T.C.

BAHÇEŞEHİR UNIVERSITY

CHURN MANAGEMENT BY USING FUZZY C-MEANS

M.S. Thesis

Evren ARİFOĞLU

İstanbul, 2011

T.C.

BAHÇEŞEHİR UNIVERSITY
The Graduate School of Natural and Applied Sciences
Computer Engineering

CHURN MANAGEMENT BY USING FUZZY C-MEANS

M.S. Thesis

Evren ARİFOĞLU

Supervisor: Assoc. Prof. Dr. Adem KARAHOCA

İstanbul, 2011

T.C
BAHÇEŞEHİR UNIVERSITY
The Graduate School of Natural and Applied Sciences
Computer Engineering

Title of the Master's Thesis : Churn Management by Using Fuzzy C-means
Name/Last Name of the Student : Evren ARİFOĞLU
Date of Thesis Defense :

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assoc. Prof. F. Tunç BOZBURA
Acting Director

This is to certify that we have read this thesis and that we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

Examining Committee Members:

Assoc. Prof. Dr. Adem KARAHOCA :

Asst. Prof. Dr. Mehmet Alper TUNGA :

Asst. Prof. Dr. Yücel Batu SALMAN :

ÖZET

FUZZY C-MEANS ALGORİTMASI KULLANILARAK MÜŞTERİ KAYIP YÖNETİMİNİN YAPILMASI

Arifoğlu, Evren

Bilgisayar Mühendisliği
Tez Danışmanı: Doç. Dr. Adem KARAHOCA

Eylül 2011, 45 sayfa

Bugünlerde, GSM (Global Service of Mobile Communication) pazarı bütün uluslarda devasa bir sektör haline gelmiştir. Ses kalitesi en önemli faktör olduğundan ve müşteriler GSM operatörlerini seçerken bu hizmete çok dikkat ettiklerinden GSM şirketleri ses kalitelerini yükseltebilmek içim 3G teknolojisini kullanmaktadırlar. Müşteri GSM operatörü seçerken etken olan başka özellikler de mevcuttur. Bu etkenler nedeni ile bir çok müşteri kullandıkları operatörleri değiştirmektedir. GSM şirketleri için müşterinin hizmetten vazgeçip, vazgeçmeyeceğini veya operatör değiştirip değiştirmeyeceğini öngörmek çok önemlidir. Bu yüzden GSM hizmeti veren şirketler herbir müşterinin anlık davranışlarını kontrol etmek ve müşterinin gelecekteki olası kararlarını tahmin etmek zorundadır. Bu çalışmada, veri madenciliği teknikleri kullanılarak bir müşterinin kullandığı operatörü değiştirip değiştirmeyeceğine dair tahminler üretmeye çalıştık. Aynı zamanda fuzzy c-means algoritması Decision Tree, Naive Bayes, Support Vector Machine ve Probabilistic Neural Network gibi algoritmalarla da karşılaştırıldı. Çalışma sonunda Fuzzy c-means algoritmasının en iyi sonucu vermesini beklemekteyiz.

Anahtar Kelimeler: Veri Madenciliği, Müşteri Kayıp Yönetimi, ANFIS, Fuzzy c-means.

ABSTRACT

CHURN MANAGEMENT BY USING FUZZY C-MEANS

Arifoğlu, Evren

Computer Engineering
Supervisor: Assoc. Prof. Dr. Adem KARAHOCA

September 2011, 45 pages

Nowadays, Global Service of Mobile Communication (GSM) market is a huge sector in nations' economies. Voice quality is an important factor for a customer to choose a GSM operator and hence GSM companies increases their voice quality via 3G technologies. Also there are other factors which affect a consumer to prefer a particular GSM operator. Due to several reasons, customers change their current GSM operators. It is very important for GSM operators to predict if a subscriber will cancel the service and switch to another GSM operator. Therefore, companies that provide GSM services have to monitor the behavior of each subscriber and predict one step ahead. In this study, using fuzzy c-means algorithm, we aim to predict whether a subscriber will change her current GSM operator or not. We also compare fuzzy c-means algorithm with Decision Tree, Naïve Bayes and Support Vector Machine and Probabilistic Neural Network. At the end of this study we expect that fuzzy c-means will give best result.

Keywords: Data Mining, Churn Management, ANFIS, Fuzzy c-means.

TABLE OF CONTENTS

LIST OF TABLES v
LIST OF FIGURES vi

1. INTRODUCTION 1

 1.1 PROBLEM DEFINITION 1
 1.2 CHURN MANAGEMENT 3
 1.3 DATA MINING 3

2. LITERATURE SURVEY 5

3. MATERIALS & METHODS 10

 3.1 MATERIALS 10
  3.1.1 Program 10
  3.1.2 Data 10
   3.1.2.1 Data Preperation 11
   3.1.2.2 Data Understanding 12

 3.2 METHODS 14
  3.2.1 C-means Algorithm 14
  3.2.2 Fuzzy C-means Algorithm 14
  3.2.3 ANFIS 16

4. FINDINGS 17
 4.1 DATA ANALYSING PHASE 17
 4.2 CLUSTERING PHASE 27
 4.3 PREDICTION PHASE 31

5. DISCUSSION AND CONCLUSIONS 40
 REFERENCES 42
 C.V. 44

LIST OF TABLES

Table 2.1 : The number of research “churn management” according to year 5

Table 2.2 : The number of research “churn prediction” according to year 6

Table 2.3 : The number of research “data mining” according to year 6

Table 2.4 : The number of research “anfis” according to year 7

Table 3.1 : The variables in data set 11

Table 4.1 : The customer marital status values 20

Table 4.2 : The customer gender values 20

Table 4.3 : The customer educational status values 20

Table 4.4 : The city values 20

Table 4.5 : The customer profession values 21

Table 4.6 : Attributes and ranking values 26

Table 4.7 : Training results for methods used 36

LIST OF FIGURES

Figure 4.1 : The analysis of the first seven attributes 17

Figure 4.2 : The analysis of the second seven attributes....................................... 18

Figure 4.3 : The analysis of the third seven attributes.. 18

Figure 4.4 : The analysis of the fourth seven attributes.. 18

Figure 4.5 : The analysis of the fifth seven attributes... 19

Figure 4.6 : The analysis of the last six attributes... 19

Figure 4.7 : Distribution of region... 22

Figure 4.8 : Distribution of tenure.. 23

Figure 4.9 : Distribution of age.. 23

Figure 4.10 : Distribution of marital status.. 23

Figure 4.11 : Distribution of address... 24

Figure 4.12 : Distribution of income.. 24

Figure 4.13 : Distribution of ed.. 24

Figure 4.14 : Distribution of employ 25

Figure 4.15 : Distribution of retire 25

Figure 4.16 : Distribution of gender 26

Figure 4.17 : Distribution of reside.. 26

Figure 4.18 : The operations to compare fuzzy c-means and k-means algorithms 27

Figure 4.19 : The distribution of clusters at the end of comparing fuzzy c-means and k-means algorithms 27

Figure 4.20 : The statistic for accuracy of two different algorithms; k-means and fuzzy c-means.. 28

Figure 4.21 : The true positives false positives true negatives and false negatives of k-means and fuzzy c-means algorithms.. 29

Figure 4.22 : The linear correlation between the results of the fuzzy c-means and k-means clustering algorithms 29

Figure 4.23 : The distribution of k-means clusters 29

Figure 4.24 : The distribution of fuzzy c-means clusters.. 30

Figure 4.25 : The statistics table of fuzzy c-means... 30

Figure 4.26 : Anfis classification of training data 31

Figure 4.27 : Anfis classification of testing data 31

Figure 4.28 : Anfis classification of checking data 32

Figure 4.29 : Anfis model of fuzzy interference 32

Figure 4.30 : Fuzzy interference diagram 33

Figure 4.31 : Membership function of equipment last month 34

Figure 4.32 : Membership function of long distance over tenure...........................34

Figure 4.33 : Membership function of toll free over tenure 34

Figure 4.34 : Membership function of equipment over tenure 35

Figure 4.35 : Membership function of log equipment 35

Figure 4.36 : The result of decision tree 37

Figure 4.37 : The rules of probabilistic neural network..37

1. INTRODUCTION

1.1 PROBLEM DEFINITION

Global Service of Mobile (GSM), with the advance in technology in 20th century, is an invention which provides communication between people in any part of the world at any time. After this form of communication has become a market, several big companies developed infrastructure either to be in this market or to increase their market share. Sonar cellular phones have become very widespread in no time. This led to an increase in the number of service providers and cellular phone manufacturers. At this point, a fierce competition had started between companies. Each company, by organizing new promotions continuously, tried to attract more consumers and convince them that their service is the best, the most qualified and the cheapest. After allowing the carryover of phone numbers from one service provider to the other in Turkey, the number of people switching their service provider (carrier) has increased. Therefore, instead of finding new ones, the first priority for GSM service providers became not losing the current customers to other service providers.

All GSM companies try to predict whether their customers will cancel the service or not via several decision support softwares. These predictions are based on the attributes of customers. The more accurate these predictions are the more customers that these companies keep and the less value they lose. It is easier and less costly to keep a current customer than attracting a customer of another GSM service provider. Therefore, customer profile information and search details are very valuable for GSM service providers. As a result, all GSM service providers put too much importance on churn management and spend their energy in this direction.

At the end of 2008, after Telecommunication Regulatory in Turkey allowed carrying the phone number over other GSM service providers, number of customers switching their service providers has increased. Therefore, GSM service providers started monitoring their customers on a regular basis. Using search data of their customers, these companies try to predict whether customers will cancel their service and this enables them to take necessary precautions beforehand.

GSM service providers focus on the demographic information, bill details, contract, and the duration and number of conversation of their customers when they organize promotions to attract their customers (Yu et al., 2005).

The behavior of a customer, her loyalty to the service provider and her service satisfaction are stored in the databases of the GSM service provider that belong to that customer (Hadden et al., 2005).

The churn of customer is a cause of huge loss of services and this is very serious problem for GSM service companies. (Huang B., Kechadi M. T., Buckley B. et al 2011)

If a GSM service provider can predict that a customer is leaving the company, GSM service provider can give new offers to convince the customer to stay. As we stated above, it is more costly and difficult for a GSM service provider to get new customers from other operators compared to convincing her own customers not to leave. Hence, preventing a current customer to leave the company is the first priority for all GSM service providers.

Customer subdivision is very important for telecommunication companies because while making decision to keep old customer in service, find new customer or be aware of churn customer. With customer subdivision companies can take decision strategically for GSM market and companies can understand different demands and different customer groups while making business decision (REN H., Zheng Y., WU Y. et al., 2009)

In this study, we support the process of churn management with the data mining tools. We use some of the frequently used data mining techniques and present each result obtained.

Moreover, we compare each technique with each other. This comparison is also present in this study.

Our objective is to find and publish which data mining techniques give better results in churn management.

In this study four different modeling methods are compared. These methods are DTNB, Naïve Bayes, Support Vector Machine and ANFIS Fuzzy c-means.

1.2 CHURN MANAGEMENT

Churning called as changing a service after lack of satisfaction feeling of customer. At the end of this process, customer changes service from one package to another package or from one GSM operator to another GSM Operator (Mozer at al. 2000).
In Telecommunication Service Sector, Churn management means keeping durability of valuable customer for the telecommunication Company (Kentrias 2001).

Similarly Berson et al. (2000) defines Churn management as the effort of GSM operator to keep its customers.

1.3 DATA MINING

Data mining is a popular tool used by many companies during the decision making process. From data that have no clear relationship, data mining extracts useful information for managers so that they can make operational or strategic decisions for the company easily.

Data mining is frequently used in all industries in the decision making process. It is the entire process of entering and analyzing the raw data, which results in valuable information.

Data mining enables detecting the relationship patterns between the data sets, which are not observed before. This can further be used to obtain numerous results.

Nowadays, companies in all sectors are very careful in collecting the data of their customers.

Together with data mining, it became very important to notice and watch relationship patterns between data sets. Using data mining, one can capture nonobvious relationships between data sets. This can be used to predict customers' demands and behaviors.

The main purpose of data mining is to extract hidden information and to detect pattern within large data sets.

2. LITERATURE SURVEY

There are many studies on Churn management. Using the key words “Churn Management”, “Churn Prediction”, “Data Mining” and “ANFIS,” we have searched Science Direct database to obtain the number of publications on each subject.

Tables 2.1 to 2.4 present our results for each year between 2001 and 2011, and 2001 and earlier.

They indicate that most popular research field is Data Mining and it is followed by Churn Management.

There are 1665 journals in science direct web site with keyword churn management.

Table 2.1 : The number of research on “Churn management” according to year.

Year	Number of journals
2011	179
2010	187
2009	188
2008	188
2007	143
2006	85
2005	65
2004	76
2003	59
2002	63
2001 and earlier	88

There are 1352 journals in science direct web site with keyword churn prediction.

Table 2.2: The number of research on "Churn prediction" according to year.

Year	Number of journals
2011	127
2010	127
2009	127
2008	104
2007	88
2006	69
2005	65
2004	45
2003	56
2002	44
2001 and earlier	115

There are 23161 journals in science direct web site with keyword "data mining".

Table 2.3: The number of research on "Data Mining" according to year.

Year	Number of journals
2011	3481
2010	3140
2009	3050
2008	2702
2007	2431
2006	2043
2005	1505
2004	1335
2003	1191
2002 and earlier	1730

There is 1602 journals in science direct web site with keyword “ANFIS”.

Table 2.4: The number of research on “ANFIS” according to year.

Year	Number of journals
2011	298
2010	225
2009	260
2008	193
2007	133
2006	112
2005	94
2004	61
2003	61
2002	31
2001 and earlier	102

Data mining is a very popular research field in computer science. There are many studies on Churn Management which use Data Mining techniques. More papers are published every year.

In telecommunication terminology, customer is defined as someone who receives a service from the telecommunication company for a few times, and Churn refers to the cancellation of the particular service by that customer. There are different Churn types which we list below.

a) Package Churn

In this Churn type, the customer switches to a different service package within the same telecommunication company. This is not operator churn and it is has not so much effects on companies (B.Q. Huang & T.-M. Kechadi et al., 2010).

b) Service Churn

The customer quits the service in this Churn type (B.Q. Huang & T.-M. Kechadi et al., 2010).

c) Operator Churn

The Customer leaves a GSM operator for another. In this Churn type, the GSM Company loses their customers. The most important aim of churn management is finding operator churn, because operator churn means that loosing customer (B.Q. Huang & T.-M. Kechadi et al., 2010).

For telecommunication sector it became more and more important not to lose a customer and hence to increase the satisfaction of current customers rather than attracting new customers (Reinartz & Kumar et al., 2003).

In the Churn management, companies use two different strategies. One of them is called as Reactive Churn Management while the other is called Proactive Churn Management. In reactive Churn management, GSM Operators make some new offers to keep the customer when he/she is not satisfied with the service and wants to cancel it. On the other hand, proactive Churn management does not wait for customer's churn declaration. The system tries to find customers who have the highest probability of churn (Burez and Van den Poel et al., 2007).

There are many reasons leading to the churn process. Among many others, the price, coverage area, call quality are very important for customer satisfaction. If one or more reasons exist, churn may happen (R.N. Bolton et al., 1998).

Customer satisfaction can be measured by monitoring customer behavior periodically. This monitoring process consists of some steps. First, customers' data such as demographical and billing information must be collected in a data warehouse so that managers can use to reach a conclusion on whether a customer will churn or not. In order to reach a conclusion, several data mining techniques are used (Berson, A., Smith, S., & Thearling, K. et al., 2000).

3. MATERIALS & METHODS

3.1 MATERIALS

3.1.1 Program

In this study, computer programs that are used are as follows:

- Weka Version 3.6: Weka is a collection of machine learning algorithms for data mining tasks. Weka contains data pre-processing, classification, regression, clustering, association rules and visualization.

- KNIME (Konstanz Information Miner) Version 2.3.0: KNIME is a useful program for data integration and analysis. KNIME is an open source program and it has also data mining tools.

- MATLAB Version 7.1: MATLAB a high-level language and interactive environment that enables you to perform computationally intensive task faster than the traditional programming language.

3.1.2 Data

In this study all of data that are used are taken from a telecommunication company. Data identifiers are hidden. Only demographic, call and billing data are in the dataset.
There are two data sets for 15.000 customers:

- Data set about demography,
- Data set about calling

3.1.2.1 Data preparation

There are two main data set: call detail data set and demographic data set. In demographic data set there is information about customer's family and financial data. On the other hand in call detail data set there is information about customer's calling habitude.

In data preparation process we tried to replace null value and miss value with meaningful values. We used Knime to solve this problem. Knime's miss value algorithm allows us to choose value of miss value that will change. We decided to give minimum value of each column to miss values.

Each row in database belongs to a different customer. Every attribute's value means a variable for a customer.

Table 3.1: The Variables in Data Set.

Name of Data	Type	Detail
customerid	numeric	ID
region	numeric	Geographic Indicator
tenure	numeric	Months with service
age	numeric	Age in years
marital	numeric	Marital Status
address	numeric	Years at current address
income	numeric	Household income in thousand
ed	numeric	Level of education
employ	numeric	Years with current employer
retire	numeric	Retired
gender	numeric	Gender
reside	numeric	Number of people in household
tollfree	numeric	Toll free service
equip	numeric	Equipment rental
callcard	numeric	Calling card service
wireless	numeric	Wireless service
longmon	numeric	Long distance last month
tollmon	numeric	Toll free last month
equipmon	numeric	Equipment last month
cardmon	numeric	Calling card last month
wiremon	numeric	Wireless last month
longten	numeric	Long distance over tenure

Table 3.1: The variables of data (continued).

Data Name	Value Type	Comment
tollten	numeric	Toll free over tenure
equipten	numeric	Equipment over tenure
cardten	numeric	Calling card over tenure
wireten	numeric	Wireless over tenure
multline	numeric	Multiple lines
voice	numeric	Voice mail
pager	numeric	Paging Service
internet	numeric	Internet
callid	numeric	Caller ID
callwait	numeric	Call waiting
forward	numeric	Call forwarding
confer	numeric	3- way calling
ebill	numeric	Electronic billing
loglong	numeric	Log-long distance
logtoll	numeric	Log-toll free
logequi	numeric	Log-equipment
logcard	numeric	Log-calling card
logwire	numeric	Log-wireless
lninc	numeric	Log income
custcat	numeric	Customer category

3.1.2.2 Data understanding

Data Sources

There are 15.000 customers in the data source. Data source can be divided two different data sets. They are Demographic information and Call Detail information.

Demographic Input Data

- Customer Region,
- Age,
- Marital Status,
- Address,
- Income,

- Marital Status,
- Educational Background,
- Gender,
- Getting Customer Date,
- Retire Status.
- Years at current address
- Employ - Profession
- Number of people in household

Call Detail Input Data

- Toll free service
- Equipment rental
- Wireless service
- Calling card service
- Long distance over tenure
- Internet
- Voice mail
- Paging service
- Call waiting
- Call forwarding
- 3-way calling
- Electronic billing
- Customer category

3.2 METHODS

3.2.1 C-means Algorithm

At the end of clustering, *N* different patterns to *c* clusters are found. While finding a cluster, the most important point is the similarity (high or low) within the data. Therefore it is necessary to find the distance between data so that we can measure the similarity. Every cluster has a centroid; therefore, number of clusters and cendroids is the same. Clusters are described by their centroids.

The c-means algorithm produces *c* non-empty subset by partitioning *N* objects X_k. As it is mentioned before every cluster has one centroid and centroids are computed as

$$v_i = \frac{\sum_{x_k \in u_i} x_k}{|c_i|} \tag{3.1}$$

In this formula:

$|c_i|$ represents the total number of objects in U_i.

The process of computing centroids and finding new membership is repeated until no new membership can't be found.

3.2.2 Fuzzy C-Means

Fuzzy C-Means (FCM) algorithm is best known and most widely used method among clustering techniques within the fuzzy division. FCM method objects can have membership in two or more than two clusters. In FCM logic every object is belong to a cluster with value of interval [0, 1]. Sum of each object's membership values must be "1". A data point closer to the centroid of a cluster belongs to that cluster.

FCM clustering algorithm, created by Dunn in 1973 and then developed by Bezdek in 1981. FCM has two main steps: In first step, the algorithm finds centroids of each cluster . In the next step every object is assigned to a cluster depending on its distance to centroids.

Algorithm is executed for minimizing the objective function which is generalization of least square method.

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2, \quad 1 \leq m < \infty , \tag{3.2}$$

where m is the fuzzifier, Algorithm is started with creating **U** membership matrix randomly. The next step is calculating centroids. Centroids are found by the following equation:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{3.3}$$

According to its cluster, centroids U matrix is calculated via the next equation. Previous and current **U** matrices are compared and this step continues until the difference between the two is less then ε

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - c_i \right\|}{\left\| x_i - c_k \right\|} \right)^{2/(m-1)}} \tag{3.4}$$

3.2.3 Adaptive Neuro Fuzzy Interference System (ANFIS)

ANFIS learning algorithm is a hybrid algorithm that contains two different methods. They are gradient descent and least square estimating methods. Rule parameters are updated iteratively via this method until actual error becomes smaller than estimated error.

There are two main and important steps in iteration process: forward and backward steps. In the forward step least square estimate method is used. On the other hand, the gradient descent method is used in the backward step.

During the forward step process, former parameters are fixed and current parameters are obtained. During the backward step process, current parameters are fixed and former parameters are updated.

While applying ANFIS model one has to pay attention to some important properties. These are the number of membership function, the number of fuzzy rules and the number of training epochs. The values of these properties are very important. If the properties are given wrong values, the algorithm will also produce wrong result.

The objective of training process is to minimize error between ANFIS output and actual objective. It is possible to produce logical rules and to apply these rules to the dataset.

The output of each rule is a linear combination of a constant term and each input.

4. FINDINGS

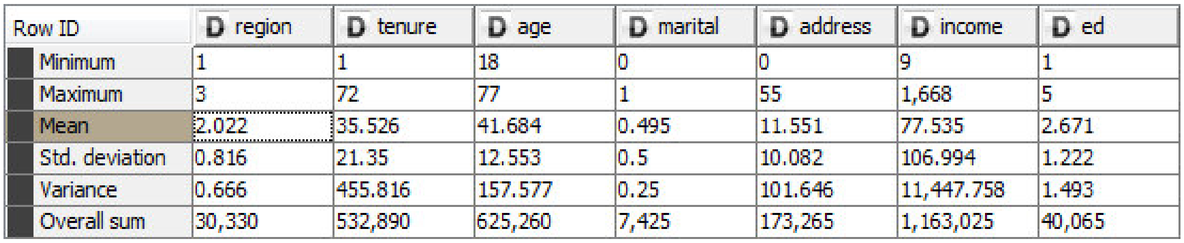
Findings will be examined in three main phases. In the first phase, data are analyzed and results of this phase are input for second phase. In second phase clusters are created and data are gathered in subgroups. The second phase which is called “Clustering Phase” is necessary for the third phase. The third phase is called “Prediction Phase”. In this phase, whether a customer will churn or not is predicted and rules are created.

4.1 DATA ANALYSING PHASE

There are 43 different fields in data set and all fields are examined with Weka analyzer.

The data is analyzed by Weka program and summary statistics (Maximum, Minimum, Mean, Standard Deviation, Variance and Overall Sum) belonging the data are shown below.

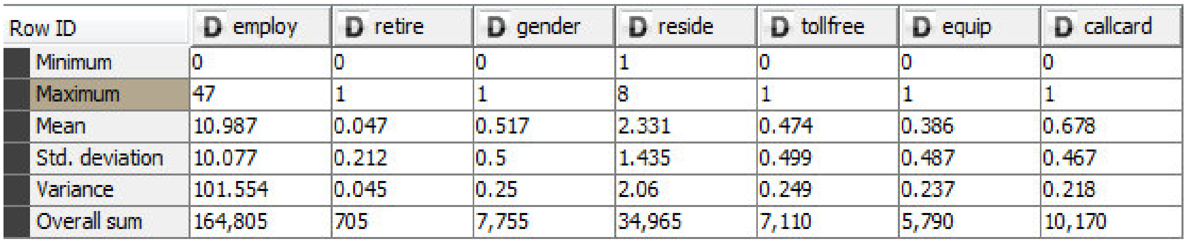
In figure 4.1 there are statistical data about first seven attributes. These attributes are region, tenure, age, marital status, address, income and ed.



Row ID	D region	D tenure	D age	D marital	D address	D income	D ed
Minimum	1	1	18	0	0	9	1
Maximum	3	72	77	1	55	1,668	5
Mean	2.022	35.526	41.684	0.495	11.551	77.535	2.671
Std. deviation	0.816	21.35	12.553	0.5	10.082	106.994	1.222
Variance	0.666	455.816	157.577	0.25	101.646	11,447.758	1.493
Overall sum	30,330	532,890	625,260	7,425	173,265	1,163,025	40,065

Figure 4.1 : The analysis of first seven attributes.

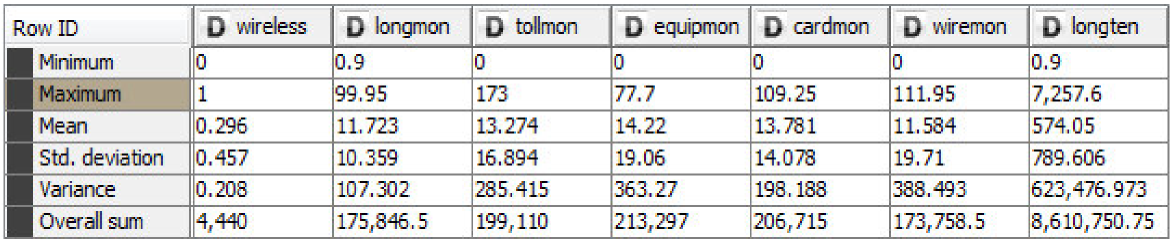
In figure 4.2 there are statistical data about second seven attributes. These attributes are employ, retire, gender, reside, tollfree, equip and callcard.



Row ID	D employ	D retire	D gender	D reside	D tollfree	D equip	D callcard
Minimum	0	0	0	1	0	0	0
Maximum	47	1	1	8	1	1	1
Mean	10.987	0.047	0.517	2.331	0.474	0.386	0.678
Std. deviation	10.077	0.212	0.5	1.435	0.499	0.487	0.467
Variance	101.554	0.045	0.25	2.06	0.249	0.237	0.218
Overall sum	164,805	705	7,755	34,965	7,110	5,790	10,170

Figure 4.2 : The analysis of second seven attributes.

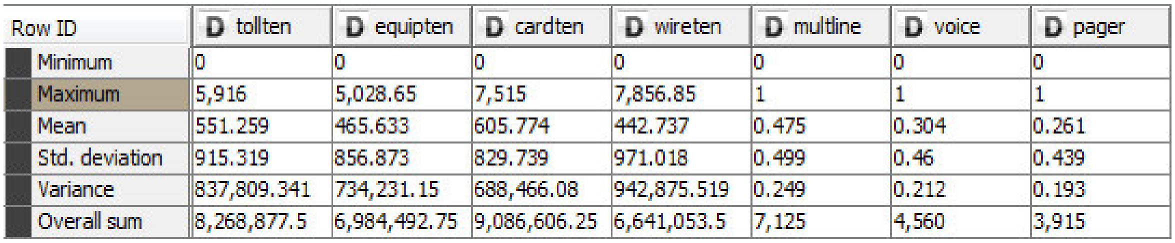
In figure 4.3 there are statistical data about third seven attributes. These attributes are wireless, longmon, tollmon, equipmon, cardmon, wiremon and longten.



Row ID	D wireless	D longmon	D tollmon	D equipmon	D cardmon	D wiremon	D longten
Minimum	0	0.9	0	0	0	0	0.9
Maximum	1	99.95	173	77.7	109.25	111.95	7,257.6
Mean	0.296	11.723	13.274	14.22	13.781	11.584	574.05
Std. deviation	0.457	10.359	16.894	19.06	14.078	19.71	789.606
Variance	0.208	107.302	285.415	363.27	198.188	388.493	623,476.973
Overall sum	4,440	175,846.5	199,110	213,297	206,715	173,758.5	8,610,750.75

Figure 4.3 : The analysis of third seven attributes.

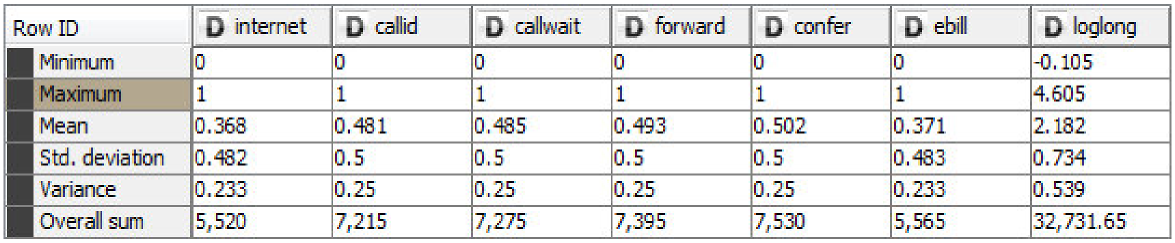
In figure 4.4 there are statistical data about fourth seven attributes. These attributes are tollten, equipten, cardten, wireten, multline, voice and pager.



Row ID	D tollten	D equipten	D cardten	D wireten	D multline	D voice	D pager
Minimum	0	0	0	0	0	0	0
Maximum	5,916	5,028.65	7,515	7,856.85	1	1	1
Mean	551.259	465.633	605.774	442.737	0.475	0.304	0.261
Std. deviation	915.319	856.873	829.739	971.018	0.499	0.46	0.439
Variance	837,809.341	734,231.15	688,466.08	942,875.519	0.249	0.212	0.193
Overall sum	8,268,877.5	6,984,492.75	9,086,606.25	6,641,053.5	7,125	4,560	3,915

Figure 4.4 : The analysis of fourth seven attributes.

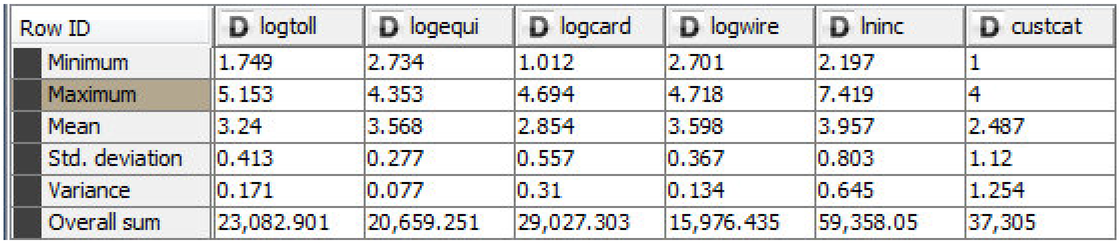
In figure 4.5 there are statistical data about fifth seven attributes. These attributes are internet, callid, callwait, forward, confer, ebill and loglong.



Row ID	D internet	D callid	D callwait	D forward	D confer	D ebill	D loglong
Minimum	0	0	0	0	0	0	-0.105
Maximum	1	1	1	1	1	1	4.605
Mean	0.368	0.481	0.485	0.493	0.502	0.371	2.182
Std. deviation	0.482	0.5	0.5	0.5	0.5	0.483	0.734
Variance	0.233	0.25	0.25	0.25	0.25	0.233	0.539
Overall sum	5,520	7,215	7,275	7,395	7,530	5,565	32,731.65

Figure 4.5 : The analysis of fifth seven attributes.

In figure 4.6 there are statistical data about last six attributes. These attributes are logtoll, logequi, logcard, logwire, lninc and custcat.



Row ID	D logtoll	D logequi	D logcard	D logwire	D lninc	D custcat
Minimum	1.749	2.734	1.012	2.701	2.197	1
Maximum	5.153	4.353	4.694	4.718	7.419	4
Mean	3.24	3.568	2.854	3.598	3.957	2.487
Std. deviation	0.413	0.277	0.557	0.367	0.803	1.12
Variance	0.171	0.077	0.31	0.134	0.645	1.254
Overall sum	23,082.901	20,659.251	29,027.303	15,976.435	59,358.05	37,305

Figure 4.6 : The analysis of last six attributes.

In data there are 42 different attributes.

In following tables there is information about analyzing of categorical variables. In following tables, variables and their values are shown.

The values of Marital Status are shown in Table 4.1.

Table 4.1 : The Customer Marital Status Values

Values	Meaning
1	Married
0	Single

The values of Gender are shown in Table 4.2.

Table 4.2 : The Customer Gender Values

Values	Meaning
1	Male
0	Female

The values of Educational Status are shown in Table 4.3.

Table 4.3 : The Customer Educational Status Values

Values	Meaning
1	Primary School
2	High School
3	University
4	Graduate Student
5	Doctorate

The values of Address are shown in Table 4.4.

Table 4.4 : The City Values

Adana	Bilecik	Erzurum	Karaman	Mersin	Tekirdağ
Adıyaman	Bingöl	Eskişehir	Kars	Muğla	Tokat

Table 4.4 : The City Values(Continued)

Afyon	Bitlis	Gaziantep	Kastamonu	Muş	Trabzon
Ağrı	Bolu	Gazimağusa	Kayseri	Nevşehir	Tunceli
Aksaray	Burdur	Giresun	Kilis	Niğde	Uşak
Amasya	Bursa	Girne	Kırıkkale	Ordu	Van
Ankara	Çanakkale	Gümüşhane	Kırklareli	Osmaniye	Yalova
Antalya	Çankırı	Hakkâri	Kırşehir	Rize	Yozgat
Ardahan	Çorum	Hatay	Kocaeli	Sakarya	Zonguldak.
Artvin	Denizli	Iğdır	Konya	Samsun	
Aydın	Diyarbakır	Isparta	Kütahya	Siirt	
Balıkesir	Düzce	İstanbul	Lefkoşa	Sinop	
Bartın	Edirne	İzmir	Malatya	Sivas	
Batman	Elazığ	Kahramanmaraş	Manisa	Şanlıurfa	
Bayburt	Erzincan	Karabük	Mardin	Şırnak	

The values of Profession are shown in Table 4.5.

Table 4.5 : The Customer Profession Values

Akademisyen / Üniversite Öğretim Görevlisi	Hatalı Girilmiş	Pilastik Sanatlar (Ressam Heykeltıraş … vb.)
Analist / Programcı	Hemşire / Ebe	Polis / Güvenlik Görevlisi
Astsubay	Hizmet / Ticaret	Reklam / Halkla ilişkiler
Aşçı / Garson / Barmen	Hostes	Sağlık Personeli
Avukat	İmalat / Sanayi	Sahne Sanatları (Bale Tiyatro … vb.)
Büyük Sanayici	İnsan Kaynakları	Sekreter / Yönetici Asistanı
Çalışmayan	İşçi	Serbest Meslek / Esnaf (Bakkal Nalbur … vb.)
Çiftçi / Balıkçı	Küçük Sanayici	Serbest Meslek / Zanaatkâr (Terzi … vb.)
Dealer / Broker	Manken / Fotomodel	Sosyal ve İdari Bilimci

Table 4.5 : The Customer Profession Values(Continued)

Diğer	Memur	Sporcu / Antrenör
Diplomat	Mevsimlik İşçi	Subay (Teğmen Yüzbaşı Binbaşı Albay … vb.)
Diş Hekimi	Mimar / İçMimar / Dekoratör	Şöför
Eczacı	Muhasebeci / Mali Müşavir	Teknisyen
Emekli	Mühendis	Temel Bilimci (Fizikçi Kimyager … vb.)
Ev Hanımı	Noter	Tercüman / Çevirmen
Finans	Öğrenci	Tıp Doktoru

In following graphs, distributions of data attributes have been shown. Distribution of each data attribute is important in analyzing the attributes and the data.

- The values of customers’ region are shown below.

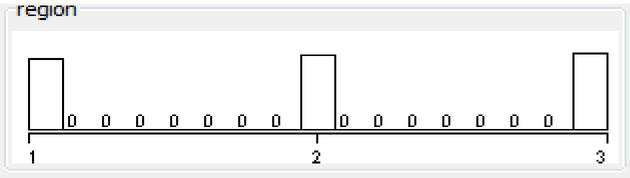


Figure 4.7: Distribution of Region

- The values of customers’ tenure are shown below.

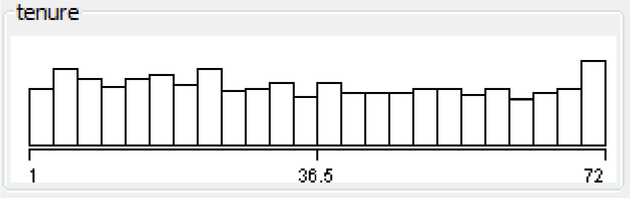


Figure 4.8 : Distribution of Tenure

- The values of customers’ age are shown below.

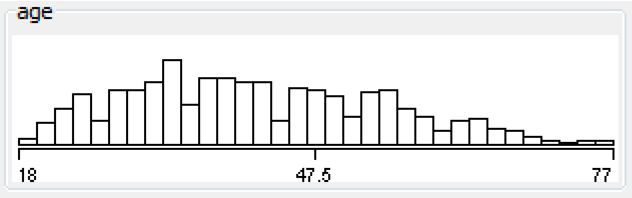


Figure 4.9: Distribution of Age

- The values of customers’ marital status are shown below.
  - “0” represents single,
  - “1” represents married,

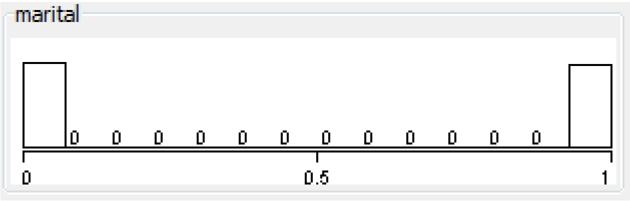


Figure 4.10 : Distribution of Marital Status

- The values of customers’ addresses are shown below.

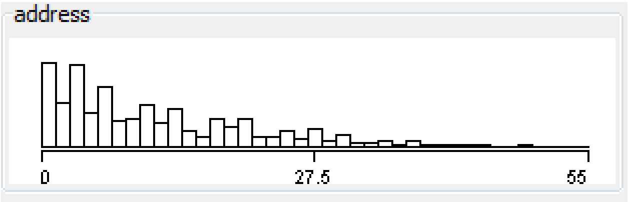


Figure 4.11 : Distribution of Address

- The values of customers’ income are shown below.

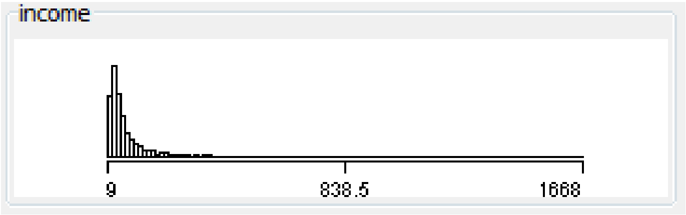


Figure 4.12 : Distribution of Income

- The values of customers’ education levels are shown below.
  - “1” represents primary school,
  - “2” represents high school,
  - “3” represents university,
  - “4” represents graduate student,
  - “5” represents doctorate,

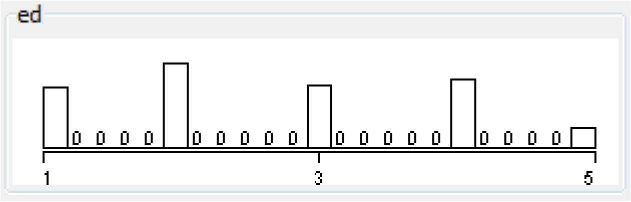


Figure 4.13: Distribution of Ed

- The values of customers' employ are shown below.

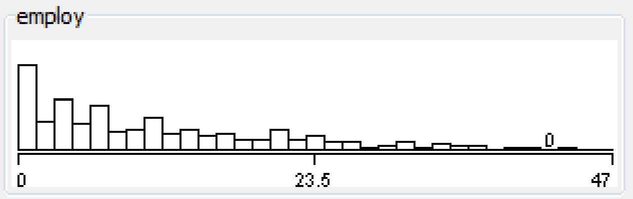


Figure 4.14: Distribution of Employ

- The values of customers' retire are shown below.
  - "0" represents not retired,
  - "1" represents retired,

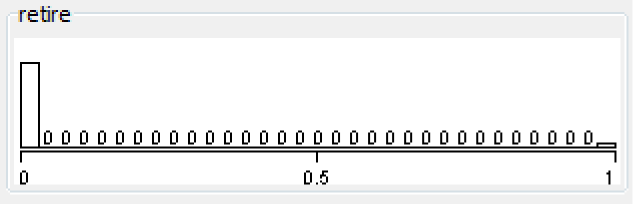


Figure 4.15: Distribution of Retire

- The values of customers' gender are shown below.
  - "0" represents female,
  - "1" represents male,

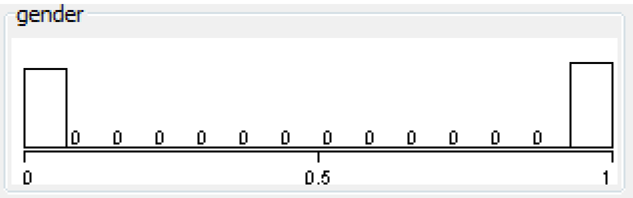


Figure 4.16 : Distribution of Gender

- The values of customers' reside are shown below.

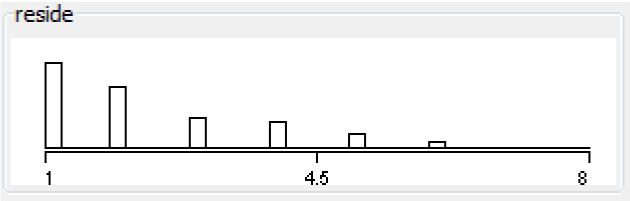


Figure 4.17: Distribution of Reside

- Ranked attributes and values are shown in Table 4.6.

Table 4.6: Attributes and Ranking values

Attribute Name	Gain Ranking Filter
Longten	0.8156362
Equipten	0.4488072
Tollten	0.3928553
Equipmon	0.3762778
Logequi	0.373694

4.2 CLUSTERING PHASE

In this paper, for clustering phase, we try to determine the proper customer groups by using customers' attributes and two different clustering algorithms. These clustering algorithms are k-means and fuzzy c-means. After clustering the same data with two different algorithms, these algorithms are compared to reach the best result for this phase.

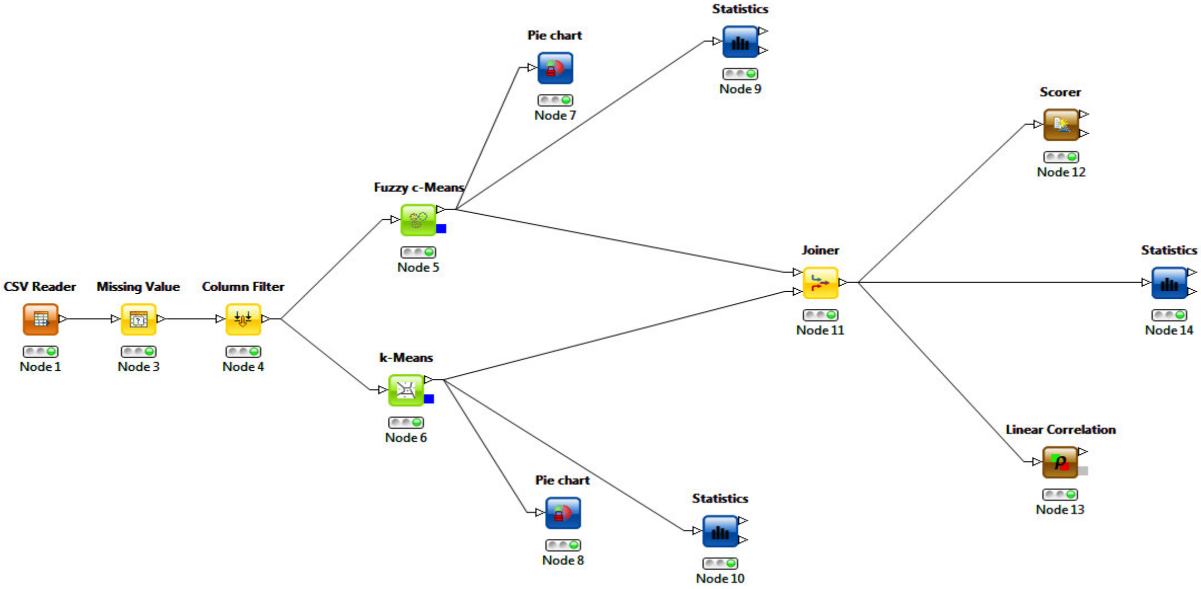
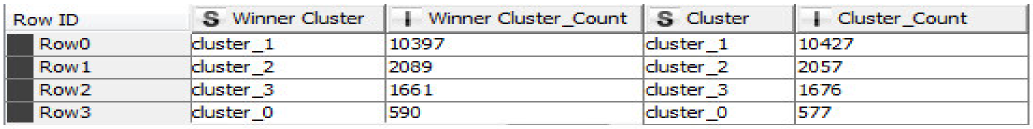


Figure 4.18 : The Operations to Compare Fuzzy C-Means and K-Means Algorithms.

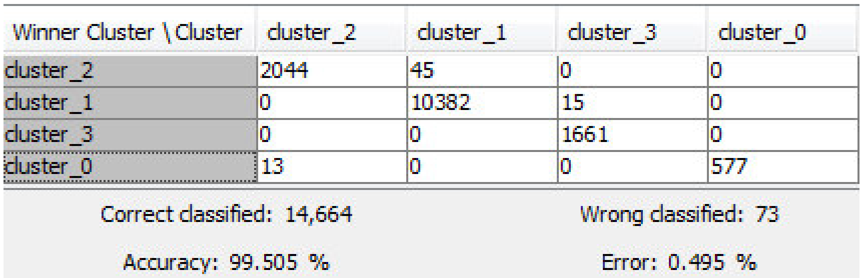
Following graph shows us the results of k-means and fuzzy c-means algorithms. "Winner Cluster" indicates result of fuzzy c-means algorithm; "Winner Cluster Count" indicates total number of objects in each cluster. "Cluster" indicates result of k-means algorithm and "Cluster Count" means that total number of objects in each cluster after result of k-means.



Row ID	S Winner Cluster	I Winner Cluster_Count	S Cluster	I Cluster_Count
Row0	cluster_1	10397	cluster_1	10427
Row1	cluster_2	2089	cluster_2	2057
Row2	cluster_3	1661	cluster_3	1676
Row3	cluster_0	590	cluster_0	577

Figure 4.19 : The Distribution of Clusters at The End of Comparing Fuzzy C-Means and K-Means Algorithms.

Following graphs illustrates the accuracy of two different algorithms for the same dataset.



Winner Cluster \ Cluster	cluster_2	cluster_1	cluster_3	cluster_0
cluster_2	2044	45	0	0
cluster_1	0	10382	15	0
cluster_3	0	0	1661	0
cluster_0	13	0	0	577

Correct classified: 14,664

Wrong classified: 73

Accuracy: 99.505 %

Error: 0.495 %

Figure 4.20 : The Statistic for Accuracy of Two Different algorithms; K-Means and Fuzzy C- Means.

The next graph is the values of True Positives, False Positives, True Negatives and False Negatives result of these two different algorithms above.

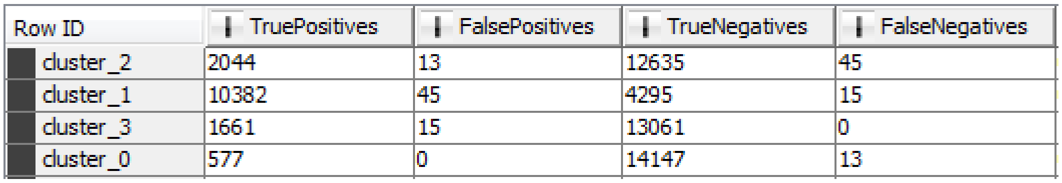
True Positives means total number of positive examples correctly estimated by the clustering.

False negative means total number of positive examples wrongly estimated as negative by the clustering.

False positive means total number of negative examples wrongly estimated as positive by the clustering.

True negative means total number of negative examples correctly estimated by the clustering.

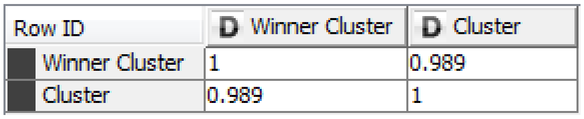
Figure 4.21 there in information about the distribution of TP, TN, FP, FN values according to clusters.



Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives
cluster_2	2044	13	12635	45
cluster_1	10382	45	4295	15
cluster_3	1661	15	13061	0
cluster_0	577	0	14147	13

Figure 4.21 : The True Positives False Positives True Negatives and False Negatives of K-Means and Fuzzy C-Means Algorithms.

The next figure shows the linear correlation between the results of K-means and Fuzzy c-means algorithms.



Row ID	Winner Cluster	Cluster
Winner Cluster	1	0.989
Cluster	0.989	1

Figure 4.22 : The linear correlation between the results of the Fuzzy c-means and K-means clustering algorithms.

In figure 4.23 distribution of clusters after applying K-means method is shown.

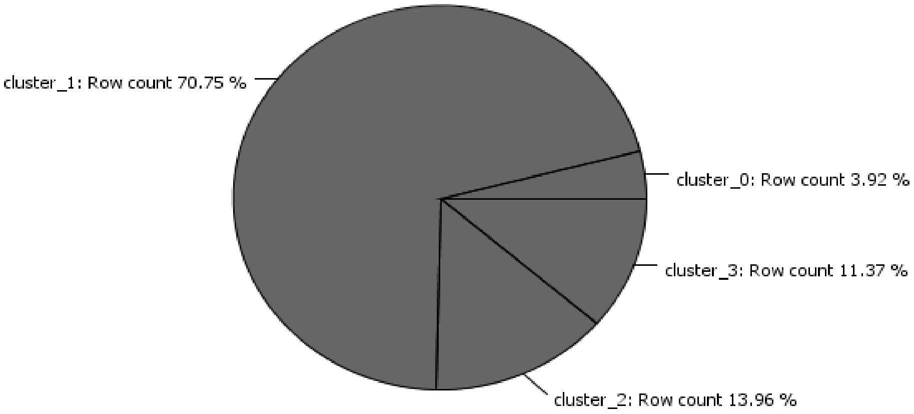


Figure 4.23 : The distribution of K-Means Clusters.

The next figure (Figure 4.24) shows the distribution of clusters obtained via fuzzy c-means algorithm.

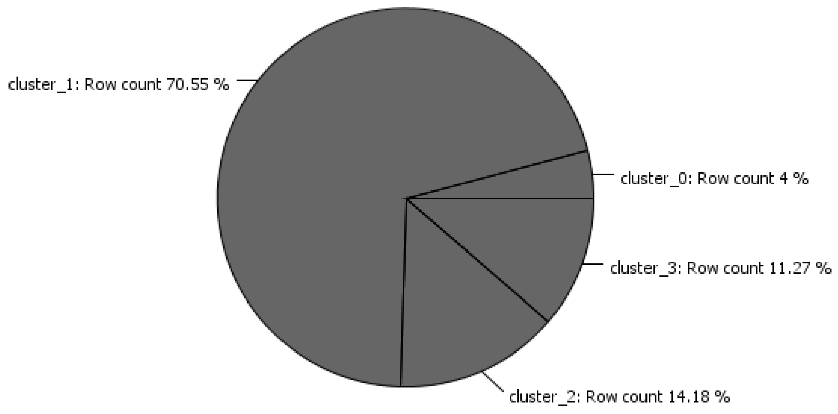
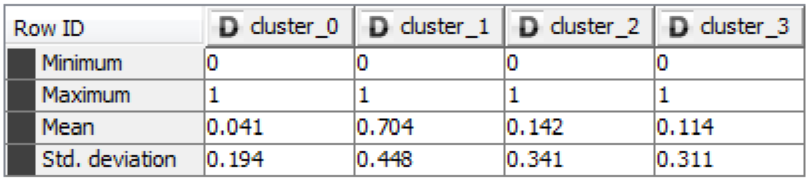


Figure 4.24 : The Distribution of Fuzzy C-Means Clusters.

All summary statistics (minimum, maximum, mean and standard deviation) for each cluster are shown in Figure 4.25. These statistical values are taken after applying Fuzzy C-Means algorithm to the dataset.



Row ID	D cluster_0	D cluster_1	D cluster_2	D cluster_3
Minimum	0	0	0	0
Maximum	1	1	1	1
Mean	0.041	0.704	0.142	0.114
Std. deviation	0.194	0.448	0.341	0.311

Figure 4.25 : The Statistics Table of Fuzzy C-Means.

4.3 PREDICTION PHASE

For the prediction phase, we use ANFIS by Matlab. In this study Sub-clustering is used to obtain the results. Here are the parameter values for training the model: The range of influence is 0.5, squash factor is 1.25, accept ratio is 0.5; rejection ratio is 0.15.

With giving these parameters ANFIS gives these results:

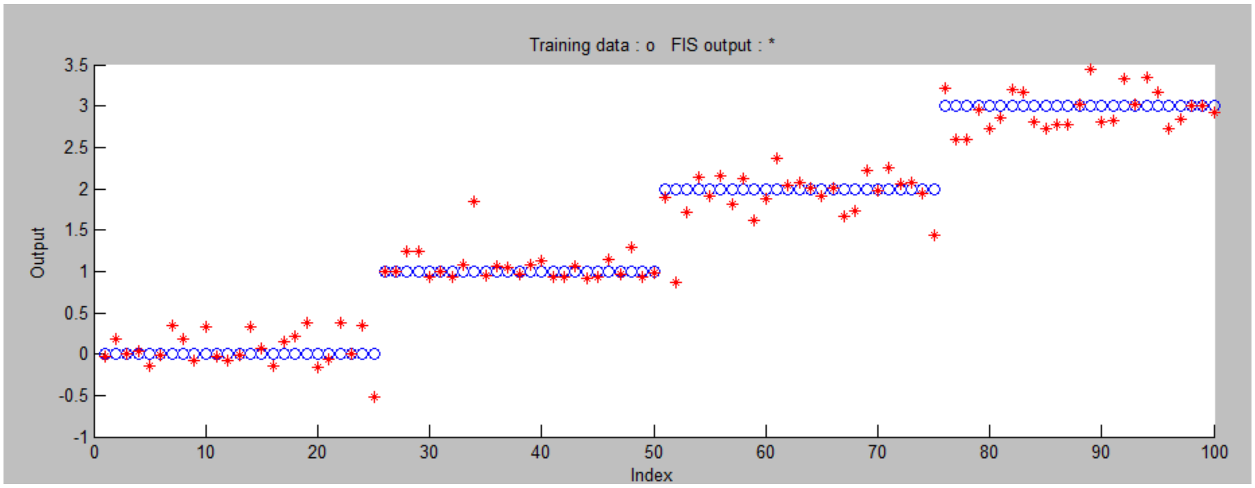


Figure 4.26 : ANFIS classification of training data

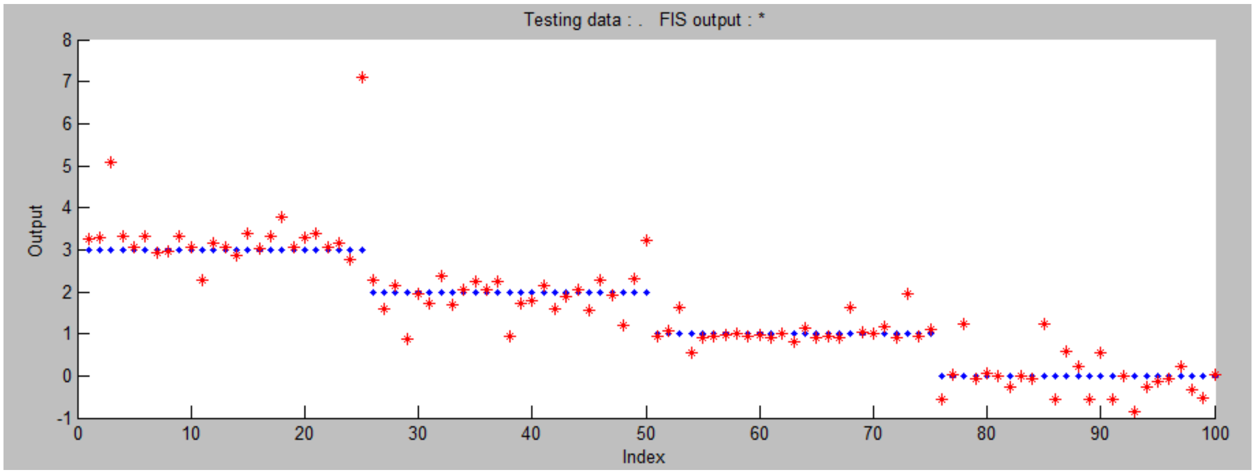


Figure 4.27 : ANFIS classification of testing data

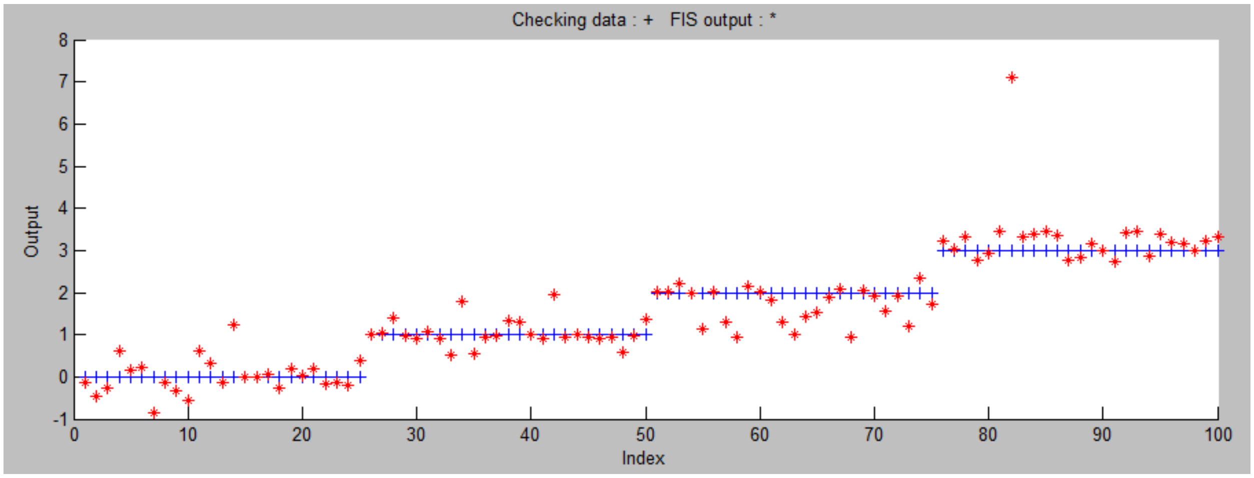


Figure 4.28 : ANFIS classification of checking data

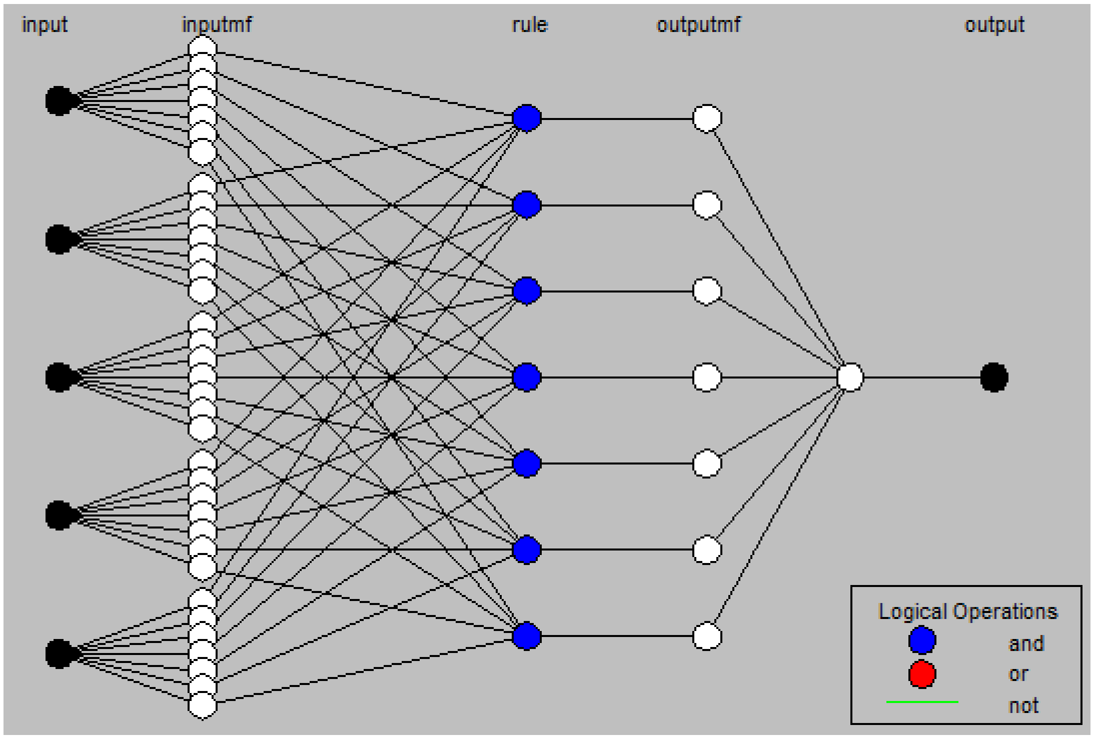


Figure 4.29 : ANFIS model of fuzzy interference

Figure 4.29. displays plot of input factors for fuzzy inference and the output results in the conditions. The fuzzy interference diagram is the composite of all factor diagrams. This diagram shows all parts of fuzzy interference process.

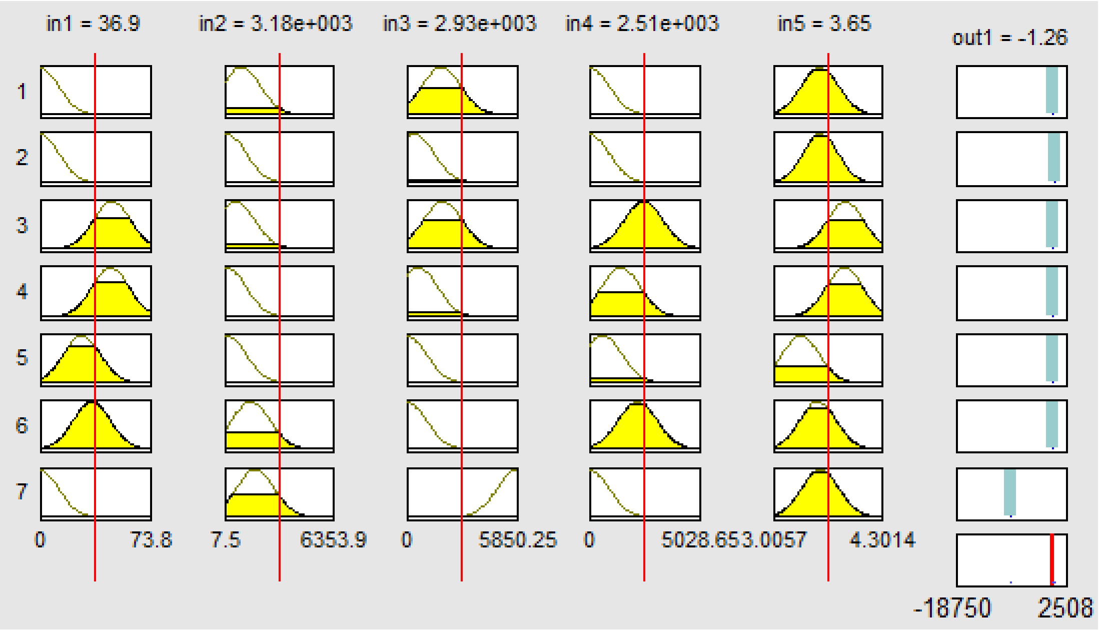


Figure 4.30 : Fuzzy interference diagram

By applying ANFIS method, ANFIS creates membership functions of each input. Following graphs, membership functions of each input are shown. These inputs are Long distance over tenure (longten), Equipment over tenure (equipten), Toll free over tenure (tollten), Equipment last month (equipmon) and Log equipment (logequi) variables. Membership function of each input can also be examined after and before training by using these graphs.

If an input has an effect on average, it causes considerable deviation from the original curve.

In figures. 4.31. – 4.35., vertical axis is the value of the membership function; horizontal axis denotes the value of input factor.

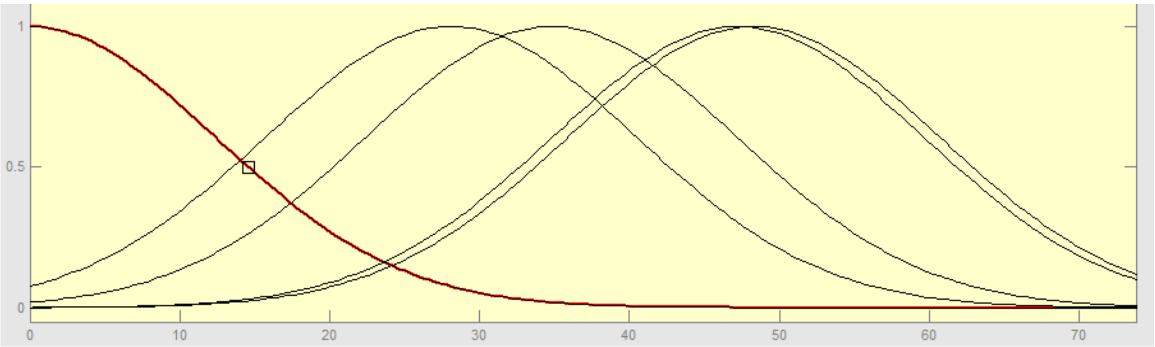


Figure 4.31 : Membership function of Equipment last month (equipmon)

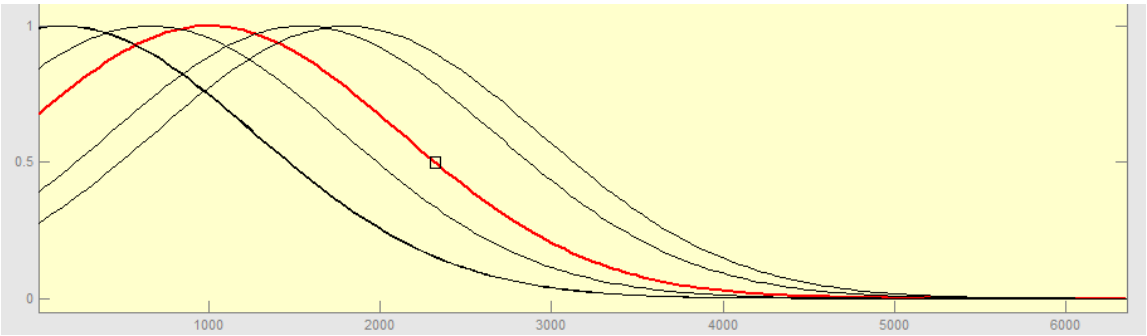


Figure 4.32 : Membership function of Long distance over tenure (longten)

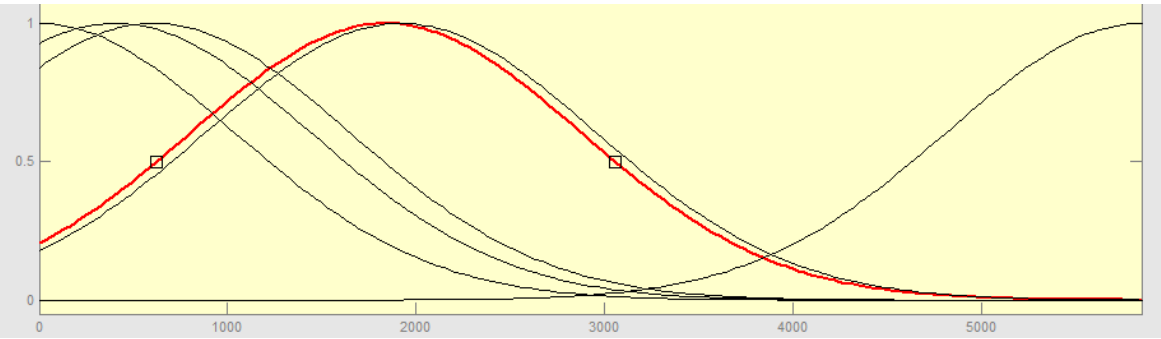


Figure 4.33 : Membership function of Toll free over tenure (tollten)

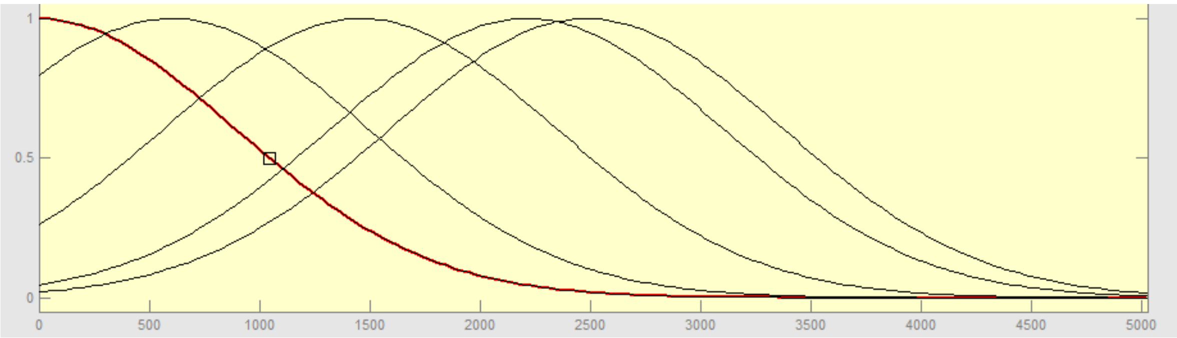


Figure 4.34 : Membership function of Equipment over tenure (equipten)

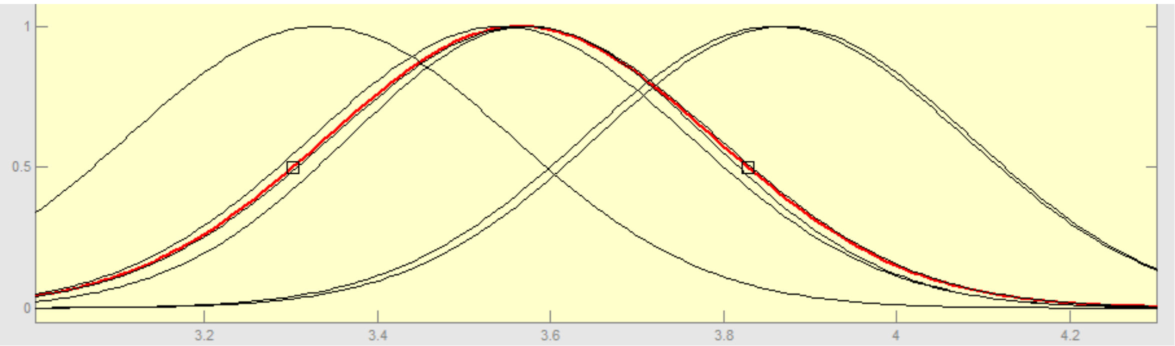


Figure 4.35 : Membership function of Log equipment (logequi)

Here some results of comparing different methods are shown in this table. These results are taken by using training data.

These formulas are used while computing sensitivity, specificity, precision and accuracy of methods.

TP : *number of true positive*
FN : *number of false negative*
FP : *number of false positive*
TN : *number of true negative*

$$sensitivity = \frac{TP}{TP + FN} \tag{4.1}$$

$$specificity = \frac{TN}{FP + TN} \tag{4.2}$$

$$precision = \frac{TP}{TP + FP} \tag{4.3}$$

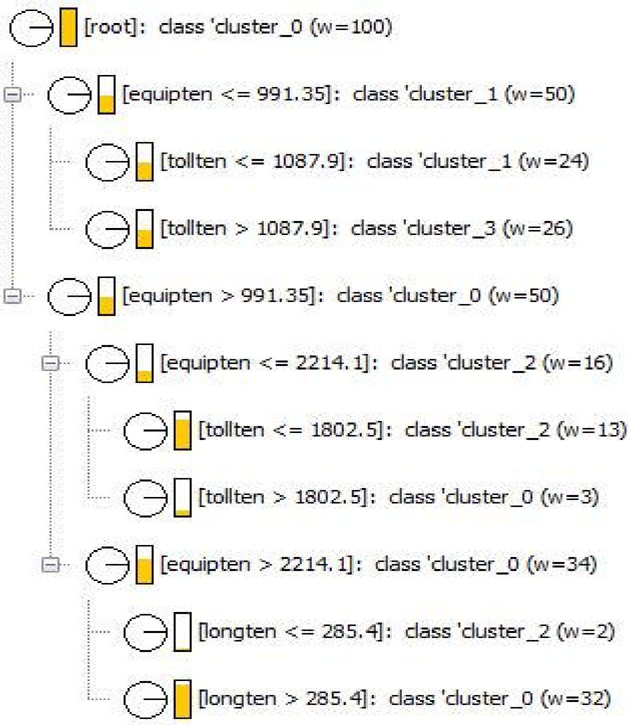
$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.4}$$

Table 4.7 : Training results for methods used.

Method	Sensitivity	Specificity	Precision	Accuracy
Decision Tree	0.83	0.92	0.83	0.89
Naïve Bayes	0.61	0.87	0.62	0.80
SVM	0.51	0.83	0.51	0.75
Probabilistic Neural Network	0.85	0.92	0.85	0.91
ANFIS	0.86	0.95	0.86	0.93

As a result of testing four different methods we can say that ANFIS is the best method for used dataset.

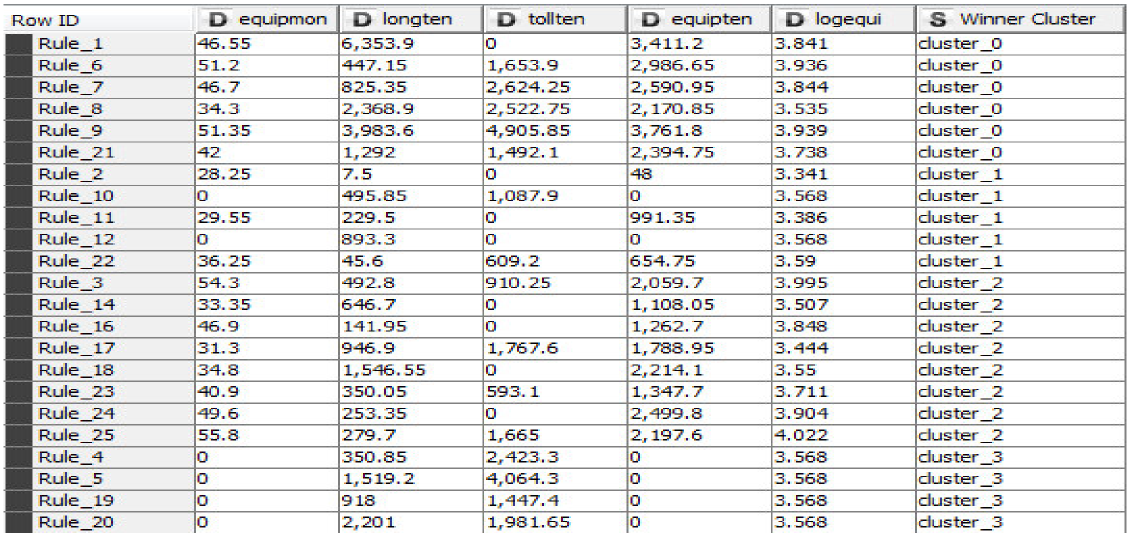
After Decision tree predictor the result is as following figure.



```
[root]: class 'cluster_0 (w=100)
    [equipten <= 991.35]: class 'cluster_1 (w=50)
        [tollten <= 1087.9]: class 'cluster_1 (w=24)
        [tollten > 1087.9]: class 'cluster_3 (w=26)
    [equipten > 991.35]: class 'cluster_0 (w=50)
        [equipten <= 2214.1]: class 'cluster_2 (w=16)
            [tollten <= 1802.5]: class 'cluster_2 (w=13)
            [tollten > 1802.5]: class 'cluster_0 (w=3)
        [equipten > 2214.1]: class 'cluster_0 (w=34)
            [longten <= 285.4]: class 'cluster_2 (w=2)
            [longten > 285.4]: class 'cluster_0 (w=32)
```

Figure 4.36 : The Result of Decision Tree

After applying Probabilistic Neural Network algorithm the rules are:



Row ID	D equipmon	D longten	D tollten	D equipten	D logequi	S Winner Cluster
Rule_1	46.55	6,353.9	0	3,411.2	3.841	cluster_0
Rule_6	51.2	447.15	1,653.9	2,986.65	3.936	cluster_0
Rule_7	46.7	825.35	2,624.25	2,590.95	3.844	cluster_0
Rule_8	34.3	2,368.9	2,522.75	2,170.85	3.535	cluster_0
Rule_9	51.35	3,983.6	4,905.85	3,761.8	3.939	cluster_0
Rule_21	42	1,292	1,492.1	2,394.75	3.738	cluster_0
Rule_2	28.25	7.5	0	48	3.341	cluster_1
Rule_10	0	495.85	1,087.9	0	3.568	cluster_1
Rule_11	29.55	229.5	0	991.35	3.386	cluster_1
Rule_12	0	893.3	0	0	3.568	cluster_1
Rule_22	36.25	45.6	609.2	654.75	3.59	cluster_1
Rule_3	54.3	492.8	910.25	2,059.7	3.995	cluster_2
Rule_14	33.35	646.7	0	1,108.05	3.507	cluster_2
Rule_16	46.9	141.95	0	1,262.7	3.848	cluster_2
Rule_17	31.3	946.9	1,767.6	1,788.95	3.444	cluster_2
Rule_18	34.8	1,546.55	0	2,214.1	3.55	cluster_2
Rule_23	40.9	350.05	593.1	1,347.7	3.711	cluster_2
Rule_24	49.6	253.35	0	2,499.8	3.904	cluster_2
Rule_25	55.8	279.7	1,665	2,197.6	4.022	cluster_2
Rule_4	0	350.85	2,423.3	0	3.568	cluster_3
Rule_5	0	1,519.2	4,064.3	0	3.568	cluster_3
Rule_19	0	918	1,447.4	0	3.568	cluster_3
Rule_20	0	2,201	1,981.65	0	3.568	cluster_3

Figure 4.37 : The Rules of Probabilistic Neural Network

PNN algorithm generated 25 different rules and there are 100 different customers' information in each data set (training and testing).

ANFIS fuzzy c-means algorithm generated 7 rules and these rules can be expressed as follows:

Rule 1: [33 660 0 5028.65 3.4965] [cluster_0]
If equipment last month is 33 and long distance over tenure is 660 and toll free over tenure is 0 and equipment over tenure is 5028.65 and log equipment is 3.4965 then output is cluster_0

Rule 2: [51.2 447.15 1238.2 2170.85 3.5351] [cluster_0]
If equipment last month is 51.2 and long distance over tenure is 447.15 and toll free over tenure is 1238.2 and equipment over tenure is 2170.85 and log equipment is 3.5351 then output is cluster_0

Rule 3: [0 7.5 0 0 3.005] [cluster_1]
If equipment last month is 0 and long distance over tenure is 7.5 and toll free over tenure is 0 and equipment over tenure is 0 and log equipment is 3.001 then output is cluster_1

Rule 4: [48.4 279.7 1802.22 1481.35 4.02] [cluster_2]
If equipment last month is 48.4 and long distance over tenure is 279.7 and toll free over tenure is 1802.22 and equipment over tenure is 1481.35 and log equipment is 4.02 then output is cluster_2

Rule 5: [49.2 447.15 1238.2 2498.35 3.895] [cluster_2]
If equipment last month is 49.2 and long distance over tenure is 447.15 and toll free over tenure is 1238.2 and equipment over tenure is 2498.35 and log equipment is 3.895 then output is cluster_2
Rule 6: [31.3 1037.65 4.176 1395.1 3.171] [cluster_2]

If equipment last month is 31.3 and long distance over tenure is 1037.65 and toll free over tenure is 4.176 and equipment over tenure is 1395.1 and log equipment is 3.171 then output is cluster_2

Rule 7: [0 2291.1 1447.4 0 3.568] [cluster_3]
If equipment last month is 0 and long distance over tenure is 2291.1 and toll free over tenure is 1447.4 and equipment over tenure is 0 and log equipment is 3.568 then output is cluster_3

5. DISCUSSION AND CONCLUSIONS

We try to analysis and find some results about customer behavior. We have a sample data in telecom industry. There are 42 different attributes and 15000 records in data. There was missing values and we used some methods to fill these values. First step of this study is to analyzing data. We look at the values of all attributes.

Then we applied data mining techniques to get meaningful information from data set.
Next step was to carry out which attributes are more efficient and effective in data set. We chose five attributes from dataset and in process of creating model we used this attributes. We compared fuzzy-cmeans and k-means methods results. The results of fuzzy c-means and k-means are used as input for classify process and we can easily say that fuzzy c-means is best algorithm for used data set.

In this paper we choose ANFIS fuzzy-cmeans method to get best result. In prediction phase we used ANFIS fuzzy c-means method and we added ANFIS classification figures, membership function of each used attributes and fuzzy interference diagram. We also try three other methods which take the same data as input. In this study there is result about comparing four different classify methods. These classify methods are Decision Tree, Naïve Bayes, Support Vector Machine and ANFIS. We try to compare four methods by using their confusion matrix. We get sensitivity, specificity, precision and accuracy of each method and these values are included in this study.

ANFIS generated 7 rules and sensitivity, specificity, precision and accuracy values are better than other methods. We have seen that Probabilistic Neural Network gave the best value after ANFIS but PNN created 25 rules. Number of rules is huge according to ANFIS. ANFIS created just only 7 rules.

In conclusion, all telecom customers have been collected in the same group according to their similarities by using different methods. Finally dataset have been an input for different classify methods. The most useful and efficient method is ANFIS Fuzzy C-means. We try to show how to apply customer segmentation and find the best method for data step by step. This paper will be a source for other publication.

REFERENCES

Berson, A., Smith, S., & Thearling, K., 2000. *Building data mining applications for CRM.* New York, NY: McGraw-Hill.

B.Q. Huang, T.-M. Kechadi, B. Buckley, G. Kiernan, E. Keogh, T. Rashid. *A new feature set with new window techniques for customer churn prediction in land-line telecommunications*

Burez, J., Van den Poel, D., 2007. *Crm at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services*. Expert Systems with Applications, 32, pp. 277–288.

Felkin M., 2007 *Comparing Classification Results between N-ary and Binary Problems.* Studies in Computational Intelligence (SCI) 43, pp. 277–301

Hadden, J., Tiwari, A., Roy, R., Ruta, D., 2005.*Computer assisted customer churn management: State of the art and future trends*. Journal of Computers and Operations Research, 34 (10), pp. 2902–2917.

Huang B., Kechadi M. T., Buckley B., 2011. Customer churn prediction in telecommunications. Expert Systems with Applications.

Karahoca A, Karahoca D., 2011, *GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system*. Expert Systems with Applications 38 pp. 1814-1822.

Kentrias, S., 2011. *Customer relationship management: The SAS perspective* www.cm2day.com,.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., Kaushanksky, H., 2000. *Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry*. IEEE Transactions on Neural Networks, 11 (3), pp. 690–696.

Reinartz, W., Kumar, V., 2003. *The impact of customer relationship characteristics on profitable lifetime duration*. Journal of Marketing, 67 (1), pp. 77–99.

Ren H., Zheng Y., Wu Y., 2009. *Clustering analysis of telecommunication customers* The Journal of China Universities of Posts and Telecommunications, 16(2), pp. 114–116

R.N. Bolton, 1998. *A dynamic model of the duration of the customer's relationship with a continuous service provider: the role of satisfaction*, Marketing Science 17 pp. 45–65.

Yu, W., Jutla, D. N., Sivakumar, S. C., 2005. *A churn management alignment model for managers in mobile telecom*. In Proceedings of the 3rd annual communication networks and services research conferences (2005) pp. 48–53.

C.V.

Name Surname	: Evren Arifoğlu
Address	: Hürriyet mh. İzci sk. No:10 D:4 Bağcılar / İstanbul
Place of Birth, Year of Birth	: Demirözü, 1985
Foreign Language	: English, French
Primary School	: ŞİRİNTEPE İLKÖĞRETİM OKULU, 1996
Middle School	: ŞİRİNTEPE İLKÖĞRETİM OKULU, 1999
High school	: BAĞCILAR LİSESİ, 2003
University	: BAHÇEŞEHİR ÜNİVERSİTESİ, 2008
Graduate School	: BAHÇEŞEHİR ÜNİVERSİTESİ, 2011
Institute	: Graduate School of Natural and Applied Sciences
Graduate Program	: M.S. Program Computer Engineering
Experience	: BAHÇEŞEHİR ÜNİVERSİTESİ, 2008 – (Present)