T.C.

BAHÇEŞEHİR UNIVERSITY

CUSTOMER SEGMENTATION FOR CHURN MANAGEMENT BY USING ANT COLONY

M.S. Thesis

Batuhan GÜLLÜOĞLU

İstanbul, 2011

T.C.

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

CUSTOMER SEGMENTATION FOR CHURN MANAGEMENT BY USING ANT COLONY

M.S. Thesis

Batuhan GÜLLÜOĞLU

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	: Customer Segmentation for Churn Management by Using Ant Colony
Name/Last Name of the Student	: Batuhan GÜLLÜOĞLU
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 (Supervisor) :

Asst. Prof. Dr. Mehmet Alper TUNGA :

Asst. Prof. Dr. Yalçın ÇEKİÇ :

ABSTRACT

CUSTOMER SEGMENTATION FOR CHURN MANAGEMENT BY USING ANT COLONY ALGORITHM

Güllüoğlu, Batuhan

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

September 2011, 41 pages

Data mining is interested in clustering, by similarities of data. Some of clustering techniques are evolutionary and optimization techniques. Characteristic selection is used for novel hybrid modeling.

Customer priorities are very important for companies. Moreover, customer priorities must be determined, and campaigns must be ordered according to these priorities. Customer segmentation was done with Ant Colony algorithm. Shortest path approach is used in Ant Colony algorithm. Moreover, clustering is done by the euclidean distance formula in Ant Colony algorithm.

Customer segmentation attributes are mostly related with the satisfaction factors, but some of them were eliminated by using ranker. These results are mostly related with the customer's income, tenure, equip, callcard and reside. These attributes are the most important satisfaction factors not to lose customers as expected. There are many reasons in changing GSM operator for subscribers and it is very important for companies to predict if subscriber will change GSM operator or not. For this reason companies that gives GSM services have to monitor subscribers behavior and predict one step forward. In this study changing subscribers' GSM operator will be predicted by using data mining techniques.

Keywords: Data Mining, Customer Segmentation, Ant Colony, Churn Management

ÖZET

KARINCA KOLONİ ALGORİTMASI KULLANILARAK MÜŞTERİ KAYIP YÖNETİMİ YAPILMASI

Güllüoğlu, Batuhan

Bilgisayar Mühendisliği

Tez Danışmanı: Doç. Dr. Adem KARAHOCA

Eylül 2011, 41 sayfa

Veri madenciliği verilerin kümelenmesi ve bu kümeleme sayesinde verilerde ki benzerliklerin ortaya çıkarılması için kullanılan bir tekniktir. Veri kümeleme için bir çok teknik bulunmaktadır. Bunlardan bazıları gelişimini tamamlayan ve optimizasyon teknikleridir. Karakteristik seçimi yeni çıkmış hibrid modellemedir. Müşteri segmentasyonu; bankaları için müşterilerinin önceliklerini belirlemede önemli rol oynar.

Müşteri öncelikleri belirlenmeli ve kampanyalar bu müşteri memnuniyetlerine göre düzenlenmelidir. Ant Koloni algoritması en kısa yol yaklaşımını içeren bir tekniktir. Ant Koloni algoritmasında öklit mesafe formülü kullanılmaktadır.

Müşteri segmentasyonu ile ilgili sonuçlar en çok müşterinin; geliri, kaç ay o şirket ile çalıştığı, ödemiş olduğu bedeli, kullanımış olduğu servis ile ilintilidir. Bu değişkenler beklenen şekilde müşteri kaybetmemek için olan memnuniyet faktörleridir. Müşterinin şirketten ayrılması ayrılmaması halini alması için bu parametreler kullanılmaktadır. Deneysel sonuçlar bu parametrelerde ki değişimlere bağlı olarak verilerin tekrar tekrar weka anfisde naive bayes methodu ile gerçekleşmektedir.

Anahtar Kelimeler: Veri Madenciliği, Müşteri Segmentasyonu, Ant Koloni, Müşteri Kayıp Yönetimi

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 PROBLEM DEFINITION

Customer management and satisfaction are the most important issues for companies. Companies have to know their customers. Moreover, companies have to know their customer's characteristics and they should determine campaign according to these characteristics. Companies have to influence their customers due to priorities. The most important profit is customer satisfaction and this will be an advantage or disadvantage according to the company's approach (Chris Rygielski, Jyun-Cheng Wang, David C. Yen.,1970).

The most important aim is providing customer satisfaction according to their priorities. Customer priorities are very important for campaigns. Moreover, customer priorities must be determined then, campaigns must be ordered according to these priorities. Another aim is to determine the characteristics of campaings. This is the main problem for the companies. Because, customer's priorities are changed day by day so customer segmentation must be done regularly by the companies. This study is related with the customer segmentation with Ant Colony Optimization algorithm and this algorithm provides clustering of the updated data. Therefore, the customer satisfaction is obtained.

There is a main problem that is mentioned about for campaigns. It affects the profit and loss ratio's of the campaigns. For this reason, campaigns have to cluster the customer data according to determine their the necessities of customers. Clustering is used for determine the customer types for campaigns. There are a lot of types of clustering algorithms.

The goal of clustering is to reduce the amount of data by categorizing or grouping similar data items together. Such grouping is pervasive in the way humans process information, and one of the motivations for using clustering algorithms is to provide automated tools to help in constructing categories or taxonomies (Jardine and Sibson, 1971).

You can divide clustering methods into two basic types: hierarchical and partitional clustering. Each type has a wealth of subtypes and there are different algorithms to the clusters. In my thesis, ant colony optimization is used for clustering the customer data sets. Ant colony is one of the most important technique to make selection from data sets (Ying Zhao, George Karypis, 1973).

There are some algorithms to search the properties. These are complete, sequential and random algorithms. As you understand, sequential and random search algorithms determine searching strategy. Moreover, there are some approaches to classify the data and these approaches determine the evaulation criteria. These evaluation criterias based on Wrapper and Hybrid approaches.

Keeping the customers that you have needs more source than gaining new customers. Churn models are developed to detect the customers about the attrite and taking action about them (Zeng H., Cheung Y. M., 2009).

There are some hypothesis that is indicated in this study. One of the hypothesis is about customer's age and income. If the customer is under eighteen years old and has no income then he or she won't spend more than five hundred turkish liras in a month. Moreover, if the customer is over eighteen years old woman and has some income, then she has expense more than her salary. Therefore, she has to have a credit card which has limit more than the customer's salary. In addition, if the customer is an employee and has constant salary, he or she will spend more than his or her salary (Ahmed Al-Ani, 2003).

In addition, wrapper ant colony clustering technique is used to have detention of customers according to clustering of customer data set. Customers data set is related with the satisfaction factor so wrapper ant colony technique has to provide clustering according to the ecludian vector.

1.2 CHURN MANAGEMENT

Churn management is one of the most important thing that provides differences in the system according to the necessities of customers. According to the customers satisfaction, the bank changes its strategy to be more successful. Strategy is usually about the products and services that the customers use (Chiz Wei, 2005).

Churn management revises the customer data and analysis the results. Results depend on customers satisfaction. Churn management makes the customers satisfaction in a stable position.

Churn management is an issue that the customers are leave from their companies or not due to their priorities and satisfaction. In order to provide satisfaction of customers, necessities have to be determined. Churn management begins with the data collecting from the customers. All of the parameters related with customers are investigated and studied. There are some methods to determine the common properties of the customers. These methods provide some clusters between these customer data. Then, comparable data sets are produced according to these clusters. There will be processes on this data to determine priorities to provide satisfaction. There are a lot of iterations according to these data and parameters are updated on each iteration step. Then, if the customer leave the company as it means churn, the company should analyse the updated parameters and produce some results according to these solutions. This is churn management.

1.3 BACKGROUND

There are a lot of methods for clustering of customer data. Filter and wrapper approaches is used for unsupervised property selection. Ant colony optimization clustering is used to solve the problems of clustering. Customer detention or customer churn is applied by some methods.

Filter clustering is related with filtering the data according to the tool that the data clustered. There are a lot of attributes in churn management data. But, clustering is related with accuracy so the most important attributes have to be determined according to the accuracy. These attributes are related with data but if the clustering isn't accurate, then, it should be abated among these attributes and this filtering should be done unless the most accurate parameters are obtained (Tsai C.F., 2009).

Moreover, customer detention should be done before filter clustering as it means that; before data analysing. There are some parameters that are not directly related with clustering data. These types of data should be eliminated before filter clustering. This is physical type of clustering.

In addition, weka, ant colony optimization and myra are some of the methods that the data is clustered by them. All of them has the same fallacy. The centroids are determined randomly then all of the objects are investigated related with distance. If the nearest distance from the centroid is determined then this object is in that cluster. Ecludian distance is used in Ant Colony Optimization. Myra and weka investigates the similarities of values on each vector and then cluster them in groups.

In conclusion, there exits some problems about the data set but the main point is to determine the valuable data among these data set and process these data. There is usually null values among these data set and these null values are set with meaningful data that is provided from the average of other columns.

Train data is obtained as a result of wrapper approach. Train data means the data that is eliminated among all data set. Then, testing the accuracy continues and the data set is eliminated until it reaches the train data. In this study, train data is two hundred and eighty rows and nineteen columns (Cotter S.F., Kreutz-Delgado K., 2001).

This is a result of wrapper filtering and it provides the most accuracy solutions related with data set and conclude as customer satisfaction. Nowadays, churn management is done by all kind of companies in Turkey by using some of these algorithms.

There are a a lot of people who works for churn management. These people always collect data by their databases and systems and then they work on data to make it valuable. Moreover, this data has some meaningless values so this data has to be manipulated according to the clusters. There should be iterations for these data and in these iterations, data should be redevelopment in each step (Shin-Yuan Hung, 2008).

2. PREVIOUS STUDIES

There are some data mining techniques to customer segmentation with ant colony optimization. I investigated the literature in Science Direct. Customer segmentation, data mining, ant colony were investigated. I investigated the number of documents related with my topic, moreover I investigated subject areas and publication years.

There are a lot of studies related with data mining techniques with ant colony. I wanted to find the documents related with customer segmentation with ant colony optimization to determine the customers that will be usefull for the campaign. There are totaly one hundred and twenty two topics related with these topics.

There are a lot of techniques that is used in Ant Colony Optimization. Firstly, Deneubourg used ACO in 1991. This study is related with robots. Deneubourg used this algorithm to provide some property to robots. Robots decide to take up or drop the objects according the this algorithm results. This is a result that is similar to the result of this study. In this study, the data are clustered according to ACO algorithm and then classification is done and there exits some rules according to classification like robots operations (Xiao-bin Z., 2009).

Gutowitz changed this algorithm in 1993. It is the same algorithm but robots schedule the tasks related with the operations of robots and respectively do these operations.

Lumer and Faieta updated Gutowitz algorithm in 1995. This study is related with tasks. Tasks are done according to the distances. These distances are obtained due to the clustered data as a result of ACO algorithm (Cheung Y. M., 2009).

De Castro improved Lumer and Faieta's study. This study is related with robots learning. He provides a determining matrix related with the distances. According to these matrix, robots do operations respectively in the fallacy of ACO algorithm. Moreover, pheromone is used in his study and similar operations are seen as a result of pheromone level (Chow, 2008).

Weili improved the study related with entropy. Entropy is related with the operations of robots. In this study, Faieta's algorithm is improved related with memory. Moreover, there exists a counter in this study and this counter occupy errors in distance matrix.

Shelokar’s study is nearly same as this study. Pheromone level is updated on each iteration and the most similar data is clustered according to the distances and pheromone level. This is determined according to the updated pheromone level.

Zeng and Cheung has an algorithm related with property selection. This algorithm provides an index related with the similar properties. This index shouldn't be pre-determined (Xiao-bin Z., 2009).

All of these methods and algorithms are has the same solution as this study mentioned. The most important topic is clustering data and determine the methodology related with data set. The operations are made to determine the similarities and then provide some properties according to these similarities.

3. MATERIALS & METHODS

3.1 MATERIALS

There are some materials related with clustering and classification. Weka is a clustering tool which classify clustered data by using naive bayes method. Myra is another tool that clusters data according to the similarities of data that use ant miner algorithm depends on java programming language.

In order to use Ant Colony Optimization Algorithm, there must be written some code related with loops and if types that calculates the distances between vectors and then the distance matrix is operated in Matlab according to the classification rules. These materials and versions are respectively mentioned as you see below.

- Weka Version 3.6.5: Weka is machine learning tool that was written in Java programming language. It provides classification, regression, clustering, pre-processing and visualization of data. Weka determine centroids randomly and then calculates the all objects distance between the centroids and these objects. Then, it clusters data and prepare a result distance matrix which is waited to classify.

- C++ Version 3.1: C++ is an "object oriented" programming language which is a superset of c programming language and provides high level programming.

- MATLAB Version 7.1.2: MATLAB is a matrix-based computation that was designed for scientific calculations which was written in c and provides high level programming.

3.1.1 Data

Data set is taken from one of the public Turkish banking company. Data set has about six months duration operational data of this bank. Data set is related with 70.000 customers.

This data set consists of some variables due to determine customer's churn attitude and to have customer segmentation. In addition, this data set includes operational and demographic structure of customer data.

There are a lot of studies related with recency, frequency and monetary variables due to determine customer's churn management state.

Recency, frequency and monetary variables are used in customer segmentation and customer behaviors.

Churn data has fourty two attributes and it has approximately fifteen thousand data. These attributes are related with customer's age, region, birth of date, birth of place, monthly income, gender, marital, tenure and thirty four attributes grow in.

These data are obtained from all customers and these data are updated regularly according to provide accuracy. Moreover, if one of these attributes has no value in the data set then the average value of that attribute is assigned to null value.

These data is related with the current customer information. Age is one of the most important attribute of this data set. Age determines the profiles of the customers and the expenses of the customers. Moreover, income is another important attribute for this data set. Income has direct proportion with the expenses of the customers.

Marital, birth of place and birth of date effects the campaigns that must be done to these customers. Because, these kinds of customers have some different properties according to their marital and birth of place and date.

Data set has numerical data. These data is related between logic. Some of the attributes has only one or zero value. If the attribute is true then it will be one in the data set and if the attribute is false then it will be zero in this study.

3.1.1.1 Data Dictionary

There are some techniques to determine the appropriate patterns. These techniques are related with the churn data set. Weka ranker is one of the most important technique to determine the attributes that is used in this study. There are more than fourty attributes and there are totally nineteen coherent attributes that are given by weka ranker technique. Another technique is;the to determine the values from past data sets related with customers. Recency, Frequency and Monetary values are used to determine these patterns. Recency attributes are respectively customerid, age, income, cardmon, tollfree, equipten, wireten, longten, wiremon, tenure, tollten (Z. Marzuki, 2004).

These values determines the processes and customer events in customer segmentation data. Moreover, these data provides the most important events that customers have on products. Due to provide euphemism about customer models, these values are used. Recency, frequency and monetary values come into existence by the data of the bank company. Monetary attributes are respectively insurance_debt, pers_loan, insurance_amt, loan_amt, rltnship_length. Moreover, multline, loglong, callid are frequecy attributes.

These values catch the customers' account information and customer transactions in the company.

Table 3.1 RFM Recency Variables

RFM TYPE	COLUMN_NAME	DATA TYPE	DISTINCT VALUES
recency	customerid	NOMINAL	{F,T}
recency	Age	NOMINAL	{F,T}
recency	income	NOMINAL	{F,T}
recency	cardmon	NOMINAL	{F,T}
recency	tollfree	NOMINAL	{F,T}
recency	equipten	NOMINAL	{F,T}

recency	wireten	NOMINAL	{F,T}
recency	longten	NOMINAL	{F,T}
recency	wiremon	NOMINAL	{F,T}
recency	tenure	NOMINAL	{F,T}
recency	tollten	NOMINAL	{F,T}

Table 3.1 RFM Recency Variables (Continued)

Table 3.2 RFM Monetary Variables

RFM TYPE	**COLUMN_NAME**	**DATA TYPE**	**DISTINCT VALUES**
monetary	insurance_debt	ORDINAL	{L,H}
monetary	pers_loan	ORDINAL	{L,H}
monetary	insurance_amt	ORDINAL	{L,H}
monetary	loan_amt	ORDINAL	{L,H}
monetary	prd3_amt	ORDINAL	{L,H}
monetary	prd5_amt	ORDINAL	{L,H}
monetary	prdtot_amt	ORDINAL	{0,L,H}
monetary	prdtot_debt	ORDINAL	{L,H}
monetary	prdtot2_amt	ORDINAL	{0,L,H}
recency	rltnship_length	NOMINAL	{}

Banking company is in contact by the some ways that with the assistance of these ways customers connect with their bank. There is a data below the table. There are three ways that customers provide connection with bank. All the ways have two subcategories which are seen as multline and cardten.

Table 3.3 Ways Connection Variables

COLUMN_NAME	**DATA TYPE**	**DISTINCT VALUES**
Multline	NOMINAL	{F,T}
Cardten_1	NOMINAL	{F,T}
Loglong	NOMINAL	{F,T}
Callid	NOMINAL	{F,T}
Cardten_2	NOMINAL	{F,T}

Customers events are provided by using variables that are demographics. Demographic values in data haven't quality so, these are not used.

3.1.1.2 Data Pre-Processing

Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data (Chiz Wei, 2005).

In this study, data cleansing is used to provide correctness in the algorithm. There are some loss data in the attributes and these values are changed with match numbers. After then attribute distributions are controlled.

There are some updates after analyzing these controlled data. Some of the patterns are mainly have simple features to control. so these data are monitored easily out of the segmentation. Other data is illustrated at random to form customer segmentation set.

3.1.1.3 Discretization

In data mining, discretization process is known to be one of the most important data preprocessing tasks. Most of the existing machine learning algorithms are capable of extracting knowledge from databases that store discrete attributes (Chiz, 2005).

If the attribute are continuous, the algorithms can be integrated with a discretization algorithms which transform them into discrete feature (Chiz, 2005).

Bar charts are watched over for both attributes. Separation points are determined due to allocation of the data set. Separation point breaks are relabelled as bins. As a result last data set comes into exist from these attributes.

3.2 METHODS

3.2.1 Ant Colony Optimization Clustering

Ant colony optimization is one of the clustering algorithms that is used in this study depends on P.S.Shelokar, V.K.Jayaraman and B.D.Kulkarni methodology.

3.2.2 ACO For Clustering

The range of the objects is decreased to solve the problem of segmentation N objects to K clusters. Data set is specified by N. The result set contains defined clusters of the data set and is specified as S.

τ : NxK matrix of pheromone detention

$\tau_{i,j}$: Detention matrix of i' th of j'th form. Pheromone sign matrix is resettled to τ_0

Result structure process is ant's randomly behaviour. Ant selects the things depend on the sign pheromone matrix. Objects are become more desirable to select if pheromone invests on them. Ant's selection chances depends on pheromone concentration of nominee objects to all sediment pheromone concentration.

$$p_{i,j} = \frac{\tau_{i,j}}{\sum_{i=1}^{k} \tau_{i,k}}, \mathrm{j} = 1..\mathrm{K} \qquad \textbf{(3.1)}$$

Building results depends on the agents that utilized the pheromone matrix info. Pheromone matrix is updated corresponding.

Local investigation utilizes fitness data values to have better results after both of the agents build the results and pheromone matrix is updated according to their degree.

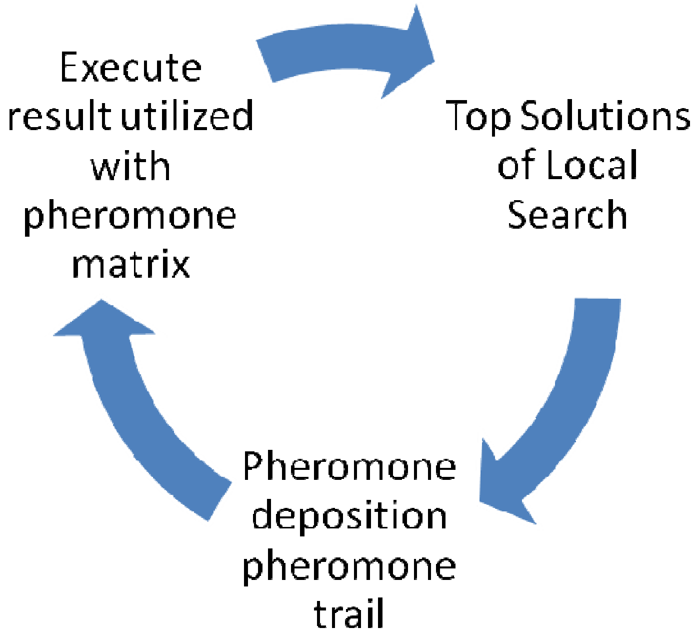

Execute result utilized with pheromone matrix
Top Solutions of Local Search
Pheromone deposition pheromone trail

Figure 3.1 Lifecycle of ants in ant colony clustering

There are two varieties of methods that are form the solution and these methods are respectively examination and actuation (Jardine and Sibson, 1971).

Permanent pheromone condensation is characterized as Q_0. This pheromone condensation permanent specify stability between examination and actuation that is $0 < Q_0 < 1$.

The solution is found by producing a number that is randomly selection.

The following thing is abated that depends on the pheromone condensation which uses $p_{i,j}$ from equation 1, when the number is smaller than Q_0 . One of the clusters of K is chosen randomly, if the number is bigger than Q_0.

Dataset objects stand as $\{C_1, C_2, ..., C_N\}$.

$c_{i,v}$: i’th sample of the v’th attribute value.

$t_{i,j}$: The value is 1 when i’th object is in cluster j. The value is 0 when i’th object is not in cluster j. $i = 1,..,N, j = 1,..,K$

When the value of i'th model of the j'th cluster is regarding to the cluster then value becomes 0. When the value of the i'th model of the j'th cluster isn't regarding to the cluster then the value becomes 1. All these values are kept in Tij 's cluster that is NxK matrix.

Shehlokar is the method that is updated in thesis statement and nominal attributes will be handled as $n_{i,j}$.

$n_{i,j}$: i’th cluster’s j’th attribute's mode.

In this thesis statement; the distance metric is handled that has some differences from the method of Shelokar. Manhattan distance can be used for the measurements similarity between things.
Manhattan distance formula is defined as:

$$\left|x_{i,v} - m_{j,v}\right| \qquad \textbf{(3.2)}$$

Scatter seperability criterion provides cluster similarity and it is the aim of the clustering algorithm.

The equation above calculates the coherence of the results. Manhattan distance summation is among things and cluster centroids.

$$\text{Min F (w, m)} = \sum_{j=1}^{K} \sum_{i=1}^{N} \sum_{v=1}^{n} w_{i,j} \left|x_{i,v} - m_{j,v}\right|$$

This algorithm's following step is about finding the values that fits with the accurate solutions. In this step, local search is used to find the accurate solutions but is not used to find all of the solutions. %25 of the results are handled for local search in Shelokar's method. Local searching is preparing in ascending order from L top quality results.

p_{ls} : is the search chance and a permanent number among 0 and 1.

Local search routine is executed after the determination of the accurate results. Then L results are produced then the chance number is respectively given in nominee result.

The results are examined and then if the number that is chosen randomly is smaller than search threshold, i'th of the case is updated to cluster that can be made. On the other hand the result stand same.

According to this coincidental solution; factors are searched. This method can execute for all the nominee results. In addition the result is completed when the changing is finished due to these results. In conclusion, exchange process is executed between the solutions when accurate value is bigger than the original result.

Ant colony optimization's last step is about the exchange process. Searching the accurate results and comparing with the original values, ant colony optimization pheromone updates these values.

Pheromone matrix is effected from the best results. Pheromone matrix is updated from L results and L results are updated from the formula below.

$$\tau_{i,j}(t+1) = (1-k)\tau_{i,j}(t) + \sum_{l=1}^{L} \Delta\tau_{i,j}^{l}$$
$$i = 1,..,N, j = 1,..,K$$

k : in this equation is the pheromone vaporization rate.

There are new results that come into being slower while the vaporization parameter rises in examination. On the other hand parameters that are related with vaporization are declined.

$\Delta\tau_{i,j}^{l}$: Pheromone matrix's pheromone amounts.

Precipitate quantity is found as $1/G_l$, that this study's aim is decrement of G_l
Moreover, this method is executed to have the maximal repetitions of objects that are clustered.

3.2.3 SBS Algorithm

Checking learning for the selection of properties; provides consecutive reflux search which is a heuristic method [11]. The property subset values are eliminated according to the worst values according to consecutive reflux selection that begins with set which has all properties.

In conclusion searching is finished if there isn't any evolution on the property set.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_N \end{bmatrix} \xrightarrow{feature-selection} \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{iM} \end{bmatrix}$$

Figure 3.2 Feature selection

There is a procedure that is about assess. This method begins on the set that has full properties. The solutions are determined according to the this method. All of the properties are taken and new solutions are formed.

By using this method, there is always feature selection on each stairs of this method. One of the feature is picked up on each step. Therefore the accurate value is determined from the data that is property set.

The feature is selected due to the results related with highest accuration rate. This process is continued to find the last solution that has no innovation that can be determined.

3.2.4 Feature Selecting Ant Colony Clustering (FS-ACO CLUSTERING)

There are three parts of wrapper's algoritm. These parts are respectively property search , unsupervised learning and assessment algorithm. The first algorithm is SBS in this thesis statement. The second one is ACO clustering algorithm and the third one is about assessment algorithm. In addition, assessment function uses unified criterion approach.

3.2.5 Unified Evaluation Criteria

Property choice and clustering process are associated in Wrapper approach. These association process requires an evaluation procedure. There is only one assessment procedure that is used in this thesis statement for property selection and clustering process.

The unified criteria Scatter seperability index is as below:

S_w is about cluster object distances from cluster's centre.

S_b is about cluster object distances from cluster's intra.

$$S_w = \sum_{j=1}^{k} \sum_{i=1}^{N} w_{ij} \left| x_i - m_j \right|$$

m_{ij} = mode of the i'th attribute's mode value in the cluster j.

$$S_b = \sum_{i=1}^{k} w_{ij} \left| m_j - M_0 \right|$$

M_0 is the mode of the sample.

$$M_0 = \sum_{j=1}^{k} w_{ij} m_j$$

The scatter seperability index is :

$$trace(S_w^{-1} S_b)$$

If the number of parameters rise then this index rises. This procedure is not related with property selection. There is a penalty glossary to control and check the values assessment.

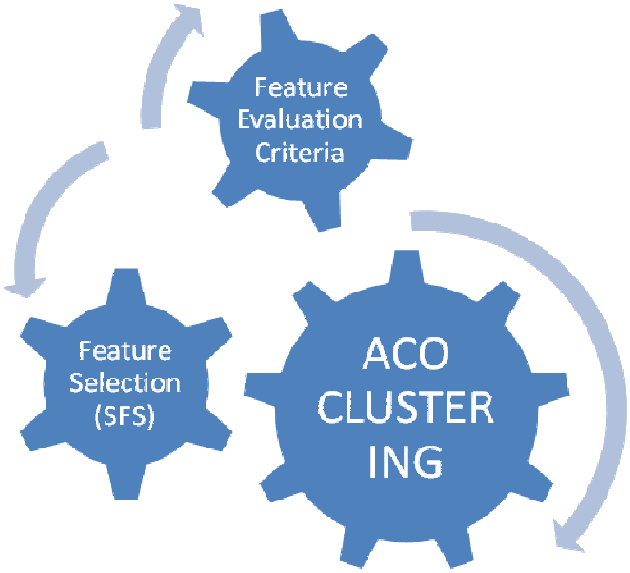


Figure 3.3 FS-ACO Clustering

Distance between objects to the center is defined as evaluation functions. Nominal attributes is used in FS-ACO clustering algorithm in this thesis statement. There is a process. Firstly, according to the format of data these attributes are taken from the record.

Secondly, in order to decrement of computational supplementary load of the method, these values are updated as binary values.

The agents respectively execute results and this is a process. The clusters are picked to calculate the centroids of them and then iteration performance is found by evaluation criteria.

Then, according to the results of that, there is a evaluation procedure to calculate the other results and the top results are chosen by this procedure. By this way the results are updated when the new results become more accurate with the oldest ones.

4. FINDINGS

This study's solution is clearly expounded in this part. The offering FS-ACO aggregation method's performance estimations on data are given.

The offered methods first section is the discretiazion of the data. The property of the data are transformed into binned assets. Numerical amounts are transformed into nominal features in respect of their distribution.

4.1 PARAMETER SETTINGS

The FS-ACO condensation algorithm is planned to have some multiples to be able to adjust the learning duration. It implies human interference to tune the multiples of the algorithm. First of all the pheromone vaporization ratio which makes the stability between exploration and the taking advantage of set to 0.01 as Dorigo operation on Dorigo's own book. (Dorigo,2004). The set of pheromone priori parameters is should be 0.1 which definition is ant to take the object together with maximum pheromone or different article. Therefore, this result is not get stuck to local is the lowest. If the parameter is minimum then the result may be stuck to local optimum. On the other hand if the parameters go up the solution might not converge [12].

The native research threshold macro is applied to definition the percentage of the samples in the solve to use the local research. The main solve rates are applied for pheromone update operation. Whether the number of the fittest solves are update the pheromone track matrix is maximum over, it happened baffling for FS-ACO algorithm to converge the very universal answer.

There are many several local research macros tunes used in trainings. FS-ACO is always developed by condensation tool is applied in this subject. The termination tests of the FS-ACO condensation algorithm is define as the highest number of iterations.

The number of effective influence the convergence of the solving. The effective numbers are rising and the algorithm converges quickly. Moreover if the digit of the effective rising too much, so the convergence rate rising, the inclination of converging to universal lowest decline. On the digit of feature to remove rising, to do studying algorithm more stable the digit of effective is rising.

The lowest digit of property the algorithm may be select also define by a macro.

4.2 TRAINING THE FS-ACO ALGORTIHM

Macros of the algorithm are optimized of the following different proving. The education dataset which is explained in before part, sampled in order to decrease the calculate complexity and education time. For example 280 samples are used for education of the FS-ACO condensation algorithm.

When they researching some sampling factor will be maximum skewed, first of all the factors are maximum skewed distribution are carried. Finally the result is 20 features are used for the beginning education of the data.

In the data set increase and there are have some overlap problem. These problems are caused suitability function to be zero and it take the local 0(zero). This case is solution by joining the term 1 to the suitability function.

First practice, FS-ACO condensation algorithm carrying 1 of the main quality form the education set.

Macros setting for the practice are given below.

Table 4.1 Macros for one characteristic removal

Parameter	Value
total number of attributes	20
total number of samples	280
total number of clusters	5
total number of agents	50
Pheromone	0,001

The Prior Pheromone	0,1
Evoporation of Pheromone	0,01
The Threshold Search	0,01
Number of Search	10
Total Iteration Numbers	5000
Selection of Property	19

Considering the highest repetition digit is given as 5000 the algorithm converges before reaching the highest number of repetition is given maximum, so the reduce the possibility of premature convergence.

Centroids of the groups are given in this table:

Table 4.2 Groups centroids and used property for one property removal

	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4	Chosen
USAGE OF C1	F	F	F	F	F	1
USAGE OF C2_1	T	T	F	T	F	1
USAGE OF C2	F	F	T	T	F	1
CUSTOMERID	F	F	F	F	F	1
AGE	F	F	F	F	F	1
INCOME	F	F	F	T	T	1
TOLLFREE	F	F	F	F	F	1
LONGTEN	F	F	F	F	T	1
WIREMON	F	F	T	T	T	1
TENURE	F	F	F	F	F	1
TOLLTEN	F	F	F	T	F	1
PERS_LOAN	H	L	0.0	H	0.0	0
PRD3_AMT	H	0.0	0.0	0.0	0.0	1
PRDTOT_AMT	0.0	0.0	L	H	H	1
PRDTOT2_AMT	0.0	0.0	L	H	H	1
INSURANCE_AMT	0.0	0.0	L	H	H	1
PRDTOT_DEBT	F	F	F	F	F	1
INSURANCE_DEBT	F	F	F	F	F	1
LOAN_AMT	F	F	F	T	T	1
PRD5_AMT	F	F	F	F	T	1

Centroids deliver us about the feature of the client profiles.

Set 0 is a one product and one channel using client profile, on the other hand having big amount on one produce.

Set 1 is also a one produce user batch, however it has a minimum amount saving tendency in the produce.

Set 2 profile clients make dissimilar channel and has more produce penetration in finally. It have minimum amount of money in those produces.

Set 3 profile client are the very active users. These customers have several types of products, so they employ all channels. Clients quantity of money in the producers is also maximum in all subjects.

Set 4 profile clients have a maximum produce penetration. Customers selection of produces are several from the client 3 profile. Moreover clients also keep high amount of money in those produce.

The measure of the bundles are given down. It should be told that the frequency of the bundles about evenly distributed.

Table 4.3 Set distribution for one characteristic removal

Clusters	**Frequency**
Percent of Cluster 0	64 (%22.85)
Percent of Cluster 1	45 (%16.07)
Percent of Cluster 2	86 (%30.71)
Percent of Cluster 3	36 (%12.85)
Percent of Cluster 4	49 (%17.5)

The other second testing, 3 characteristic are chosen to be removed from the characteristic cluster. The macros setting for this proving is given down table.

Table 4.4 Macros for three characteristic removal

Parameter	Value
total number of attributes	20
total number of samples	280
total number of clusters	5
total number of agents	50
Pheromone	0,001
The Prior Pheromone	0,1
Evoporation of Pheromone	0,01
The Threshold Search	0,01
Number of Search	10
Total Iteration Numbers	4000
Selection of Property	17

This table point the centroids of the set and the characteristic selected in this instance.

Table 4.5 Selected characteristic for three characteristic removal

ATTRIBUTE	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4	Chosen
USAGE OF C1	F	F	F	F	F	1
USAGE OF C2_1	T	F	T	T	T	1
USAGE OF C2	F	F	F	T	T	0
CUSTOMERID	F	F	F	F	F	1
AGE	F	F	F	F	F	1
INCOME	T	F	F	F	F	1
TOLLFREE	F	F	F	F	F	1
LONGTEN	F	F	F	F	F	1
WIREMON	T	T	F	F	F	1
TENURE	F	F	F	F	F	1
TOLLTEN	F	F	F	F	F	1
PERS_LOAN	0.0	0.0	L	H	H	0
PRD3_AMT	0.0	0.0	0.0	H	H	0
PRDTOT_AMT	H	L	0.0	0.0	0.0	1
PRDTOT2_AMT	H	L	0.0	0.0	0.0	1
INSURANCE_AMT	H	L	0.0	0.0	0.0	1
PRDTOT_DEBT	F	F	F	F	F	1
INSURANCE_DEBT	F	F	F	F	F	1
LOAN_AMT	T	F	F	F	F	1

According the graphics deliver us insights concerning the FS-ACO sets algorithm's wrapper features.

If the x axis is the characteristics in range [0-20], y-axis is the number of repetition and the z axis given us if the characteristics used or not with Boolean values (0,1).

It should be made that in the first repetition the 10, 13, 15 indexed characteristics were used. As the repetition continues and several characteristics cluster become the optimal characteristics set. The characteristics set also gives directions to the optimality of the answer.

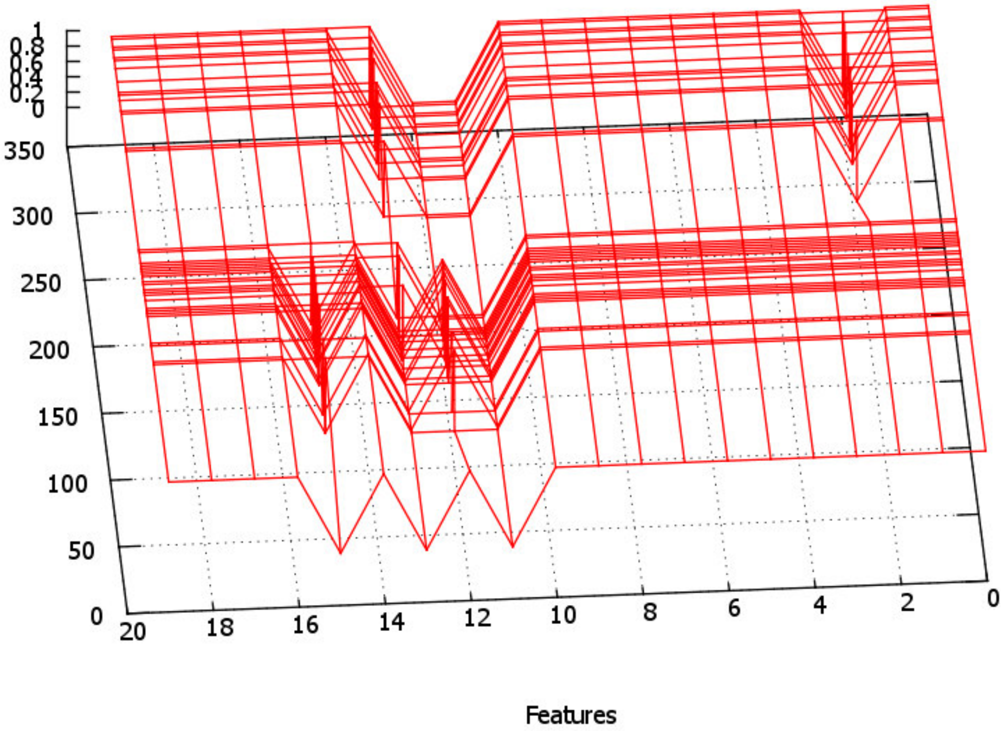


Figure 4.1 Selected characteristics vs. Repetition for 1 characteristics removal

Frequency of the calculated clusters table is seen below.

Table 4.6 Cluster distribution for one feature removal

Cluster	Frequency
Percent of Cluster 0	81 (%28.92)
Percent of Cluster 1	88 (%31.42)
Percent of Cluster 2	52 (%18.57)
Percent of Cluster 3	25 (%8.92)
Percent of Cluster 4	34 (%12.14)

5 properties are deleted in the third experiment from property set. Moreover, there are a lot of iterations that provides to cognize the complication of property space, so the maximum iterations macro rises.

Table 4.7 Parameters for five feature removal

Parameter	Value
total number of attributes	20
total number of samples	280
total number of clusters	5
total number of agents	50
Pheromone	0,001
The Prior Pheromone	0,1
Evoporation of Pheromone	0,01
The Threshold Search	0,01
Number of Search	10
Total Iteration Numbers	6000
Selection of Property	15

Property selection and Centroids of clusters are seen below.

Table 4.8 Selected features for five feature removal

ATTRIBUTE	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4	Chosen
USAGE OF C1	F	F	F	F	F	1
USAGE OF C2_1	T	T	F	T	T	1
USAGE OF C2	F	F	F	T	T	1
CUSTOMERID	F	F	F	F	F	1
AGE	F	F	F	F	F	1
INCOME	F	F	T	F	F	1
TOLLFREE	F	F	F	F	F	1
LONGTEN	F	F	F	F	F	1
WIREMON	F	T	T	F	F	1

TENURE	F	F	F	F	F	1
TOLLTEN	F	F	F	F	F	1
PERS_LOAN	L	H	0.0	M	H	0
PRD3_AMT	0.0	L	0.0	H	H	0
PRDTOT_AMT	0.0	L	H	0.0	0.0	0
PRDTOT2_AMT	0.0	L	L	0.0	0.0	0
INSURANCE_AMT	0.0	L	L	0.0	0.0	0
PRDTOT_DEBT	F	F	F	F	F	1
INSURANCE_DEBT	F	F	F	F	F	1
LOAN_AMT	F	F	T	F	F	1
PRD5_AMT	F	F	F	F	F	1

Figure indicate the iterations about the alternation of selected property.

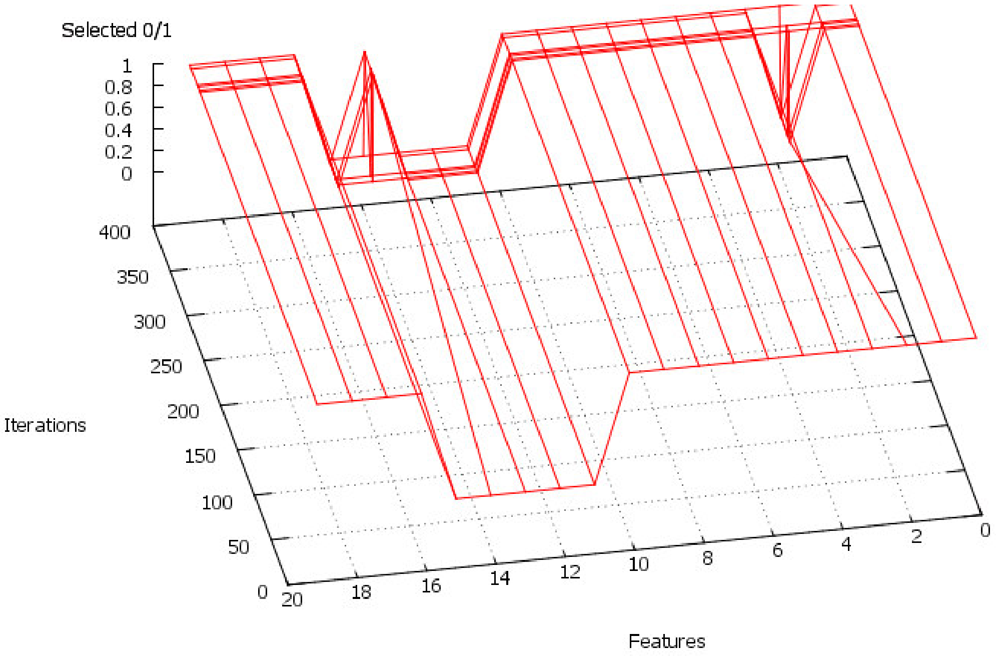


Figure 4.2 Selected features vs. Iterations for 3 feature removal

Frequency of the calculated clusters table is seen below.

Table 4.9 Cluster distribution for five feature removal

Cluster	**Frequency**
Percent of Cluster 0	69 (%24.64)
Percent of Cluster 1	26 (%9.28)

Percent of Cluster 2	113 (%40.35)
Percent of Cluster 3	41 (%14.64)
Percent of Cluster 4	31 (%11.07)

7 properties are deleted in the fourth experiment from property set. Moreover, there are 6000 iterations which is the maximum number which provides to cognize the complication of property space. The macro setting is seen below.

Table 4.10 Parameters for eight feature removal

Parameter	**Value**
total number of attributes	20
total number of samples	280
total number of clusters	5
total number of agents	50
Pheromone	0,001
The Prior Pheromone	0,1
Evoporation of Pheromone	0,01
The Threshold Search	0,01
Number of Search	10
Total Iteration Numbers	6000
Selection of Property	13

Property selection and Centroids of clusters are seen below.

Table 4.11 Selected features for eight feature removal

ATTRIBUTE	**cluster 0**	**cluster 1**	**cluster 2**	**cluster 3**	**cluster 4**	**Chosen**
USAGE OF C1	F	F	F	F	F	1
CUSTOMERID	F	F	F	F	F	1
TENURE	F	F	F	F	F	1
TOLLTEN	F	F	F	F	F	1
AGE	F	F	F	F	F	1
INCOME	F	F	F	T	F	1
LONGTEN	F	F	F	F	F	1
WIREMON	F	T	T	T	F	1
INSURANCE_DEBT	F	F	F	F	F	1
PRD5_AMT	F	F	F	F	F	1
PRDTOT_AMT	0.0	L	L	H	0.0	1
PRDTOT_DEBT	F	F	F	F	F	1
PRDTOT2_AMT	0.0	L	L	H	0.0	1
USAGE OF C2_1	T	T	F	T	T	0
USAGE OF C2	F	F	T	F	F	0
TOLLFREE	F	F	F	F	F	0

INSURANCE_AMT	0.0	L	L	H	0.0	0
LOAN_AMT	F	F	F	T	F	0
PERS_LOAN	L	M	0.0	0.0	H	0
PRD3_AMT	0.0	0.0	0.0	0.0	H	0

Frequency of the calculated clusters table is seen below.

Table 4.12 Cluster distribution for five feature removal

Cluster	**Frequency**
Percent of Cluster 0	54 (%19.28)
Percent of Cluster 1	22 (%7.85)
Percent of Cluster 2	84 (%3)
Percent of Cluster 3	66 (%23.57)
Percent of Cluster 4	54 (%19.28)

There are fifth experience and according to this experience the feature number is updated as 9. The problems are more impure as the property number rises. As a result agent number is risen to 70 as seen below.

Table 4.13 Parameters for eight feature removal

Parameter	**Value**
total number of attributes	20
total number of samples	280
total number of clusters	5
total number of agents	**70**
Pheromone	0,001
The Prior Pheromone	0,1
Evoporation of Pheromone	0,01
The Threshold Search	0,01
Number of Search	14
Total Iteration Numbers	4000
Selection of Property	11

Property selection and Centroids of clusters are seen below.

Table 4.14 Selected features for eight feature removal

ATTRIBUTE	**cluster 0**	**cluster 1**	**cluster 2**	**cluster 3**	**cluster 4**	**Chosen**
USAGE OF C1	F	F	F	F	F	1
CUSTOMERID	F	F	F	F	F	1
AGE	F	F	F	F	F	1

INCOME	F	F	F	T	T	1
TOLLFREE	T	F	F	F	F	1
LONGTEN	F	F	F	T	F	1
WIREMON	T	T	F	T	T	1
PRDTOT_AMT	L	L	0.0	H	H	1
INSURANCE_DEBT	F	F	F	F	F	1
LOAN_AMT	F	F	F	T	T	1
PRD5_AMT	F	F	F	T	F	1
USAGE OF C2_1	T	T	T	T	T	0
USAGE OF C2	T	F	F	F	T	0
TENURE	F	F	F	F	F	0
TOLLTEN	F	F	F	F	F	0
PERS_LOAN	0.0	0.0	L	0.0	0.0	0
PRD3_AMT	0.0	0.0	0.0	0.0	0.0	0
PRDTOT2_AMT	L	L	0.0	H	H	0
INSURANCE_AMT	L	L	0.0	H	H	0
PRDTOT_DEBT	F	F	F	F	F	0

Calculated frequency of the clusters are seen below.

Table 4.15 Cluster distribution for eight feature removal

Cluster	**Frequency**
Percent of Cluster 0	52 (%18.57)
Percent of Cluster 1	42 (%15)
Percent of Cluster 2	105 (%37.5)
Percent of Cluster 3	34 (%12.14)
Percent of Cluster 4	47 (%16.78)

Classification is being done by Weka which is a data mining tool,with naive bayesian method. Naive bayesian algorithm produces ten rules for classification. There is a list below that includes the rules.

Rule 1: This rule contains the customers age which are under eighteen or over eighteen years old. If the customers are under eighteen years old, the company lose their customers easily. If the customers are older than eighteen years old, the company don't lose their customers.

Rule 2: This rule contains the customers age which are under fifty years old and over thirty years old. If the customers are over thirty years old and below fifty years old, the company lose their customers easily. If the customers over fifty years old, the company don't lose their customers.

Rule 3: This rule contains the customers age and gender information which are between thirty and fifty years and male or female informations. If the customer is under fifty years old and over thiry years old and gender is female; the company lose their customers when there is a limit under their income on their credit cards. This company has to give credit limit over their incomes.

Rule 4: This rule contains the customers income and monthly expenses. If the customer's income is more than his or her monthly expense; the bank should prepare some campaigns to this group customers not to lose them.

Rule 5: This rule contains the customers region, age, gender, income. If the customer lives in east and between eighteen, thirty years old, female and has the salary between two thousand and five thousand turkish liras, then the company should provide him or her credit over his or her income. If the customer lives in west and between eighteen thirty years old, female and has the salary between two thousand and five thousand turkish liras, the company doesn't have to provide credit over his or her income.

Rule 6: This rule contains information about customers maritual status and gender. If the customer is married and male, the company doesn't have to provide some extra campaigns but if the customer is married and female, the company has to provide extra campaign not to lose their customers.

Rule 7: This rule contains information about customers tenure. If the customer works below five years at the same company, the campaign doesn't need to provide credit campaigns. But if the customer works between two and five years at the same company has to provide extra campaigns to this group of customers. Because, this type of customers are not reliable to all banks so they prefer the banks that provides them credit campaigns.

Rule 8: This rule contains the information about customers age and retirement. If the customers are over fifty years old and retired, the company should provide some campaigns which includes extra installments.

Rule 9: This rule contains the information about customers credit card number and limits. If the customer has more than two credit cards and each of them has high limits, the company should provide extra bonus campaigns for their customers to assess this group of customers.

Rule 10: This rule contains the information about customers' bank information. If the customer is in the same bank below five years, the customers think to change their bank's, so it should have some extra campaigns to customers which are in their banks below five years.

4.3 FS-ACO CLUSTERING VS. WEKA COMPARISON

K-means algorithm is utilized to evaluate the wrapper algorithm with filtering methods. By using K-means then algorithm all the dataset is aggregated to provide the properties. Clustering attribute chosen labels is utilized to the data. Weka property chosen is utilized to get ahead this situation. Property ranking solutions are seen below.

Table 4.16 Feature Ranking for Weka feature selection

ATTRIBUTE	RANK
PRDTOT2_AMT	1.03
INSURANCE_AMT	0.9634
PRDTOT_AMT	0.9469
WIREMON	0.7686
PERS_LOAN	0.4788
USAGE OF C2	0.4426
TOLLFREE	0.4323
LOAN_AMT	0.3719
INCOME	0.37
PRD3_AMT	0.3363
LONGTEN	0.2329
PRD5_AMT	0.2329

USAGE OF C2_1	0.2283
TENURE	0.1064
PRDTOT_DEBT	0.0648
TOLLTEN	0.0487
RECORD1	0.0469
INSURANCE_DEBT	0.0439
AGE	0.0439
USAGE OF C1	0.0392

The same number of the attributes are deleted to detect the comparison of the results of WEKA and FS-ACO clustering algorithms. The Weka clustering solutions related with frequency distributions are seen as below.

Table 4.17 Cluster distribution for Weka clustering

# Of Attributes	**19**	**17**	**15**	**13**	**11**
Sum Error	793	745	682	612	491
Percent of Cluster 0	74 (26%)	74 (26%)	74 (26%)	74 (26%)	74 (26%)
Percent of Cluster 1	48 (17%)	48 (17%)	48 (17%)	48 (17%)	44 (16%)
Percent of Cluster 2	53 (19%)	53 (19%)	53 (19%)	53 (19%)	70 (25%)
Percent of Cluster 3	61 (22%)	61 (22%)	61 (22%)	61 (22%)	62 (22%)
Percent of Cluster 4	44 (16%)	44 (16%)	44 (16%)	44 (16%)	30 (11%)

The summation of the failures to the properties that are deleted is seen as below. Manhattan distance among objects and centroids as definesd as summation of the errors provides decrement of the number of the properties. Attribute determination with Weka is used by FS-ACO to determine clustering algorithm.

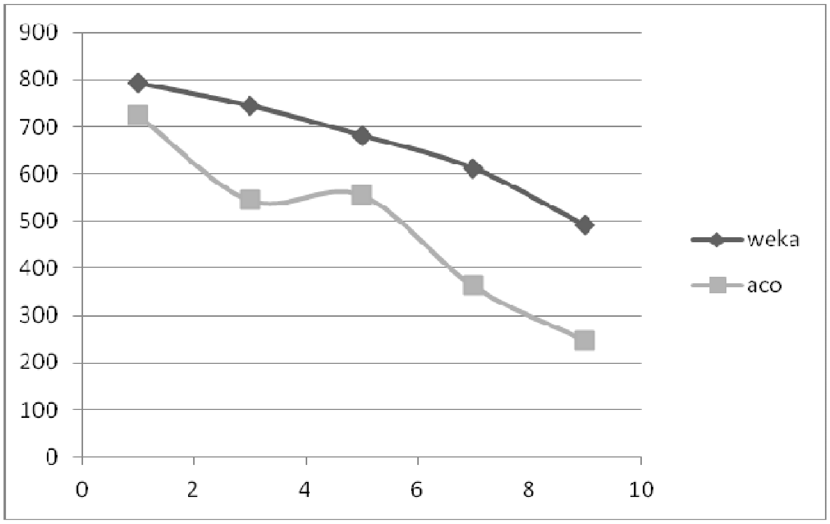
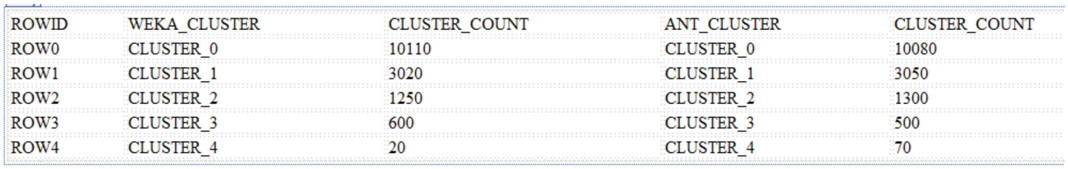


Figure 4.3 Comparison of FS-ACO vs WEKA

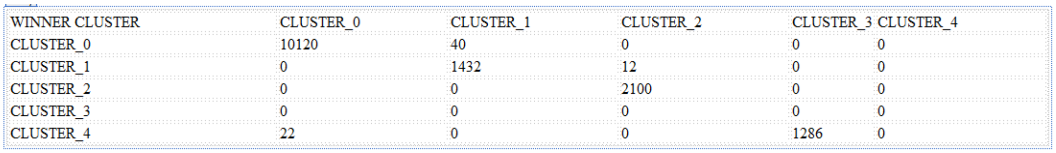
Ant Colony Algorithm and Weka Clustering results are seen below. "WEKA_CLUSTER" shows the weka clustering algorithm and the first "CLUSTER_COUNT" shows the number of objects in the clusters. "ANT_CLUSTER" shows Ant Colony Optimization clustering algorithm and the second "CLUSTER_COUNT" shows the number of objects in the cluster ant colony.



ROWID	WEKA_CLUSTER	CLUSTER_COUNT	ANT_CLUSTER	CLUSTER_COUNT
ROW0	CLUSTER_0	10110	CLUSTER_0	10080
ROW1	CLUSTER_1	3020	CLUSTER_1	3050
ROW2	CLUSTER_2	1250	CLUSTER_2	1300
ROW3	CLUSTER_3	600	CLUSTER_3	500
ROW4	CLUSTER_4	20	CLUSTER_4	70

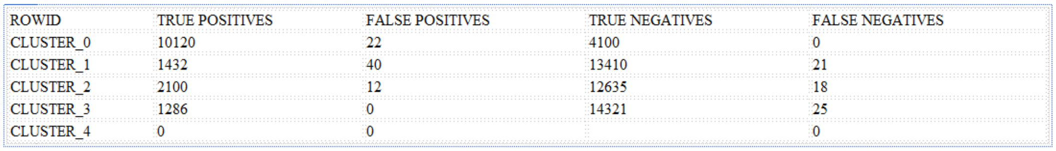
Figure 4.4 : The Distribution of Clusters; Ant Colony, Weka

The following graphs shows the accuracy of two different algorithms.



WINNER CLUSTER	CLUSTER_0	CLUSTER_1	CLUSTER_2	CLUSTER_3	CLUSTER_4
CLUSTER_0	10120	40	0	0	0
CLUSTER_1	0	1432	12	0	0
CLUSTER_2	0	0	2100	0	0
CLUSTER_3	0	0	0	0	0
CLUSTER_4	22	0	0	1286	0

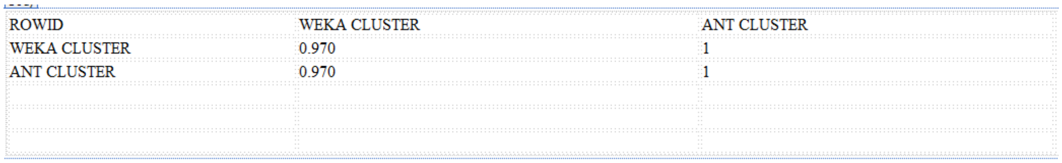
Figure 4.5 : Different Algorithms Accuracy Ant Colony and Weka



ROWID	TRUE POSITIVES	FALSE POSITIVES	TRUE NEGATIVES	FALSE NEGATIVES
CLUSTER_0	10120	22	4100	0
CLUSTER_1	1432	40	13410	21
CLUSTER_2	2100	12	12635	18
CLUSTER_3	1286	0	14321	25
CLUSTER_4	0	0		0

Figure 4.6 : Ant Colony and Weka Clustering Variables ; The True Positives and False Positives, True Negatives and False Negatives

Linear correlation between Ant Colony and Weka clustering results is seen below.



ROWID	WEKA CLUSTER	ANT CLUSTER
WEKA CLUSTER	0.970	1
ANT CLUSTER	0.970	1

Figure 4.7 : Linear Correlation Between Ant Colony and Weka Clustering Results

5. DISCUSSION AND CONCLUSIONS

Clustering is one of the most important data mining problem that is studied in this thesis statement. There are lots of techniques are implemented for data sets to provide the most suitable parts of data.

Unsupervised property elimination is a technique to these problems. Unsupervised property selection has two approaches that are respectively Filter and Wrapper. Moreover ,there is better accuracy in Wrapper approach.

Ant Colony Optimization clustering is used with SBS method for the elimination of properties of data in this thesis statement. Clustering algorithm and property elimination are joined and Wrapper approach comes into being by this way.

Shelokar's ant colony clustering model contains ant colony optimization method that is utilized in this thesis statement. Manhattan distance is provided from the distance metric by using nominal attributes. Moreover, features mean values are utilized. CRIT criteria application is provided innovation by the original function.

There isn't any improvements in ant colony optimization and SBS wrapper approach before this thesis statement. The most accurate solutions is utilized to determine the features and separation of data sets.

There is respectively an iteration according to determine to find data sets. In all steps SBS algorithm is utilized. All of the aims of this thesis statement is to determine the recoup the customers. There is a process and this process is related with choosing the data and converting them from numerical to nominal entities.

There are two hundred and eighty items that are illustrated in this thesis statement. FS-ACO clustering uses this data. In conclusion, we can provide customer's characteristics according to this results.

In conclusion, FS-ACO algorithm is used for evaluation of the results as filtering approach. Property set is clustered. The attributes are deleted according to their ranks in the related algorithm. Label's of the class are utilized to gain information by using Weka ranking algorithm.

Finally, all of the error's from the output of these algorithms are collated and K-means algorithm is outperformed according to FS-ACO algorithm.

REFERENCES

Chris Rygielski, Jyun-Cheng Wang, David C. Yen, *Data mining techniques for customer relationship management*.

Jardine and Sibson, 1971 Jardine, N. and Sibson, R. (1971) *Mathematical Taxonomy*. Wiley, London.

Ying Zhao, George Karypis and Usama Fayyad, *Hierarchical Clustering Algorithms for Document Datasets*.

Zeng H., Cheung Y. M., *A new feature selection method for Gaussian mixture clustering*. Pattern Recognition, Volume 42, Issue 2, February 2009, pp. 243-250

Ant Colony Optimization for Feature Subset Selection, Ahmed Al-Ani.

Turning telecommunications call details to churn prediction: a data mining approach Chih-Ping Wei, Chih-Ping Wei

Data Cleaning: Problems and Current Approaches Erhard Rahm Hong Hai Do University of Leipzig, Germany

Data mining: practical machine learning tools and techniques, Ian H. Witten,Eibe Frank

Data Mining Discretization Methods and Performances Z. Marzuki, F. Ahmad

Marco D. 2004 *Ant Colony Optimization*. Massachusetts Institute of Technology ISBN 0-262-04219-3

Xiao-bin Z., Feng G., & Hui H. *Customer-churn Research Based on Customer Segmentation.* Electronic Commerce and Business Intelligence, ECBI 2009. International Conference on 2009, pp. 443 – 446

Anderson ET. *Sharing the wealth: when should firms treat customers as partners?*.
Management Science 2002;48(8): 955–71

Xiao-bin Z., Feng G., & Hui H. *Customer-churn Research Based on Customer Segmentation.* Electronic Commerce and Business Intelligence, ECBI 2009. International Conference on 2009, pp. 443 – 446

Zeng H., Cheung Y. M., *A new feature selection method for Gaussian mixture clustering.* Pattern Recognition, Volume 42, Issue 2, February 2009, pp. 243-250

Chow, T.W.S., Wang P., Ma E.W.M., *A New Feature Selection Scheme Using a Data Distribution Factor for Unsupervised Nominal Data*. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions Vol. 38 , Issue: 2 2008 , pp. 499 – 509

Cotter S.F., Kreutz-Delgado K., & Rao B.D. *Backward sequential elimination for sparse vector selection*. Signal Processing 81 (2001). pp. 1849–1864

Tsai C.F., & Lu Y.H., *Customer churn prediction by hybrid neural network.* Expert Systems with Applications, Volume 36, Issue 10, December 2009, pp. 12547-12553

Shin-Yuan Hung, David C. Yen, Hsiu-Yu Wang, *Applying data mining to telecom churn management.*

C.V.

Name Surname	: Batuhan .GÜLLÜOĞLU
Address	: Süleyman Demirel cd. Erguvan Evleri C Blok. No:15 Esenkent / İstanbul
Place of Birth, Year of Birth	: Fatih, 1985
Foreign Language	: English
Primary School	: CUMHURİYET İLKÖĞRETİM OKULU, 1996
Middle School	: CUMHURİYET İLKÖĞRETİM OKULU, 1999
Highschool	: KDZ. EREĞLİ ANADOLU 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 – (continued)