.4847450772576435
9	0.4777786892066459
10	0.45095514541607357

The values are interpretation and validation of consistency within clusters of data. The average silhouette scores in the table for each k shows that the best choice for k is 4 since it has the maximum score that shows validation of consistency within clusters.

5.4 Discussions

Out of 100,000 samples taken from telecom CDR weekly records that are summarized based on the type of value store in the attributes used for testing, the data is categorized into four clusters. Those tuples with null values was not selected in the sampling stage. We have analyzed that these sampled customer records contain values of different type of telecom user starting from vary low usage history to great customers with high usage history. During data transformation process, the values for features are transformed to the same scale to consider all features on an even operating interval. This process of transforming numerical features to use the same scale is feature scaling. Those scaled features are used in k-means clustering algorithm to cluster the sampled data.

Totally 100,000 sampled CDR data are clustered to their respective categories based on the usage history values. To find the optimal number of groups for our sampled data we used the common evaluation methods elbow method and silhouette coefficient. As shown in plotted chart for elbow method, the optimal turning point elbow can be considered as class number four and we considered k to be four. In silhouette coefficient method that is a metric used to calculate the goodness of a clustering technique, value ranges from -1 to 1. Even though it takes much time, we recorded the mean silhouette coefficient over all samples. The Larger numbers indicate that samples are nearer to their clusters than they are to other cluster (high intra-cluster similarity). As shown in the table the very significant difference in value is seen clearly and the value of silhouette coefficient is greater when cluster size is four with this we clustered the sampled data into four categories. The cluster identification numbers are assigned for each sample. When we see the number of customer's distribution in each class, in the first class there are 19,959 customers with related usage history. 19.959 % of the sample was grouped in one. In the second class there are 603 customers with related usage characteristics. I.e. 0.603% of the sample was grouped together. In the 3rd class there are 4383 customers with related usage characteristics. I.e. 4.383% of the sample was grouped together. In the 4th class there are 75,054 customers with related usage characteristics. I.e. 75.054% of the sample was grouped together.

During this study, we have tried to see the sample relationships between cluster usage history and user's summarized attribute values for each class of customer using graphs. As we have tried to show in the figure 5.5 using SMS usage frequency and the total number of calls. The group can be considered as customers with low usage history of the service,

common usage history of the service, customers with very good usage history of the service and customers with very midst history. This can be visualized for each attribute in each class that are used in testing process. The ultimate goal of any business would be to have as many customers up there in the service category. As we are ready with a new decisions and we can target the group of customers as per our analysis. This analysis can be further used in decision making to have many number of customers to our needed group i.e. a group with very good usage behavior of services.

Therefore, having clustering model for telecom CDR data can be used in marketing management which is one of the important aspects in business. In this work, a model is proposed for clustering telecom customers into four categories based on usage behavior of customers. A class label for each customer is determined based upon summarized CDR data and a clustering model is built for predicting different class of customers. The clustering model was verified with CDR test data. During clustering extracted information is helpful for planning for new customer services and providing personalized and grouped customer services by understanding their usage behavior.

Chapter 6 : Conclusion and Recommendation

6.1 Conclusion

Recent advancement in mobile phone and mobile networking, data processing, and related technologies have significantly eased the process of generating and collecting large amounts of CDR data. Currently ethiotelecom [16] has 48.9M Mobile Service Subscribers, 23.5M Data and Internet users, 309.4K Fixed Broadband subscribers, 981K Fixed Service Subscriber and 50.7M Total Customer as of January 2021. CDR offers key metadata on how and when to use your business phone system. Moreover, a call detail record (CDR) provides information about calls made over a phone service. Wide collections of anonymized CDRs may extract surprisingly useful knowledge about human actions. CDR allows phone companies to generate your phone bills, and lets you keep definite records of how and when your phone system was used. Computer science agencies also use CDRs to assess telephone service. There are many works done in the area of segmenting mobile customers in different researchers' cluster customers based on different attributes like call detail records, customer profiling. customer behavior and others with purpose of marketing analysis, customer management, for improving profitability of one company and used different data analysis techniques for different mobile operators like china, Pakistan, Korea, India [10,33].

This research work had attempted to look into the techniques of unsupervised customer clustering by using telecom customer detail records (CDR). Throughout this study the basic elements of the model for customer clustering are presented. The designed model consists of preprocessing module that is essential for better data preparation. Feature scaling transforms features relevant to the clustering algorithm, so as to reduce the dimensionality and to fit for the model. This process of transforming numerical features to use the same scale is feature scaling. Those scaled features are used in k-means clustering algorithm to cluster the sampled data.

Total of 100,000 sampled CDR data from telecom are used in experimentation and clustered to their respective categories based on the usage history values. To find the optimal number of groups for our sampled data we used the common evaluation methods elbow method and silhouette coefficient. As shown in plotted chart for elbow method, the optimal turning point elbow can be considered as class number four and we considered k to be four. In silhouette coefficient method that is a metric used to calculate the goodness

of a clustering technique, value ranges from -1 to 1. Even though it takes much time, we recorded the mean silhouette coefficient over all samples. The Larger numbers indicate that samples are nearer to their clusters than they are to other cluster (high intra-cluster similarity). As shown in the table the very significant difference in value is seen clearly and the value of silhouette coefficient is greater when cluster size is four with this we clustered the sampled data into four categories. The cluster identification numbers are assigned for each sample. When we see the number of customer's distribution in each class, in the first class there are 19,959 customers with related usage history. 19.959 % of the sample was grouped in one. In the second class there are 603 customers with related usage characteristics. I.e., 0.603% of the sample was grouped together. In the 3rd class there are 4383 customers with related usage characteristics. I.e., 4.383% of the sample was grouped together. In the 4th class there are 75,054 customers with related usage characteristics. I.e. 75.054% of the sample was grouped together. During this study, we have tried to see the sample relationships between cluster usage history and user's summarized attribute values for each class of customer using graphs as we have tried to show in the above figure using SMS usage frequency and the total number of calls. The group can be considered as customers with low usage history, normal or common usage history and customers with very good usage history. This can be visualized for each attribute in each class that are used in testing process. The ultimate goal of any business would be to have as many customers up there in the service category. As we are ready with a new decision and we can target the group of customers as per our analysis. This analysis can be further used in decision making to have many numbers of customers to our needed group i.e., a group with very good usage behavior of services.

Therefore, having clustering model for telecom CDR data can be used in marketing management which is one of the important aspects in business. It is vital for the company to adopt different methodologies by which high valued customers can be identified, in order to perform suitable target marketing effectively. In this work, a model is proposed for clustering telecom customers into four categories based on usage behavior of customers. A class label for each customer is determined based upon summarized CDR data and a clustering model is built for predicting different class of customers. The clustering model was verified with CDR test data. During clustering extracted information is helpful for planning for new customer services and providing personalized and grouped customer

services by understanding their usage behavior. Furthermore, having more trained model with more attributes, then the performance of the system will be improved significantly.

6.2 Contribution of the Study

The main contributions of the study are listed below:

- A model is designed for unsupervised telecom customer clustering that takes advantage of customer detail records.
- The customer detail record data was collected from the company that can be used for further works related to customer detail analysis.
- This study showed the optimal number of clusters using the elbow and Silhouette coefficient method for clustering telecom customers.
- This study contributes the algorithm for clustering telecom customers using their customer detail records.
- This study showed the attribute relationships after clustered based on their distribution during clustering.
- This study showed the association between the cluster size and the evaluation methods in customer clustering model.

In addition, the study contributes to the growth of customer detail record data analysis for business management and decision making.

6.3 Recommendations

The designed system i.e. unsupervised clustering model using telecom customer detail records attempted grouping of customers based on network usage behaviors. However, it is also learnt that further research and developmental effort is needed so as to enable CDR data analysis and organization more accurate. Furthermore, there are some components that should be added and integrated for better performance of the system. Some of the future research issues and features that needed to provide a better result include:

- Using subscriber information and customer detail records (CDR) for different data analysis tasks is believed to result in a significant improvement of the task. In this study we tried only customer clustering using CDR. However, it is also believed to result a significant improvement in time serious activity analysis and other customer data analysis tasks.
- It is interesting to validate the effectiveness of using subscriber information and customer detail records (CDR) that offers information related to customers with different attributes would be essential. It has additional potential that would be used with scaled features. Having subscriber information without violating privacy with its CDR data is believed to result a significant improvement of the clustering process.

References

[1] Krishan Kumar et al, "Vulnerability Detection of International Mobile Equipment Identity Number of Smartphone and Automated Reporting of Changed IMEI Number," *Journal of Computer Science and Information Technology*, 2015.

[2] Feven Fesseha and Mesfin Kifle, " Predictive SIM Box Fraud Detection Model for ethio telecom", *Journal of Computer Science and Technology*, December 2017.

[3] Internet Usage World Stats- Internet and Population Statistics 2019, http://www.internetworldstats.com/, last Acc. on May 10, 2019.

[4] Qining LIN and Yan WAN, "Mobile Customer Clustering Based On Call Detail Records For Marketing Campaigns", *International Conference on Management and Service Science* (IEEE), 2009.

[5] Parre amarnath1 and M. chandini, "A Two-Layer Clustering Model for Mobile Customer Analysis", *International journal of Scientific Engineering and Technology research*, May 2018.

[6] So Young Sohn and Yoonseong Kim, "Searching customer patterns of mobile service using clustering and quantitative association rule", *ScienceDirect Expert Systems with Applications*, 2008.

[7] Shohin Aheleroff and Gholamian, "Customer Segmentation for a Mobile Telecommunications Company Based on Service Usage Behavior", *The 3rd International Conference on Data Mining and Intelligent Information Technology Applications*. October 2011.

[8] Lulu Deyu, "data mining approach to analyze mobile telecommunications network quality of service: the case of ethio-telecom", *Master thesis in Addis Ababa University*, May 2014.

[9] Dr. Rajan Vohra, "Segmentation of Mobile Customers for Improving Profitability Using Data Mining Techniques", *International Journal of Computer Science and Information Technologies*, 2014,

[10] Y.Gopi, "Tele Comm. Customer Data Analysis using Multi-Layer Clustering Model", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2018.

[11] Lulu Deyu, "data mining approach to analyze mobile telecommunications network quality of service: the case of ethio-telecom", *Master thesis in Addis Ababa University*, May 2014.

[12] S.M.H. Jansen, |Customer Segmentation and Customer Pro-filing for a Mobile Telecommunications Company Based on Usage Volume 3, Issue 1, January-February-2018 | www.ijsrcseit.com | UGC Approved Journal 311 Behavior: A Vodafone Case Study, 17 July 2007.

[13] PennState Eberly Collage of Science Online courses, retrieved from https://onlinecourses.science.psu.edu/stat505, Last accessed on, January 1, 2018.

[14] Internet usage Mobile networks - Wiki https://computersciencewiki.org/index.php/Mobile_networks last Acc. on Feb 10, 2020

[15] Arasan, Ezhil. A Review on mobile technologies: 3G, 4G and 5G. Second International Conference on Recent Trends and Challenges in Computational Models, Tindivanam, (2017).

[16] Internet usage Ethio telecom 2013 EFY (2020/21) First Half Business Performance Summary Report https://www.ethiotelecom.et / last Acc. on Feb 10, 2021.

[17] Giridhar Maji1, Soumya Sen2, Data Warehouse Based Analysis on CDR to Retain and Acquire Customers by Targeted Marketing, *Institute of Electrical and Electronics Engineers,* September 2016.

[18] Annett Saarik, Trajectory Reconstruction and Mobility Pattern Analysis Based on Call Detail Record Data, Masters thesis, 2017.

[19] Zhang Tianyuan " telecom customer segmentation and precise package design by using data mining" Master thesis in ISCTE Bbusiness school, October, 2018.

[20] Anna Huang, David Milne, Eibe Frank and Ian H.Witten, "Clustering Documents with Active Learning using Wikipedia", *Eighth IEEE International Conference on Data Mining,* February 10, 2009.

[21] JunkaiYi, Yacong Zhang, Xianghui Zhao and Jing Wan, "A Novel Clustering Approach Using Deep-Learning Vocabulary Network", *Hindawi Publishing Corporation,* 15 March 2017.

[22] Wael H. Gomaa and Aly A. Fahmy, "A Survey of Text Similarity Approaches", *International Journal of Computer Application*s, April 2013.

[23] N. Sandhya, Y.Sri Lalitha, A.Govardhan and K.Anuradha, "Analysis of Similarity Measures for Text Clustering". *Semantic scholar*, 2013.

[24] Michael Steinbach, George Karypis and Vipin Kumar, "A Comparison of Document Clustering Techniques", *Proceedings of the International KDD Workshop on Text Mining,* June 2000.

[25] Fidan Kaya Gülağız and Suhap Şahin, “Comparison of Hierarchical and Non Hierarchical Clustering Algorithms”, *International Journal of Computer Engineering and Information Technology*, January 2017.

[26] Data Science, K-means Clustering ,retrieved from *https://www.datascience.com* , Last accessed on, January 2018.

[27] Shi Zhong, “Efficient Online Spherical K-means Clustering”, *IEEE International Joint Conference on Neural Networks,* August 2005.

[28] Shenghong Yang and Yongheng Wang, “Density-Based Clustering of Massive Short Messages using Domain Ontology”, in *Asia-Pacific Conference on Information Processing,* May 2009.

[29] Stuti Karol and Veenu Mangat, “Evaluation of Text Document Clustering Approach Based on Particle Swarm Optimization”, *Central European Journal of Computer Science,* September 2012.

[30] Towards data science, Word Embeddings, Choosing the Optimal Number of Clusters, retrieved from *https://towardsdatascience.com/,* Last accessed on, *January 21, 2020.*

[31] Lailil Muflikhah and Baharum Baharudin, “Document Clustering Using Concept Space and Cosine Similarity Measurement”, *International Conference on Computer Technology and Development,* 2009.

[32] Masood, Salar & Moaz, Ali & Arshad, Faryal & Qamar, Ali & Kamal, Aatif & Rehman, Ahsan. Customer Segmentation and Analysis of a Mobile Telecommunication Company of Pakistan using Two Phase Clustering algorithm, IEEE International Conference on Digital Information Management (ICDIM), Pakistan, 2013.

[33] Kim, Su-Yeon & Jung, Tae-Soo & Suh, Euiho & Hwang, Hyunseok., "Customer Segmentation and Strategy Development Based on Customer Lifetime Value: A Case Study," Expert Systems with Applications, vol. 31, 2006, pp. 101-107, July 2006.

[34] Ding, Z. & Jia, Y. & Zhou, B. Survey of data mining for microblogs. Jisuanji Yanjiu yu Fazhan/ Computer Research and Development, April 2014.

[35] Hamka, Fadly & Bouwman, Harry & de Reuver, Mark & Kroesen, Maarten. Mobile customer segmentation based on smartphone measurement. Telematics and Informatics, May 2014.

[36] Kim, Hee-Su & Yoon, Choong-Han. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications Policy, October 2004.

[37] Maha Alkhayrat, Mohamad Aljnidi and Kadan Aljoumaa, A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA, Springer, 2020.

Annex A: snapshot of sampled data

	A	B	C	D	E	F	G	H	I	J	K	L	M
61	2.52E+11	1	1	1	1	100	238	1	1	2	2	N	2
62	2.52E+11	1	1	1	1	20	240	1	1	0	0	N	2
63	2.52E+11	2	1	2	2	60	120	0.5	1	1	1	N	3
64	2.52E+11	2	1	2	2	60	120	0.5	1	1	1	N	1
65	2.52E+11	1	1	1	1	100	238	1	1	2	2	N	3
66	2.52E+11	2	1	2	2	100	119	0.5	1	0	0	N	1
67	2.52E+11	2	1	2	2	100	119	0.5	1	3	1	N	2
68	2.52E+11	1	1	1	1	20	240	1	1	2	1	N	1
69	2.52E+11	1	1	1	1	20	240	1	1	1	1	N	1
70	2.52E+11	2	2	2	2	120	119	1	1	1	1	N	2
71	2.52E+11	2	1	2	2	100	119	0.5	1	1	1	N	5
72	2.52E+11	1	1	1	1	100	238	1	1	2	2	N	3
73	2.52E+11	3	2	3	3	200	79	0.67	1	1	1	N	2
74	2.52E+11	1	1	1	1	20	240	1	1	1	1	N	1
75	2.52E+11	1	1	1	1	90	239	1	1	3	1	N	8
76	2.52E+11	1	1	1	1	90	239	1	1	1	1	N	16
77	2.52E+11	21	1	1	1	30	240	1	1	1	1	N	3
78	2.52E+11	1	1	1	1	40	239	1	1	0	0	N	6
79	2.52E+11	1	1	1	1	40	239	1	1	0	0	N	4
80	2.52E+11	1	1	1	1	40	239	1	1	0	0	N	2
81	2.52E+11	1	1	1	1	200	237	1	1	0	0	N	1
82	2.52E+11	1	1	1	1	20	240	1	1	1	1	N	3
83	2.52E+11	12	1	1	1	20	240	1	1	1	1	N	3

	MSISDN	TOTAL_OU	TOTAL_DI	TOTAL_CE	TOTAL_DI	TOTAL_DU	CALL_GAP	RATIO_DI	RATIO_DI	TOTAL_IN	TOTAL_DI	F_STATUS	RECIVING_FREQU
	2.52E+11	2	1	2	2	100	119	0.5	1	0	0	N	1
	2.52E+11	2	1	2	2	100	119	0.5	1	0	0	N	1
	2.52E+11	2	1	2	2	60	120	0.5	1	0	0	N	1
	2.52E+11	12	1	1	1	30	240	1	1	0	0	N	1
	2.52E+11	2	1	2	2	60	120	0.5	1	0	0	N	1
	2.52E+11	1	1	1	1	30	240	1	1	0	0	N	2
	2.52E+11	11	1	1	1	180	237	1	1	0	0	N	1
	2.52E+11	5	4	4	1	340	47	0.8	0.2	0	0	N	2
0	2.52E+11	5	4	4	1	340	47	0.8	0.2	0	0	N	2
1	2.52E+11	4	4	3	1	160	59	1	0.25	0	0	N	1
2	2.52E+11	5	4	4	1	340	47	0.8	0.2	0	0	N	5
3	2.52E+11	11	2	11	4	320	21	0.18	0.36	4	1	N	1
4	2.52E+11	11	4	11	3	320	21	0.36	0.27	4	1	N	2
5	2.52E+11	2	2	2	1	60	120	1	0.5	1	1	N	7
5	2.52E+11	9	4	9	3	280	26	0.44	0.33	4	1	N	1
7	2.52E+11	5	3	5	2	120	48	0.6	0.4	4	2	N	1
8	2.52E+11	12	2	12	7	400	19	0.17	0.58	1	1	N	6
9	2.52E+11	4	1	4	1	80	60	0.25	0.25	2	1	N	3
0	2.52E+11	4	1	4	1	80	60	0.25	0.25	2	1	N	1
1	2.52E+11	2	1	2	1	60	120	0.5	0.5	0	0	N	7
2	2.52E+11	4	1	4	1	80	60	0.25	0.25	2	1	N	6
3	2.52E+11	4	1	4	1	80	60	0.25	0.25	2	1	N	3

samples

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
99981	2.52E+11	3	3	3	2	300	78	1	0.67	0	0	Y	1		
99982	2.52E+11	6	4	6	2	620	38	0.67	0.33	4	3	Y	2		
99983	2.52E+11	2	1	2	2	480	116	0.5	1	0	0	Y	2		
99984	2.52E+11	2	1	2	1	760	114	0.5	0.5	0	0	Y	1		
99985	2.52E+11	1	1	1	1	30	240	1	1	1	1	Y	1		
99986	2.52E+11	5	3	5	1	1020	45	0.6	0.2	1	1	Y	3		
99987	2.52E+11	2	2	2	1	240	118	1	0.5	0	0	Y	2		
99988	2.52E+11	6	2	3	3	300	39	0.33	0.5	2	2	Y	2		
99989	2.52E+11	1	1	1	1	330	235	1	1	0	0	Y	1		
99990	2.52E+11	3	3	3	1	1920	69	1	0.33	0	0	Y	4		
99991	2.52E+11	3	2	3	3	120	79	0.67	1	1	1	Y	1		
99992	2.52E+11	2	2	2	2	120	119	1	1	0	0	Y	11		
99993	2.52E+11	1	1	1	1	20	240	1	1	0	0	Y	5		
99994	2.52E+11	1	1	1	1	20	240	1	1	6	4	Y	1		
99995	2.52E+11	1	1	1	1	30	240	1	1	0	0	Y	2		
99996	2.52E+11	3	1	3	1	150	79	0.33	0.33	2	1	Y	1		
99997	2.52E+11	3	2	3	1	120	79	0.67	0.33	3	1	Y	1		
99998	2.52E+11	1	1	1	1	60	239	1	1	2	1	Y	1		
99999	2.52E+11	15	15	14	1	2220	14	1	0.07	4	2	Y	2		
100000	2.52E+11	4	3	4	2	120	60	0.75	0.5	4	2	Y	1		
100001															
100002															
100003															

Annex B: Sample Codes for feature scaling and clustering

```
import pandas as pd

from sklearn.metrics import silhouette_score
from sklearn.preprocessing import StandardScaler
from sklearn import preprocessing
from sklearn.cluster import KMeans
from sklearn import metrics
from sklearn.cluster import AgglomerativeClustering


      data=dataframe[cols_of_interest]
      ss = StandardScaler()
      scaler = StandardScaler()
      scaled_features = scaler.fit_transform(data)
      ss.fit_transform(data)
      X = StandardScaler().fit_transform(data)
      kmeans = KMeans(
                  init="random",
                  n_clusters=4,
                  n_init=10,
                  max_iter=300,
                  random_state=42
        )

      print(X[:10])
      kmeans.fit(scaled_features)
      print(kmeans.inertia_)
      print(kmeans.cluster_centers_)

      clusters=data.copy()
      clusters['cluster_pred']=kmeans.fit_predict(dat
      a)
      countc1=clusters['cluster_pred'].value_counts()
      print(clusters['cluster_pred'].value_counts())
      print("prediction ID - class
      values\n",clusters['cluster_pred'].tail(100))
      x = [0,2,4,6,8,10]
      print(data['RECIVING_FREQUENCY'].max())
```

```
plt.scatter( data['TOTAL_OUT'],
data['RECIVING_FREQUENCY'], c=
clusters['cluster_pred'].astype(int),
cmap='rainbow')
plt.xlabel("Total number calls")
plt.ylabel("SMS_FREQUENCY")
plt.show()
kmeans_kv = {
        "init":"random",
        "n_init":10,
        "max_iter":300,
        "random_state":42
}
sse = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, **kmeans_kv)
    kmeans.fit(scaled_features)
    sse.append(kmeans.inertia_)


plt.style.use("fivethirtyeight")
plt.plot(range(1, 11), sse)
plt.xticks(range(1, 11))
plt.xlabel("Number of Clusters")
plt.ylabel("SSE")
plt.show()

silhouette_coefficients = []
  # start at 2 clusters for silhouette
coefficient
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, **kmeans_kv)
    kmeans.fit(scaled_features)
    score = silhouette_score(scaled_features,
kmeans.labels_)
    print(score)
    silhouette_coefficients.append(score)
```

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all source of materials used for the thesis have been duly acknowledged.

Declared by:

Name: __ Banchalem Abebaw___

Signature: _____

Date: _____

Confirmed by advisor:

Name: ___ Mesfin Kifle (PhD)_____

Signature: _____

Date: _____