# DECISION MAKING SYSTEM ALGORITHM ON MENOPAUSE DATA SET

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

HİKMET ÖZGE BACAK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

SEPTEMBER 2010

Approval of the thesis:

**DECISION MAKING SYSTEM ALGORITHM ON MENOPAUSE DATA SET**

submitted by **HİKMET ÖZGE BACAK** in partial fulfilment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen ----------
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İsmet Erkmen ----------
Head of Department, **Electrical and Electronics Engineering**

Prof. Dr. M. Kemal Leblebicioğlu ----------
Supervisor, **Electrical and Electronics Engineering Dept., METU**

Assist. Prof. Dr. İlkay Ulusoy ----------
Co-Supervisor, **Electrical and Electronics Engineering Dept., METU**

**Examining Committee Members:**

Prof. Dr. Uğur Halıcı ----------
Electrical and Electronics Engineering Dept., METU

Prof. Dr. M. Kemal Leblebicioğlu ----------
Supervisor, Electrical and Electronics Engineering Dept., METU

Prof. Dr. Sinan Beksaç ----------
Faculty of Medicine, Hacettepe University

Assoc. Prof. Dr. Aydın Alatan ----------
Electrical and Electronics Engineering Dept., METU

Assist. Prof. Dr. Emre Tuna ----------
Electrical and Electronics Engineering Dept., METU

**Date:** 02.09.2010

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Last name: Hikmet Özge BACAK

Signature:

**ABSTRACT**

DECISION MAKING SYSTEM ALGORITHM ON MENOPAUSE DATA SET

Bacak, Hikmet Özge

Ms.C., Department of Electrical and Electronics Engineering

Supervisor: Prof. Dr. M. Kemal Leblebicioğlu

Co-Supervisor: Assist. Prof. Dr. İlkay Ulusoy

September 2010, 118 pages

Multiple-centered clustering method and decision making system algorithm on menopause data set depending on multiple-centered clustering are described in this study. This method consists of two stages. At the first stage, fuzzy C-means (FCM) clustering algorithm is applied on the data set under consideration with a high number of cluster centers. As the output of FCM, cluster centers and membership function values for each data member is calculated. At the second stage, original cluster centers obtained in the first stage are merged till the new numbers of clusters are reached. Merging process relies upon a "similarity measure" between clusters defined in the thesis. During the merging process, the cluster center coordinates do not change but the data members in these clusters are merged in a new cluster. As the output of this method, therefore, one obtains clusters which include many cluster centers.

In the final part of this study, an application of the clustering algorithms – including the multiple centered clustering method – a decision making system is constructed using a special data on menopause treatment. The decisions are based on the clusterings created by the algorithms already discussed in the previous chapters of the thesis. A verification of the decision making system /

decision aid system is done by a team of experts from the Department of Department of Obstetrics and Gynecology of Hacettepe University under the guidance of Prof. Sinan Beksaç.

*Keywords: Multiple-centered clustering, decision making system algorithm, menopause, fuzzy C-means clustering, hard C-means clustering, K-means clustering, similarity based clustering, missing data*

# ÖZ

MENOPOZ VERİLERİ HAKKINDA KARAR VEREBİLEN
BİR SİSTEMİN GELİŞTİRİLMESİ ALGORİTMASI

Bacak, Hikmet Özge

Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü

Tez Yöneticisi: Prof. Dr. M. Kemal Leblebicioğlu

Ortak Tez Yöneticisi: Yrd. Doç. Dr. İlkay Ulusoy

Eylül 2010, 118 sayfa

Bu çalışmada çok merkezli kümeleme yöntemi ve bu yöntemin sonucunda menopoz verileri hakkında karar verebilen bir sistemin geliştirilmesinden bahsedilmiştir. Bu yöntem iki bölümden oluşmaktadır. İlk bölümde, yüksek sayıda küme sayısı seçilerek veri setine bulanık C-ortalamalar (FCM) kümeleme yöntemi uygulanmaktadır. FCM'nin sonucunda küme merkezleri ve her verinin aitlik fonksiyon değeri hesaplanmaktadır. İkinci bölümde ise, ilk bölümde elde edilen küme merkezleri, kullanıcı tarafından belirlenen en son küme adedine erişinceye kadar küme birleştirme işlemi uygulanır. Bu birleştirme işlemleri, kümeler arasındaki "benzerlik ölçütü"'ne dayanmaktadır. Birleştirme işlemleri sırasında küme merkezlerinde bir değişiklik olmazken, birleşen kümelere ait veriler yeni bir küme altında birleşirler. Böylece bu yöntemin sonucunda, birden fazla merkeze sahip kümeler elde edilmektedir.

Bu çalışmanın son kısmında – çok merkezli öbekleme algoritması da dahil olmak üzere öbekleme algoritmalarının bir uygulaması olarak – menapoz tedavisinden elde editen özel bir veri kümesi üzerinden bir karar verici sistem tasarlanmıştır. Kararlar bu tezin daha önceki bölümlerinde anlatılan algoritmaların elde ettiği öbeklemeler kullanılarak elde edilmektedir. Karar verici / karar vermeye yardımcı olan bu sistemin bir doğrulaması ise başında Prof. Dr. Sinan Beksaç'ın bulunduğu Hacettepe Üniversitesi Kadın Hastalıkları ve Doğum anabilim dalından bir uzman grubu tarafından yapılmıştır.

*Anahtar Kelimeler: Çok merkezli kümeleme, karar verebilen sistemin geliştirilmesi algoritması, menopoz, bulanık C-ortalamalar, katı C-ortalamalar, K-ortalamalar kümelemesi, benzerlik tabanlı kümeleme, eksik veriler*

*To My Beloved Family…*

**ACKNOWLEDGMENTS**

I would like to express my deepest gratitude to my supervisor, Prof. Dr. Kemal Leblebicioğlu, for his knowledge transfer, guidance, advices and helpfulness throughout this research work. I would like to thank my supervisor especially for his valuable contributions to my career and show me the way to follow the best. His encouragement and continuous support throughout this research work are the main reasons for the success of the study.

I wish to express my special thanks to Prof. Dr. Sinan Beksaç for his valuable comments and spending his valuable time for my thesis study.

I also would like to thank my co-supervisor, Assist. Prof. Dr. İlkay Ulusoy for her helps and advices during my thesis duration.

I wish to express my special thanks to research assistant Örsan Aytekin for his technical support, helpfulness and friendship. I would like to thank him especially for his critical advices, encouragement and continuous support.

I would like to thank Burak Durmaz, for his continuous encouragement, great patience, trust and support during the whole process.

I also would like to thank all other my friends who have been with me during this period. Their friendship, their encouragement and their trust in me are always very important for me.

And finally, I would like to thank my beloved family, my father Cengiz, my mother Sezaver and my brother Özkan for their continuous support, but especially for their trust in me. I always feel very lucky to be with them, I love you!!

## TABLE OF CONTENTS

## LIST OF TABLES

TABLES

## LIST OF FIGURES

FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1. Problem Definition

Humans have an ability to make decisions by using their emotions, experiences and intelligence. These specialties are the result of being human. Since machines do not have these specialties, by using hardware and software capabilities of machines, people are trying to furnish some of these abilities on machines. In general, there are machines / systems produced and designed to help people to make decisions and give recommendations for their work. Because of the success of these systems and especially the real time decision making requirements in some fields, decision making and recommendation systems were developed.

Development of decision making systems has an important place in the artificial intelligence field. In these days, the subject of designing computer programs, machines or robots are very popular. Every machine, every robot or every algorithm that has an ability to make a decision, is called as a “decision making system”. These systems behave or decide similar to the person who designs them. The commands produced by these systems are made of some theoretical and mathematical formulations and explanations.

In this study, decision making processes are designed depending on the clusters in the data sets. Consequently, as a natural result of clustering processes, the system gains an ability to produce a decision by using distance measure between the data points and cluster centers. Therefore, the main work in this study is to design a robust clustering algorithm.

Clustering is one of the most important unsupervised learning techniques in data mining [11], [24]. Due to its nature of being unable to produce a ground truth, this problem is still one of the most studied in data mining field. Clustering analysis is used in many fields, such as statistical analysis, machine learning, image and pattern recognition, medical scanning areas.

In clustering analysis, the aim is to find the similar and dissimilar data groups in the data sets. Similar data members should be located in the same group and dissimilar data members should be located in different groups. The similarity between data members and cluster centers is measured by using distance metrics. There are many distance metrics used for the clustering analysis. The most used metrics can be defined as the following:

- Euclidean distance metric,
- Manhattan distance metric,
- Mahalonobis distance metric,
- Hamming distance metric.

The results of the clustering are affected from the use of different distance metrics. So, the best distance metric for the clustering can be determined by clustering validity.

In general, clustering methods are sensitive to the initialization of the cluster centers and total number of cluster centers. The initialization part causes big problems for the clustering since different initializations result in different cluster centers. For every different initialized cluster centers, algorithms may find different cluster center coordinates. This fact is the main disadvantage of clustering algorithms. Because of this, many initialization techniques were produced for the clustering analysis. After a good initialization, the algorithms can find better results and perform efficient clustering.

The clustering algorithms discussed so far are single-centered. Even if there are relatively dissimilar data or data groups in the same cluster, they are connected to only one center in the cluster. To specify dissimilar data groups in the same cluster, the proposed multiple centered fuzzy clustering algorithm is constructed to find out different cluster centers in the same cluster group. Furthermore, and maybe more importantly, the proposed algorithm is able to distinguish clusters having a non-convex structure. This feature of the proposed algorithm makes it very useful in the identification of clusters which are usually impossible to be constructed with classical clustering algorithms since they employ mostly Euclidean based distance measures.

In thesis study, multiple-centered fuzzy clustering (MCFC) based decision making system algorithm is studied in detail. The main contribution of this study is the design of the MCFC algorithm. After designing of MCFC as an extension of the standard fuzzy c-means (FCM) clustering algorithm, a decision making system is proposed. MCFC method depends on the FCM algorithms which are described in Chapter 2 in details.

MCFC algorithm includes a standard FCM clustering algorithm in its initialization procedure. FCM method can be applied on a complete data set (i.e., data vectors are not supposed to have missing components). So, if data vectors include missing components, they should be imputed by various methods of imputation. After imputation, the data set can be used in FCM method.

In many data sets, components of the data vectors may range in very large and different interval. This condition has adversities on clustering results. To eliminate this condition, normalization techniques should be applied to data sets. In the normalization process, the aim is to arrange the data set components in a similar range of numbers. After normalization, data set is ready for the clustering process. It has been frequently observed that, clustering algorithms produce more sensible results when compared to clustering without normalization.

In the MCFC algorithm, FCM algorithm with a high number of cluster centers is applied on the data set. FCM algorithm is very sensitive to initialization of the cluster centers. The initial cluster centers are assigned by using fuzzy ant initialization algorithm. After fuzzy ant initialization, the standard FCM algorithm is performed on the data set. Then, membership function values of all data members and cluster centers are found and recorded. They will be utilized throughout the MCFC algorithm application. Next, the MCFC algorithm is applied on the data set with a selected number of final clusters. In order to apply the MCFC algorithm efficiently and robustly, a similarity threshold should be defined for the merging processes between the clusters. During merging processes, the data members in the clusters are merged but the coordinates of the cluster centers do not change. So, the new clusters include more than one cluster center.

After the development of the MCFC algorithm, a decision making system on a menopause treatment data set is designed. The cluster centers, which are found by applying the MCFC algorithm as well as the other clustering algorithms, are used in this process. The Euclidean distance between the cluster centers and patient's laboratory test results is considered for deciding the treatment. The patient should be assigned to the cluster with the minimum distance. The treatment of the assigned cluster should be applied to the patient in general.

## 1.2. Literature Survey

In the literature, there are many studies and research about the clustering analysis. However the clustering methods are similar, the small differences in some techniques changes the results of the clustering.

One of the earliest clustering algorithms is the so-called "k-means" algorithms. The term "k-means" was first used by James MacQueen in 1967 [12]. Although the term was first used by James MacQueen, the idea belonged to Hugo Steinhaus in 1956 [26]. The k-means algorithm was proposed by Stuart Lloyd in 1957 as a technique for pulse code modulation since this technique was not published until 1982 [13].

There are many researchers studying about the k-means algorithms. A. Likas, N. Vlassis and J. Verbeek were studying about the global k-means algorithm [5]. This algorithm is an incremental approach in the clustering field which is an approach to k-means clustering by dynamically adding one cluster center at a time through a deterministic global search procedure. This procedure consists of N (size of the data set) executions of the k-means algorithm from random initial positions. Some specific modifications on the k-means algorithm decrease the computational cost of the method.

Although the pre-assigned number of clusters is used in the k-means algorithm in general, some researchers work on the k-means algorithm without a known number of clusters. In the work of Krista Rizman Zalik [14], the k-means algorithm is redesigned without the need of the number of clusters by using the mean-square error cost function of the k-means. When the cost function reaches a global minimum, the correct number of clusters is determined. Therefore, by using the method of minimization of the cost function, there is no need to assign the number of clusters initially in the k-means algorithm.

During the studies on k-means algorithm, it was understood that the optimization methods used in the clustering analysis effect the results of the clustering. Therefore, method of ants type evolutionary optimization techniques based clustering algorithms became popular in the clustering field. R.J. Kuo, H.S. Wang, Tung-Lai Hung and S.H. Chou worked on the ant k-means algorithm in their study [15]. The ant k-means algorithm locates the data members in a cluster with a probability which is updated by pheromone excreted by the ants.

After development of k-means algorithm, the name “fuzzy” became popular in the unsupervised learning field. In fuzzy clustering, the data members are assigned to more than one cluster with suitable membership function values. The value of the membership function shows how much a data point belongs to a particular cluster. Fuzzy clustering is a process of assigning the membership values and data members to the clusters according to these values. One of the most widely used fuzzy clustering algorithms is the fuzzy c-means algorithm (FCM), which was first reported in the literature by Joe Dunn in 1974 [27]. Dunn used the fuzzification degree m = 2 in his study. Following that study, Jim Bezdek improved Dunn’s study for a general case in his Ph.D. thesis in 1981. Thereafter, Bezdek is assumed as the creator of the general FCM algorithm.

FCM algorithm is sensitive to the initialization of the cluster centers, membership functions and total number of clusters. Therefore, after the design of the FCM algorithm, some techniques were developed to improve the FCM algorithm. Parag M. Kanade and Lawrence O. Hall studied about the initialization techniques for the FCM clustering [1]. The clustering stage of their method is the classical FCM algorithm but their initialization stage was an original development by using the motions of ants. Ants correspond to the features of the cluster center. Each ant moves independently from the other ants. The best arrangement of the ant colony was determined according to the reformulation of the fuzzy optimization functions. After finding the best arrangement, the classical FCM algorithm can be performed on the data set. So, the initial cluster centers found by the ants affect the outputs of the FCM algorithm.

In the work of Swagatam Das, Ajith Abraham and Amit Konar, multi-ellitist particle swarm optimization (MEPSO) based automatic kernel clustering was proposed [16]. The particles include the activation threshold values of cluster centroids and coordinates of the cluster centroids. According to the threshold values, the cluster centroids are activated (or not) during the iterations. After the iterations on PSO, the globally best particle, which has the highest value of fitness function, can be obtained and the number of cluster centers and their coordinates can be computed. Therefore, the algorithm does not need to know the total number of cluster centers initially.

Miin-Shen Yang and Kuo-Lung Wu had a research on similarity based clustering [25]. In this clustering method, all data members in the data set are assumed as initial cluster centers. Then, correlation comparison and similarity clustering algorithm are applied on the data set. After that, agglomerative hierarchical clustering is applied and location of the cluster centers and the number of them is found according to the hierarchical clustering tree. There is no initial information needed for this method in general.

In addition to complete set clusterings, some clustering techniques which are applicable to data sets having data vectors with missing components have been designed. The techniques are divided into two groups in general. The methods in the first group try to fill the missing values before the clustering analysis and the methods in the second group try to fill the missing values during the clustering analysis.

Manish Sarkar and Tze-Yun Leong were studied on the FCM clustering with missing values [4]. The method that they designed does not require any initial imputation before the clustering process. The missing values are filled at each iteration. Therefore, there are no constant values for the missing data members. The mean-value imputation is applied on the data set during FCM iterations. So, it is not required to use initial imputation of the missing values before the clustering analysis.

In the study of Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein and Russ B. Altman, missing value estimation methods for DNA microarrays were suggested [17]. Gene expression analysis requires complete data set without missing values. The analysis depends on the clustering techniques in the literature. Singular value decomposition based method, weighted K-nearest neighbors and row average methods are described as missing value estimation methods for DNA microarrays in their study.

There is an innumerable work on expert systems; so we will only mention a few of them somewhat similar to our approach. Arash Ghanbari, S. Farid Ghaderi and M. Ali Azadeh proposed the clustering based genetic fuzzy expert system [18]. At first, k-means clustering algorithm is applied to the data set and the data set is divided into k-clusters. Then, all clusters are fed into independent genetic fuzzy system models with the ability of rule base and data base extraction. At the end, the results with genetic fuzzy systems without clustering and clustering based genetic fuzzy systems are compared. In this study, the evaluations indicate that clustering based genetic fuzzy systems provides more accurate results than genetic fuzzy systems without clustering. The study provided by them shows that the clustering based genetic fuzzy systems will have higher interpretability.

J. Kobayashi, H. Asaka, H. Mitsui, M. Sone, I. Terayama and Y. Kenmotsu studied on the expert systems with fuzzy clustering methods [19]. Their study is about the selection of the limit for nondestructive test of life time of transformers. At that time, expert systems about the transformer life time determination were based on a threshold value. The experts systems using fuzzy clustering have the judgement with consideration of relationship between data members. By using fuzzy clustering, the system understands the relationship between the conditions and can establish the optimum judgement. Therefore, in this study, it is required to use fuzzy clustering based expert systems because of their learning capability and their relational abilities.

## 1.3. Scope of The Thesis

The aim of this thesis study is to design a decision making system by using the MCFC, which is originally suggested in thesis study and which naturally is different from the well-known clustering algorithms defined in the chapter 1.2. The main contribution of this study is the construction of the MCFC algorithm and its application in the expert system development named as "CoRUM Version 1.1", which is an expert system that is supposed to help doctors to make a decision about the menopause treatment of the patient.

This study consists of seven chapters. It starts with Chapter 2 with a theoretical background of well-known clustering techniques used in the literature. The fuzzy C-means, hard C-means, K-means, similarity based clustering and fuzzy C-means clustering with missing values are described in this chapter. Also the normalization of the data set and missing value imputation are presented in this chapter. Chapter 3 includes the real and artificial data sets used in the simulations. The artificial data sets are made of mixtures of Gaussian distributions and the real data set is composed of laboratory test results of the women who has undergone menopause treatment. Chapter 4 describes the applications of clustering algorithms defined in Chapter 2 by using the data sets defined in Chapter 3. Chapter 5 presents the proposed clustering method MCFC briefly. Chapter 6 includes the design of a decision making system. The computer program "CoRUM V. 1.1" is made for decision making and especially for recommending. The details of the program are presented in this chapter. And at last, Section 7 concludes with the recommendations and results of this study.

# CHAPTER 2

# THEORETICAL BACKGROUND

The following sections give brief information about the fuzzy ant initialization of cluster centers, clustering methods and imputation of missing values in a given data set.

## 2.1 Fuzzy Ant Initialization

Initialization is the preliminary process before performing clustering analysis. This process affects the clustering quality and accuracy of the clustering. If better initial cluster centers are selected at the beginning of the process, better cluster centers can be reached. While initialization clusters centers, the best strategy is to select them as far as possible from each other.

In the literature, there are many studies about selecting good cluster centers at the beginning of the clustering. These studies show that initialization process effects the clustering of the data sets. Accordingly, the researchers study on the initialization techniques as much as they study the clustering itself.

In this study, the fuzzy ant initialization technique is used as an initialization technique before the clustering processes. In the fuzzy ant initialization algorithm, since the ant motions are somewhat stochastic, they more or less determines the best initial cluster centers to be used in FCM and HCM algorithms which are very sensitive to initial cluster center selection.

The ant motions are stochastic and it is simulated to obtain good cluster centroids. They move randomly in the feature space and each ant carries only one feature of each cluster center. After a fixed number of iterations, the cluster centers are calculated using the reformulation equations of FCM and HCM algorithm.

The reformulation of HCM is given below:

$$R_1(xclus) = \sum_{k=1}^{n} \min(D_{1k}, D_{2k}, ..., D_{ck}) \tag{2.1}$$

where;

$c \geq 2$ : number of clusters

n : number of data points

$x_k$: data member

$D_{ik}$: distance of $x_k$ from ith cluster center

$xclus_i$ : ith cluster prototype

$m \geq 1$: fuzzification degree

The reformulation of FCM is given below:

$$R_m(xclus) = \sum_{k=1}^{n} \left( \sum_{i=1}^{c} D_{ik}^{1/1-m} \right)^{1-m} \tag{2.2}$$

where;

$D_{ik}$: distance of $x_k$ from ith cluster center

$m \geq 1$: fuzzification degree

$c \geq 2$ : number of clusters

n : number of data points

Groups of ants cooperate in finding the initial optimal cluster center values. They cooperate to find the best location but otherwise they work independently during the iterations.

The fuzzy ants initialization process starts with creating a partition which can be descriced by a “c” cluster centers in the feature space. Each cluster center is of dimension “s”. Each unique ant is assigned to a component of a cluster. So, we have s * c ants in the partition and they become the structure of the partition. When the ants are moving, they position the new centroids and create the new partition. Each ant has a memory which containes the five most optimal positions visited. When the ant moves and stops, it will replace the least good rank ordered stored position if the current position is a more optimal partition.

Each ant moves and stops independently from the other ants. Their actions are independent but they cooperate to find the best partition which gives the best initial cluster center for the clustering algorithms. At first, the components of the partition are normalized between 0 and 1 and each ant is assigned to a particular component. When ants are moving, they can not change their dimension. A fixed number of iterations can be assigned for the ant movement. After a fixed number of iterations, ants stop and creates a new partition, which is called an “epoch”. If the current partition is better than the other previous partitions in the ant’s memory, the worst partition is deleted from the memory and the new partition is added to the memory. Else, the ant goes back to a better partition with a given random probability or continues from the new partition.

There are two directions for the ant movement. These are positive and negative directions. Positive direction is the ant movement from 0 to 1 and negative direction is the ant movement from 1 to 0. With a value between $D_{min}$ and $D_{max}$, the ant moves in the selected direction. During the iterations, if the ant reaches the end of the feature space, then it changes the direction. After a fixed number of epochs, the ants stop.

The evaluations of the partitions are obtained from the $R_m$ value for each partition. The worst known $R_m$ value is deleted from the ant's memory. After a fixed number of epochs, the initial cluster centers are acquired and the clustering process for FCM and HCM algorithms can be performed.

By using fuzzy ant initialization, it is possible to find a set of good initial cluster centers compared to the random initialization. Therefore, this method is useful for the initialization process before clustering.

### 2.2 Clustering Methods

Clustering is one of the most important unsupervised learning techniques in data mining [2], [3], [11], [24]. Due to its nature of being unable to produce a ground truth, this problem is still one of the most studied in data mining field.

In general, there are many clustering approaches used in literature. The most popular methods are fuzzy C-means [1], [4], [8], [9], [10], hard C-means [1], [7] and K-means [5], [6] clustering methods. In this thesis, FCM (standard), HCM, KM, SBC and FCM with missing data clustering were studied. The following sections give brief information about these clustering methods used in the clustering area.

### 2.2.1 Fuzzy C-means Clustering (Standard)

The fuzzy clustering is a method of clustering, which depends on the fuzzy membership values of each data in the data set and which finds the cluster centers using the fuzzy memberships of the data [11]. Fuzzy clustering algorithms have an ability to assign data members to multiple clusters. Membership of each data member decides the cluster that the data can be assigned to. The clustering method suggested in this study depends on a merging procedure which utilizes a similarity measure. This measure can only be used if the clustering method is fuzzy.

FCM algorithm is a well known algorithm in fuzzy clustering field [1], [8], [9], [10]. This algorithm is very sensitive to the number of cluster centers; therefore, it is required to know the number of cluster centers initially. In FCM [1], the aim is to minimize the objective function shown below.

$$J_m(U, xclus) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} D_{ik}(x_k, xclus_i) \tag{2.3}$$

where;

$m \geq 1$: fuzzification degree

$xclus_i$: ith cluster prototype

$c \geq 2$: number of clusters

n: number of data points

$D_{ik}(x_k, xclus_i)$: distance of $x_k$ from ith cluster center

For minimizing the objective function, the FCM algorithm can be described by the following steps:

1. Initialize the cluster centers as $xclus_0$,
2. Update the membership values according to equation (2.4). Membership values are defined as a membership matrix.

$$U_{ik} = \frac{D_{ik}^{2/(m-1)}}{\sum_{j=1}^{c} D_{jk}^{2/(m-1)}} \tag{2.4}$$

where;

$U_{ik}$: membership of kth data member in the ith cluster

$m \geq 1$: fuzzification degree

$c \geq 2$: number of clusters

$D_{ik}(x_k, xclus_i)$: distance of $x_k$ from ith cluster center

3. Calculate the updated cluster centers at the tth step.

$$xclus_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}} \tag{2.5}$$

where;

$U_{ik}$: membership of kth data member in the ith cluster

$m \geq 1$: fuzzification degree

n: number of data points

$xclus_i$: ith cluster prototype

4. If the Euclidean distance between at tth and (t-1)th cluster centers is smaller than the predefined threshold ϵ, stop the algorithm. Otherwise, continue from Step 2.

With respect to the algorithm, it is clear that a prior knowledge of the number of clusters and initial cluster centers are required for the FCM algorithm.

### 2.2.2 Hard C-means Clustering

Hard C-means (HCM) algorithm is one of the simplest and most used unsupervised clustering algorithms and it is very easy to implement [1]. The major problem for the algorithm is that, it is very sensitive to selecting initial cluster centers. After selecting, nearest neighbor algorithm is used to assign each data to a cluster. After clusters are obtained, new centers can be calculated. The steps are repeated till there is no significant change in the cluster centers.

The objective function for the HCM algorithm is given below:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{n} D_{ik}(x_k, xclus_i) \tag{2.6}$$

c: number of clusters (c≥2)
n: number of data points
$xclus_i$: *ith* cluster prototype
$x_k$: *kth* data vector
$D_{ik}(x_k, xclus_i)$: distance of $x_k$ from *ith* cluster center

In general, the hard C-means algorithm is defined by the following steps:

1. Initialize the each cluster centers $xclus_0$
2. Update the membership matrix U, $U^t$, $U^{(t-1)}$
   $U_{ik} = 0$, if $D_{ik} > \min (D_{1k}, D_{2k}, D_{3k},\ldots,Dc_k)$
   $U_{ik} = 1$, otherwise
3. At the $t^{th}$ step, calculate the new cluster centers by using the equation (2.7).

$$xclus_i = \frac{\sum_{k=1}^{n} U_{ik}^m x_k}{\sum_{k=1}^{n} U_{ik}^m} \tag{2.7}$$

4. If $\left| xclus^t - xclus^{t-1} \right| < \varepsilon$ then stop; otherwise go to step 2

### 2.2.3 K-means Clustering

K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean [2], [5], [6]. In k-means clustering, the main idea is to assign one cluster center to each cluster, and totally k-cluster centers for k-clusters. The number "k", which defines the number of clusters and the initial cluster centers, is the a priori knowledge for k-means clustering.

In k-means clustering, the aim is to minimize the objective function shown below:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i - xclus_j \right\|^2 \tag{2.8}$$

where;

$xclus_j$: jth cluster center

$x_i$: data member

K-means assigns each observation to clusters based upon the mean of the cluster. The cluster's mean is computed and the process continues with the same steps again. The k-means algorithm is composed of the following steps:

1. Place k points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the k centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The k-means algorithm is not necessarily able to find the ideal number of clusters. This fact is the main disadvantage of the algorithm. To find the best solution, experimenting with different values of k to identify the value that best suits the data set should be performed.

### 2.2.4 Similarity Based Clustering

Similarity based clustering method (SCM) includes correlation comparison algorithm (CCA), similarity clustering algorithm (SCA) and agglomerative clustering algorithm (AHC) [25].

The data set x = $\{x_1,x_2.....,x_n\}$ where $x_j$ is a feature vector in the s-dimensional Euclidean space $R^s$ and c is the number of clusters. The algorithm uses Euclidean distance to find the distance between the data vectors. The Euclidean norm $\|x_j-z_i\|^2$ is used as the dissimilarity measure between $x_j$ and the $i^{th}$ cluster center $z_i$. If we want to cluster data set into c clusters, we may find $z_i$ to minimize the total dissimilarity objective function.

$S(x_j,z_i)$ is the similarity measure between $x_j$ and *ith* cluster center $z_i$. We show the formula of $S(x_j,z_i)$ as:

$$S(x_j, z_i) = \exp\left(-\frac{\|x_j - z_i\|^2}{\beta}\right) \tag{2.9}$$

The total similarity measure $J_s(z)$ can be found by using the following formula:

$$J_s(z) = \sum_{i=1}^{c}\sum_{j=1}^{n} \exp\left(-\frac{\|x_j - z_i\|^2}{\beta}\right)^{\gamma} \tag{2.10}$$

The sample variance, β can be defined as:

$$\beta = \frac{\sum_{j=1}^{n}\left\|x_j - \bar{x}\right\|^2}{n} \tag{2.11}$$

where;

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} \tag{2.12}$$

The parameter γ is very important to determine the ideal number of clusters. To analyze the effect of γ, we can arrange the total similarity of the data point $x_k$ to all data points with the given equation.

$$\bar{J}_s(x_k) = \sum_{j=1}^{n} \exp\left( -\frac{\left\| x_j - x_k \right\|^2}{\beta} \right)^{\gamma} \text{, k = 1,2,…,n} \tag{2.13}$$

A large value of $\bar{J}_s(x_k)$ means that the data point $x_k$ is close to a set of cluster centers.

In general, the objective function will have only one peak when γ is small even though the data set actually has many cluster centers. The SCM objective function has only one peak when γ is small. So, we start the correlation comparison algorithm with γ = 5. In the program, we select a threshold value as 0.97.

We have used different values of γ and change it according to "m" value. This is shown as $\gamma_m = 5m$, m = 1, 2, 3… .

Then the objective function can be rewritten as:

$$\bar{J}_s(x_k)_{\gamma_m} = \sum_{j=1}^{n} \exp\left( -\frac{\left\| x_j - x_k \right\|^2}{\beta} \right)^{\gamma_m} \text{ k = 1, 2, …, n} \tag{2.14}$$

In the correlation comparison algorithm, the aim is to find $\gamma_m$ and $\beta$ values. The algorithm starts with selecting m = 1 and threshold value $\varepsilon_1 = 0.97$. After selecting, the correlation of the values $\bar{J}_s(x_k)_{\gamma_m}$ and $\bar{J}_s(x_k)_{\gamma(m+1)}$ are calculated. If the correlation values are greater than or equal to the specified threshold value, we can choose the $\gamma_m$ as an estimate of $\gamma$. If the correlation is not greater than the specified threshold value, we choose m = m+1 and perform this process till the correlation is greater than or equal to the threshold value.

After estimating the $\gamma$, the next step is to find the zi value which maximizes $J_s(z)$. If the derivative of $J_s(z)$ is taken, the necessary condition that maximizes $J_s(z)$ can be found as:

$$z_i = \frac{\sum_{j=1}^{n} x_j (\exp - \frac{\left\| x_j - z_i \right\|^2}{\beta})^{\gamma}}{\sum_{j=1}^{n} (\exp - \frac{\left\| x_j - z_i \right\|^2}{\beta})^{\gamma}} \tag{2.15}$$

If we take the similarity relation as

$$S_{ij} = S(x_j, z_i) = \exp(-\frac{\left\| x_j - z_i \right\|^2}{\beta}) \tag{2.16}$$

the necessary condition becomes,

$$z_i = \frac{\sum_{j=1}^{n} S_{ij}^{\gamma} x_j}{\sum_{j=1}^{n} S_{ij}^{\gamma}} \tag{2.17}$$

For initial values of $z(0) = (z_1(0), z_2(0), \ldots, z_n(0))$, if the original data set values are assigned to z, the algorithm is robust to outliers.

After initializing $z_i(0)$, i = 1,2, …,c, we select threshold as 0.97 which is the same as the CCA algorithm. We estimate $S_{ij}(l+1)$ then estimate $z_i(l+1)$. We increase till $\max_i \left\| z_i^{(l+1)} - z_i^{(l)} \right\| < \varepsilon$. At the end of the algorithm, we reach data set $z_i(0)$, i = 1,2, …,c. This shows the possible cluster centers and optimal number of the clusters c* for the original data set. In some conditions, the optimal cluster number c* can be observed by the view of sight. But, a precise method called AHC is chosen with the final states of all cluster centers to find the optimal c*.

The dissimilarity measure for the AHC is the Euclidean norm and this measure is used in the objective function calculations. Single linkage method is used in the AHC as a linkage way. In some conditions, Ward's method is used during the processes for the AHC. The hierarchical clustering trees show the final states of all data points and in the trees, the increase in y-coordinate represents the distance between the clusters.

At last, these three algorithms continuously become together and they create SCM algorithm. In this algorithm, there is no need to identify the number of clusters but in some condition (e.g., in high dimensional data sets), the number of clusters can be fixed in the AHC stage.

### 2.3 Missing Values and Clustering

The FCM, HCM, K-means and similarity based clustering methods can not be applied to data sets that contain vectors having missing component values. The missing values imply that the values of some of the attributes of the pattern are unknown.

The approaches to deal with missing values can be categorized into the following groups:

- **Deductive imputation:** Missing values are deduced with certainty, or with high probability from the other information of the pattern [22].

- **Hot-deck imputation:** Missing values are replaced with values from the closest matching patterns [22].

- **Mean-value imputation:** The mean of the observed values is used to replace the missing values [22].

- **Regression-based imputation:** Missing values are replaced by the predicted values from a regression analysis [21].

- **Imputation using Expectation-Maximization:** Missing values are repaired in two steps. In the E-step, the expected value of the log likelihood is calculated, and in the M-step, the missing values are substituted by the expected values. Then the likelihood function is maximized as if no data were missing [20], [22], [23].

The imputation is a procedure that aims to fill the missing values with estimated ones. Imputation methods are required when the analysis method needs complete data set for clustering. There are many imputation methods used in the literature. The approach that is used in this study is the mean imputation method to fill the missing data members.

In mean imputation, the missing values are replaced with the mean of the known patterns. Therefore, the estimated value is a mean value of the pattern. If the missing values are rare, this method can be appropriate for the data set.

Another approach used in this study is fuzzy clustering algorithm with no imputation. The details of the algorithm are given in the following sub chapter.

### 2.3.1 Fuzzy Clustering Algorithm with Missing Data Members

As already stated, fuzzy clustering methods divide data sets into a set of clusters based on a similarity function. Most of the fuzzy clustering algorithms need complete data set during clustering processes. In this section FCM clustering with no imputation without imputation of missing values at the initial stage [4] is described for clustering processes.

The data set X, which includes "n" members is required for the applications ($X = x_1, x_2, \ldots, x_n$). If the kth attribute of $x_j$ is missing then it is proposed as,

$$x_{jk} = \frac{\sum_{i=1}^{c} u_{ij}^{m} xclus_{ik}}{\sum_{i=1}^{c} u_{ij}^{m}} \tag{2.18}$$

where;
m: degree of fuzzification
$xclus_{ik}$: *kth* attribute value of cluster center $xclus_i$
$x_{jk}$: *kth* attribute value of $x_j$
$u_{ij}$: *ith* cluster membership of *jth* member

Equation (2.18) can be obtained from the objective function of FCM algorithm [4]. Cluster centers, values of the membership functions and missing values can be iteratively calculated minimizing the objective function (2.3) [23].

The computation equation for the update of cluster centers should be modified as follows:

$$c_{ik} = \frac{\sum_{j=1}^{n} u_{ij}^{m} i_{jk} x_{jk}}{\sum_{j=1}^{n} u_{ij}^{m} i_{jk}} \tag{2.19}$$

$i_{jk}$: kth attribute of index vector $i_j$,
If the attribute k of $x_j$ is not missing, then $i_{jk} = 1$. Otherwise, $i_{jk} = 0$.

Since the imputation of estimates for the missing values is avoided if they are already available, the distances of incomplete feature vectors to the cluster centers must be estimated. This can be done by assuming that the analyzed data is clustered. Then, all of the attribute-specific distances behave basically in the same way. The following equation is proposed for estimating the Euclidean distance between incomplete feature vectors and cluster centers.

$$d^2(x_j, c_i) = \frac{p}{\sum_{k=1}^{p} i_{jk}} \sum_{k=1}^{p} i_{jk} (x_{jk} - c_{ik})^2 \tag{2.20}$$

Blessed with estimates for the distance $d^2(x_j,c_i)$, probabilistic membership degrees $u_{ij}$ can be computed as:

$$u_{ij} = \frac{1}{\sum_{l=1}^{c} \left( \frac{d^2(x_j, c_i)}{d^2(x_j, c_l)} \right)^{1/(m-1)}} \tag{2.21}$$

Obviously, the cluster memberships $u_{ij}$ for incomplete feature vectors are finally determined based on the observed data only.

$$u_{ij} = \frac{1}{\sum_{l=1}^{c} \left( \frac{\sum_{k=1}^{p} i_{jk} \left( x_{jk} - c_{ik} \right)^2}{\sum_{k=1}^{p} i_{jk} \left( x_{jk} - c_{lk} \right)^2} \right)^{1/(m-1)}} \tag{2.22}$$

With respect to the equation (2.22) defined above, the fuzzy clustering algorithm uses the membership function for missing values. So, at every iteration, missing values are computed and as an output of this, new membership values and cluster centers are computed. Because of computation of missing values at all iterations, there is no need to impute the missing values before clustering.

# CHAPTER 3

# ARTIFICIAL AND REAL DATA SETS

The data sets that were handled in the clustering processes are described in this section.

## 3.1 Real Data Set

The real data set includes 179 data members which were the laboratory test results collected from women in menopause at Hacettepe University, Faculty of Medicine. This data set was provided by Prof. M. Sinan Beksaç.

The data set is a high-dimensional data set and includes the information about the laboratory test results including height, weight, age, menstruation period, FSH, LH, estradiol, t3, t4, TSH, glucose, cholesterol, triglyceride, hdl, ldl, vldl, hemoglobin, hematocrit, menopause type, HRT, HRT duration. The data set is shown in the Appendix A in detail. The explanations of the components in the data set are given briefly in table 1.

Table 1: Explanations of components in real data set

| Components | Explanation |
|---|---|
| **Age** | Years |
| **Length** | cm |
| **Weight** | kg |
| **Menopause Duration** | Months |
| **Menopause Type** | 1: natural<br>2: by surgery<br>3: premenopause duration<br>4: perimenopause duration |
| **HRT Duration** | Hormone Replacement Therapy by months |
| **HRT** | 1: HRT is applied<br>2: HRT is not applied |
| **FSH Level(mlu/ml)** | Laboratory test results |
| **LH Level(mlu/ml)** | Laboratory test results |
| **T3 Level (ng/ml)** | Laboratory test results |
| **T4 Level (ug/dl)** | Laboratory test results |
| **TSH Level (ulu/ml)** | Laboratory test results |
| **Estradiol Level (pg/ml)** | Laboratory test results |
| **Glucose Level** | Laboratory test results |
| **Cholesterol Level** | Laboratory test results |
| **Triglyceride Level** | Laboratory test results |
| **HDL Level** | Laboratory test results |
| **LDL Level** | Laboratory test results |
| **VLDL Level** | Laboratory test results |
| **Hemoglobin Level** | Laboratory test results |
| **Hematocrit Level** | Laboratory test results |

The data set is not a complete data set and it includes missing values in it. Before clustering processes, the missing values were filled with the mean imputation method described in chapter 2.3.

The data set components are ranging from 0.32 to 384. Thereof, the clustering with the original values can deflect the expected results of clusters and cluster centers. Therefore, normalization between 0 and 1 was required for the data set. For each component, maximum value of the component is selected. Then %2 higher value of the maximum value is calculated. This value is assigned as maxima. After that, minimum value of the component is selected. Then %2 lower value of the minimum value is calculated. This value is assigned as minima. After finding maxima and minima, the difference between them is calculated for the component. Then, the minima is subtracted from the data members of the component. At last, this value is divided by the difference value, which is the result of a subtraction of maxima and minima. These computations should be applied to the all components. Therefore, the data set was normalized between 0 and 1. After normalization, the clustering processes can be performed.

After performing clustering analysis, the normalized the data set was reintegrated and the cluster centers found are represented with respect to their original values.

### 3.2 Artificial Data Sets

This section describes the artificial data sets used in the clustering analysis. The artificial experimental data sets were created by using a mixture of Gaussian distributions and uniform distributions.

The data set1, which includes the Gaussian distributed data, includes 300 members as shown in figure 1. This data set is a mixture of three Gaussians and Gaussian mixtures have the same standard deviation but different mean values.

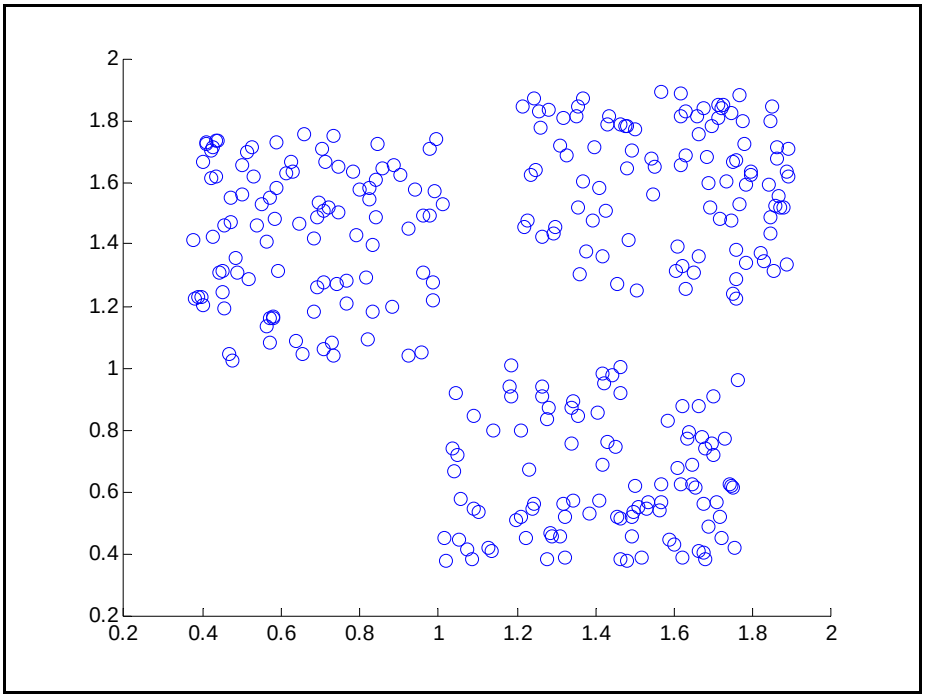


Figure 1: Data set 1

The data set2, which is shown in figure 2, includes the mixture of forty two Gaussian distributions. The mean and standard deviation of the mixtures are changeable in the data set.

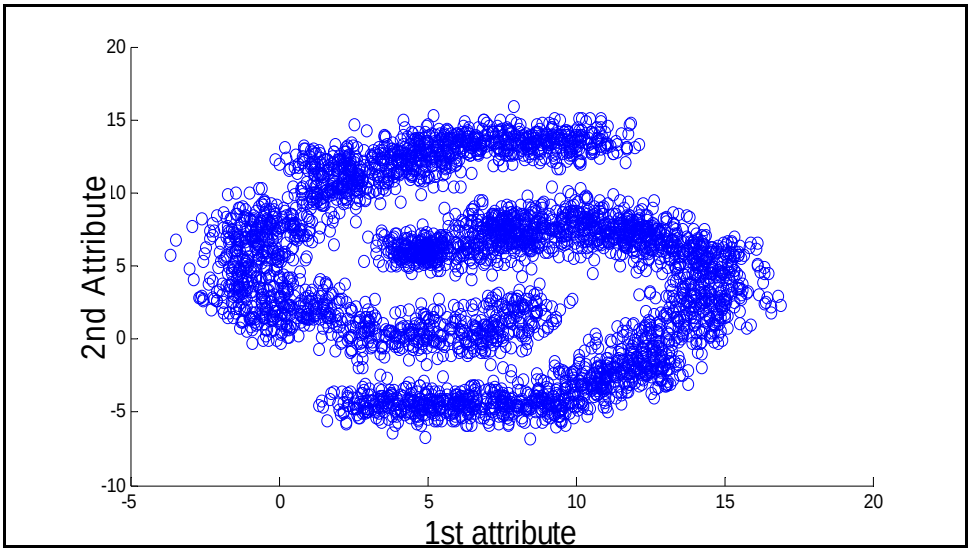


Figure 2: Data set 2

The data set3 includes two-dimensional 300 data members and created by using Gaussian distribution. Every 100 data members are generated by the Gaussian distributions. The distributions have different mean but same standard deviation values. The data set3 is shown in figure 3.

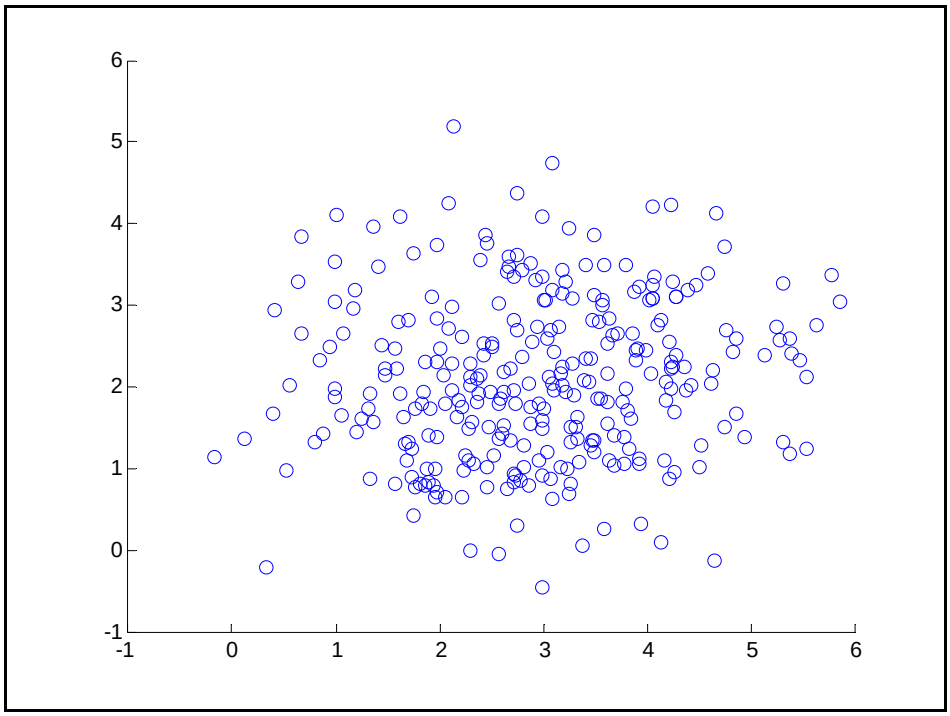


Figure 3: Data set 3

The data set 4 includes 300 data members and each data member is two-dimensional. Each 100 data member group has uniform distribution and they become together for creating data set 4. The following figure shows the data set which includes uniformly distributed three groups.

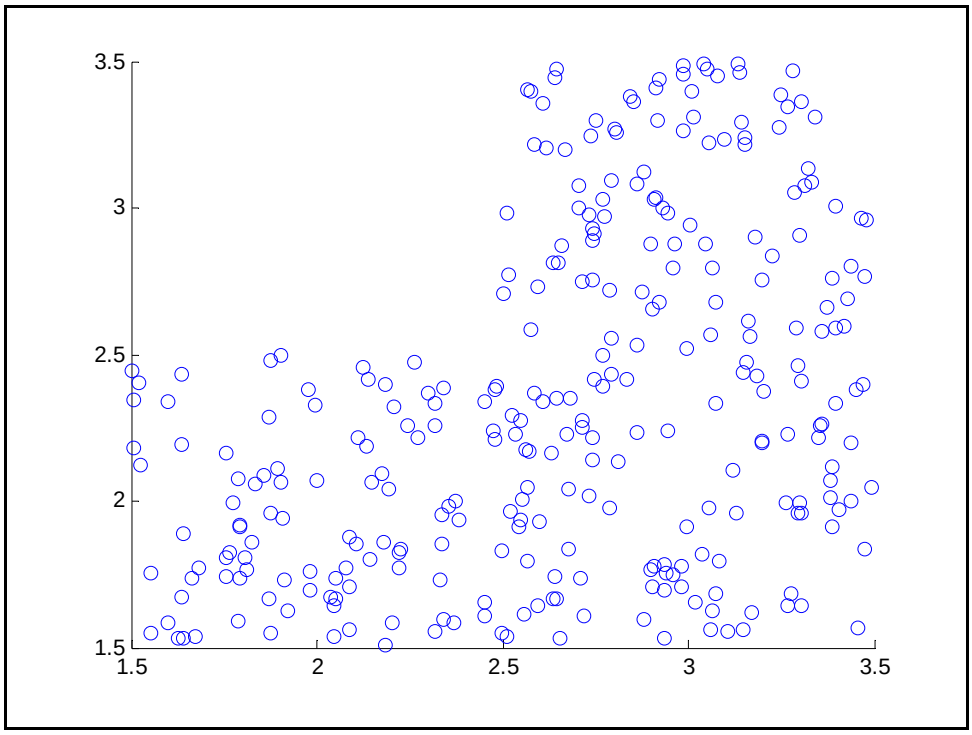


Figure 4: Data set 4

# CHAPTER 4

# APPLICATIONS OF CLUSTERING ALGORITHMS

This section demonstrates selection of the number of clusters and applications of the clustering algorithms defined in Chapter 2 to the data sets given in Chapter 3.

## 4.1 Fuzzy C-means (FCM) Results

FCM algorithm was applied to the four artificial data sets described in chapter 3. Fuzzy ants initialization technique was used as an initialization technique for the algorithm.

The FCM algorithm is sensitive to the selected number of cluster centers. To find good clusters, the aim is to minimize the objective function shown in the equation (2.3).

The following chapters describe the selection of ideal cluster centers by using objective function calculations and clustering applications on the data sets

### 4.1.1 Applications on data set 1

When we apply FCM by changing the number of cluster centers to 2-10, the objective function values which were calculated are shown in the following table:

Table 2: FCM-Changes in objective function according to the number of clusters for data set1

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 132.3178 |
| 3 | 80.8394 |
| 4 | 108.8445 |
| 5 | 137.9888 |
| 6 | 169.0842 |
| 7 | 187.0101 |
| 8 | 218.9352 |
| 9 | 246.8464 |
| 10 | 230.3289 |

According to the table 2, the graph of objective function vs. number of clusters is shown in the following figure.

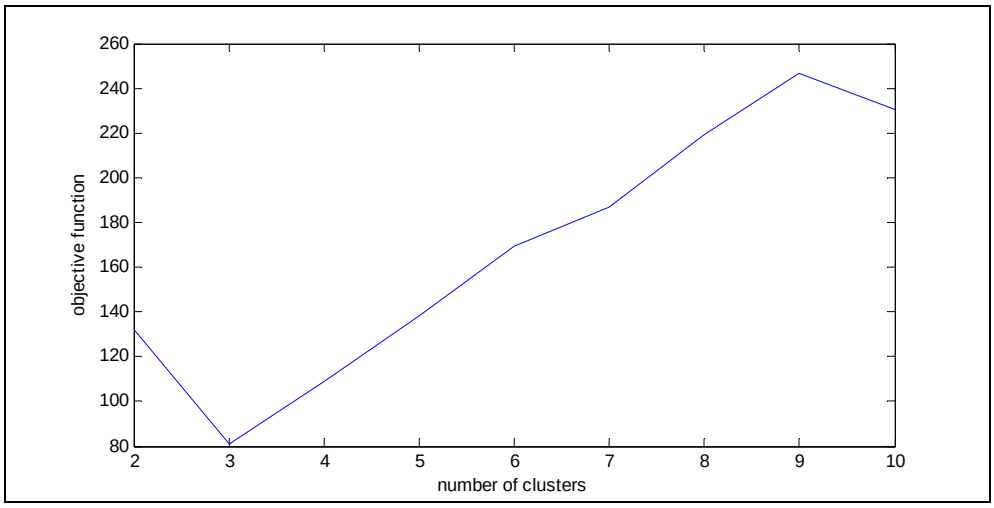


Figure 5: FCM - Objective function vs. the number of clusters for data set1

According to figure 5, it can be easily seen that three clusters is the ideal case for data set1 since, the objective function is minimum for three clusters.

The result of the FCM clustering for data set1 is shown in the following figure.

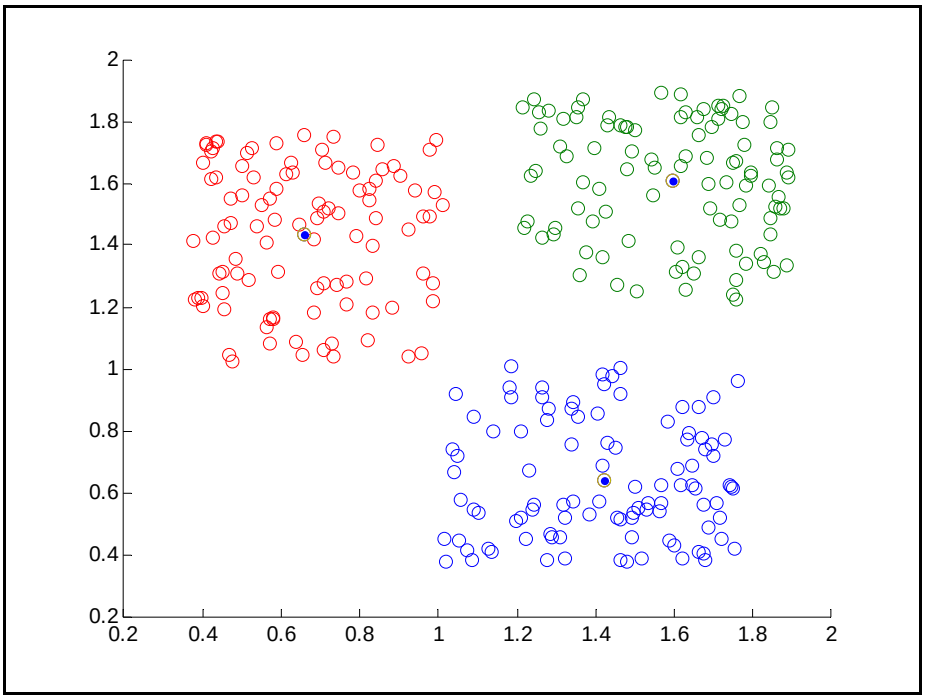


Figure 6: FCM results of data set 1

### 4.1.2 Applications on data set 2

When we apply FCM algorithm to data set 2 by changing the number of cluster centers between 2 to 10, the objective function values which were calculated for data set2 are shown in the following table:

Table 3: FCM-Changes in objective function according to the number of clusters for data set 2

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 8.6420e+003 |
| 3 | 9.2537e+003 |
| 4 | 9.6667e+003 |
| 5 | 1.2130e+004 |
| 6 | 1.2395e+004 |
| 7 | 1.4088e+004 |
| 8 | 1.2378e+004 |
| 9 | 1.0858e+004 |
| 10 | 9.2537e+003 |

According to the table 3, the graph of objective function vs. number of clusters is shown in the following figure.

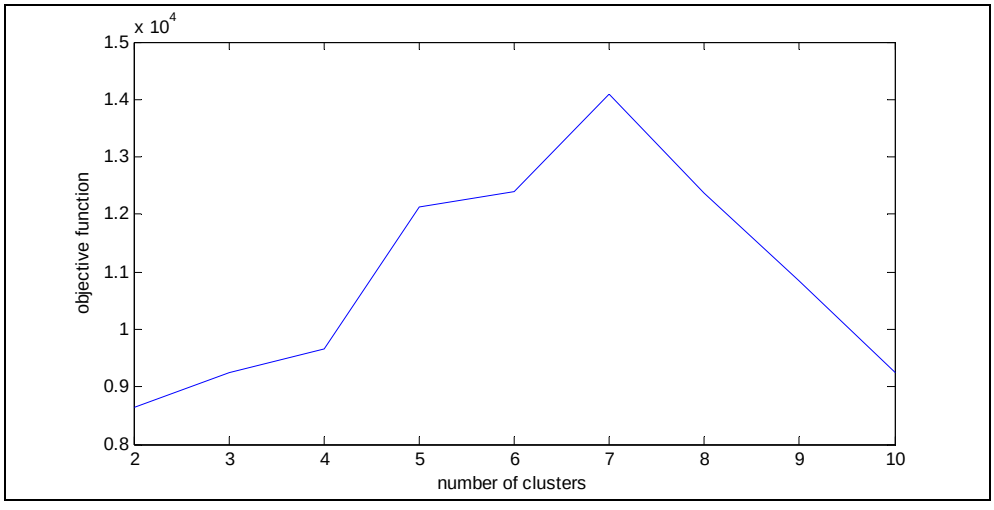


Figure 7: FCM - Objective function vs. the number of clusters for data set2

According to figure 7, it can be easily seen that two clusters is the ideal case for data set2 since, the objective function is minimum for two clusters.

The result of the FCM clustering for data set2 is shown in the following figure.

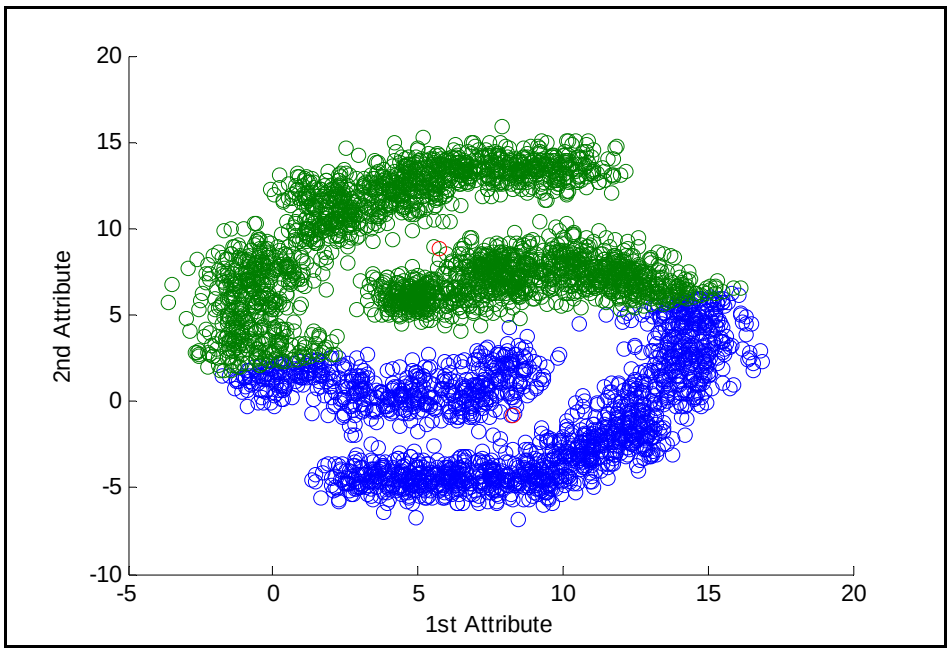


Figure 8: FCM results of data set 2

### 4.1.3 Applications on data set 3

When we apply FCM algorithm to data set3 by changing the number of cluster centers between 2 to 10, the objective function values which were calculated for data set3 are shown in the following table:

Table 4: FCM-Changes in objective function according to the number of clusters for data set 3

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 300.0512 |
| 3 | 252.5398 |
| 4 | 217.5794 |
| 5 | 262.7853 |
| 6 | 344.0006 |
| 7 | 320.9131 |
| 8 | 306.9790 |
| 9 | 394.9617 |
| 10 | 502.2777 |

According to the table 4, the graph of objective function vs. number of clusters for data set 3 is shown in the following figure.

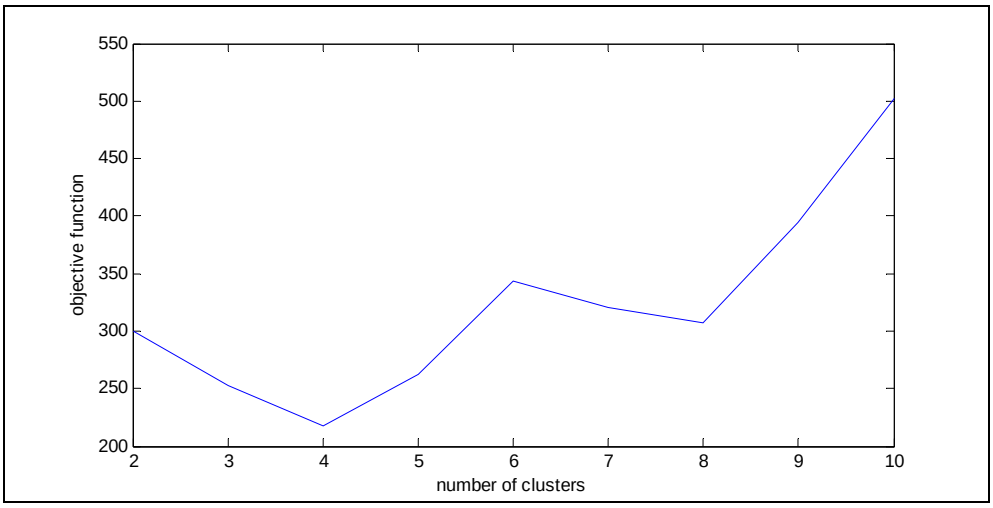


Figure 9: FCM - Objective function vs. the number of clusters for data set3

According to figure 9, it can be easily seen that four clusters is the ideal case for data set3 since, the objective function is minimum for four clusters.

The result of the FCM clustering for data set3 is shown in the following figure.

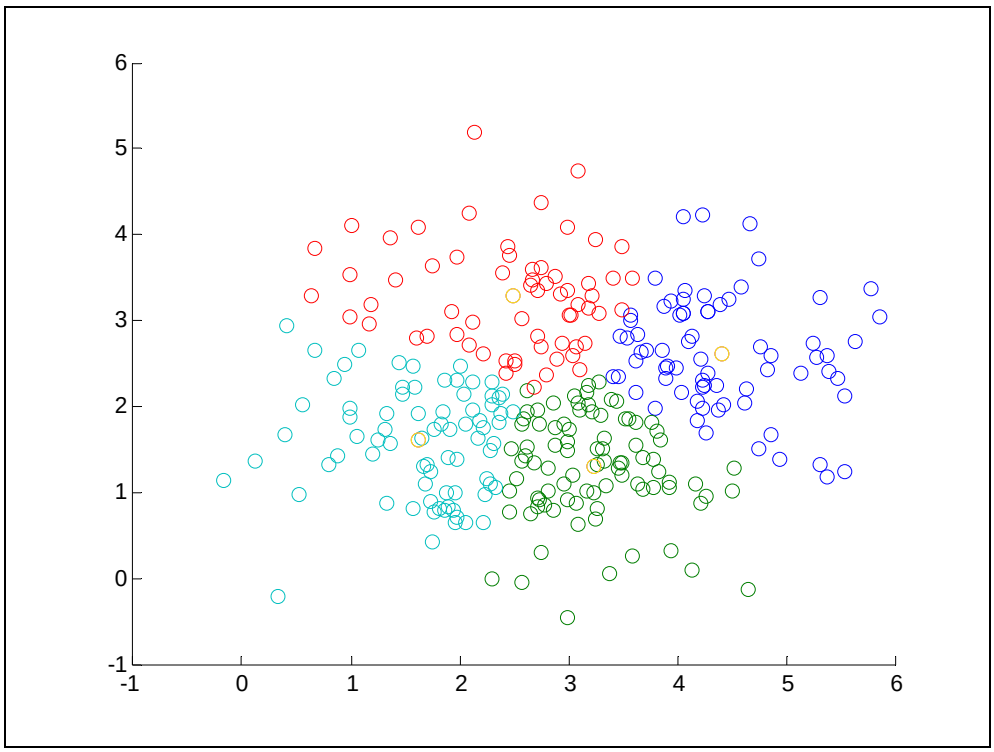


Figure 10: FCM results of data set 3

### 4.1.4 Applications on data set 4

When we apply FCM algorithm to data set4 by changing the number of cluster centers between 2 to 10, the objective function values which were calculated for data set4 are shown in the following table:

Table 5: FCM-Changes in objective function according to the number of clusters for data set 4

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 418.3406 |
| 3 | 94.4488 |
| 4 | 896.6811 |
| 5 | 1115.9543 |
| 6 | 1335.5489 |
| 7 | 1514.2654 |
| 8 | 1733.4245 |
| 9 | 1952.5120 |
| 10 | 2241.7837 |

According to the table 5, the graph of objective function vs. number of clusters for data set 4 is shown in the following figure.

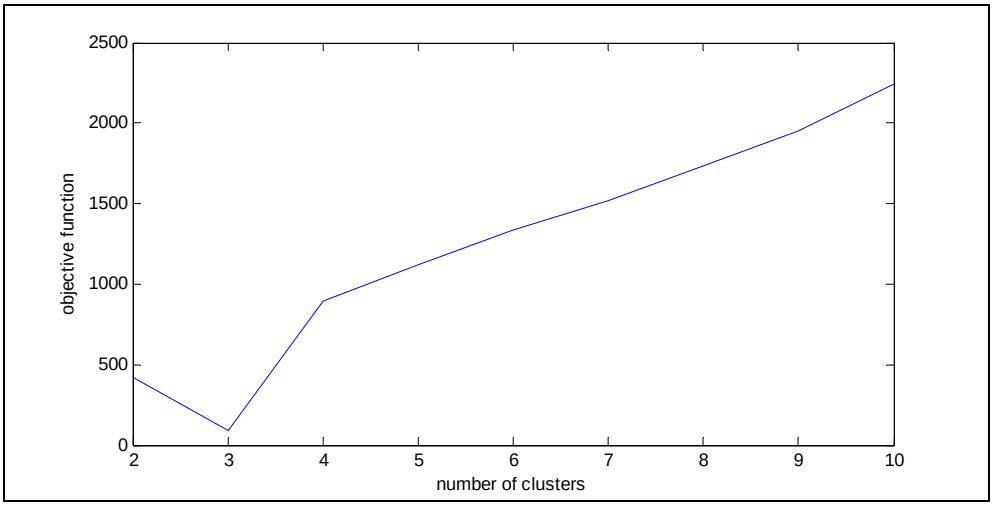


Figure 11: FCM - Objective function vs. the number of clusters for data set4

According to figure 11, it can be easily seen that three clusters is the ideal case for data set4 since, the objective function is minimum for three clusters.

The result of the FCM clustering for data set4 is shown in the following figure.

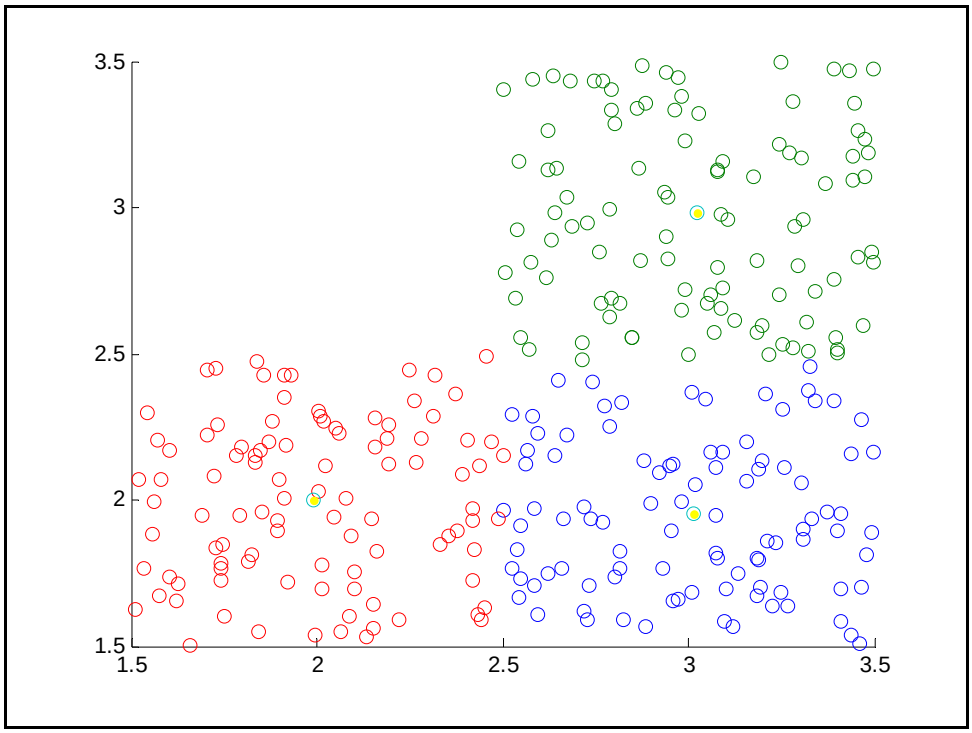


Figure 12: FCM result of data set 4

### 4.2 Hard C-means (HCM)

HCM algorithm was applied to the four artificial data sets described in chapter 3. Fuzzy ants initialization technique was used as an initialization technique for the algorithm.

HCM algorithm is sensitive to the selected number of cluster centers. To find good clusters, the aim is to minimize the objective function shown in the equation (2.6).

The following chapters describe the selecting of ideal cluster centers by using objective function calculations and clustering applications on the data sets.

### 4.2.1 Applications on data set 1

When we apply HCM algorithm to data set 1 by changing the number of cluster centers between 2 to 10, the objective function values which were calculated for data set1 are shown in the following table:

Table 6: HCM-Changes in objective function according to the number of clusters for data set 1

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 375.6007 |
| 3 | 80.8584 |
| 4 | 107.5949 |
| 5 | 134.3313 |
| 6 | 161.0678 |
| 7 | 187.8042 |
| 8 | 214.5407 |
| 9 | 241.2771 |
| 10 | 233.8716 |

According to the table 6, the graph of objective function vs. number of clusters is shown in the following figure.

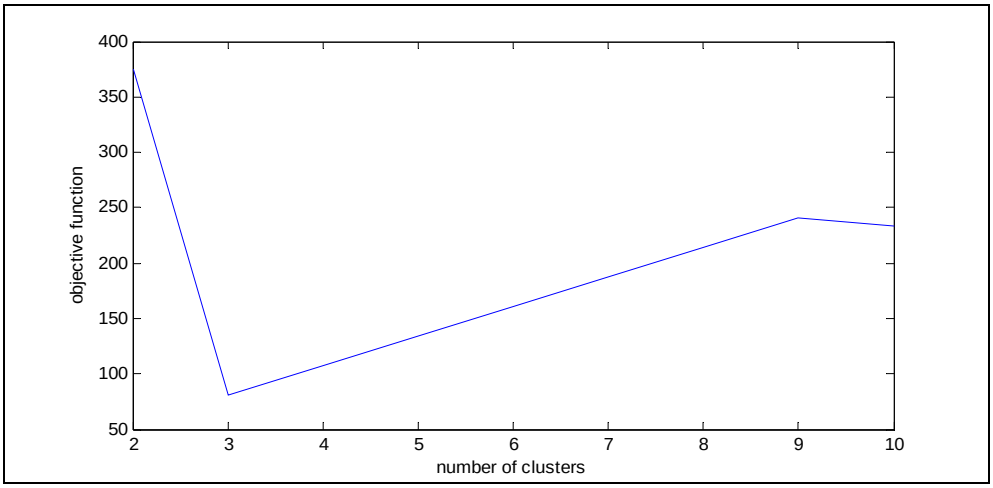


Figure 13: HCM - Objective function vs. the number of clusters for data set1

According to figure 13, it can be easily seen that three clusters is the ideal case for data set1 since, the objective function is minimum for three clusters.

The result of the HCM clustering for data set1 is shown in the following figure.

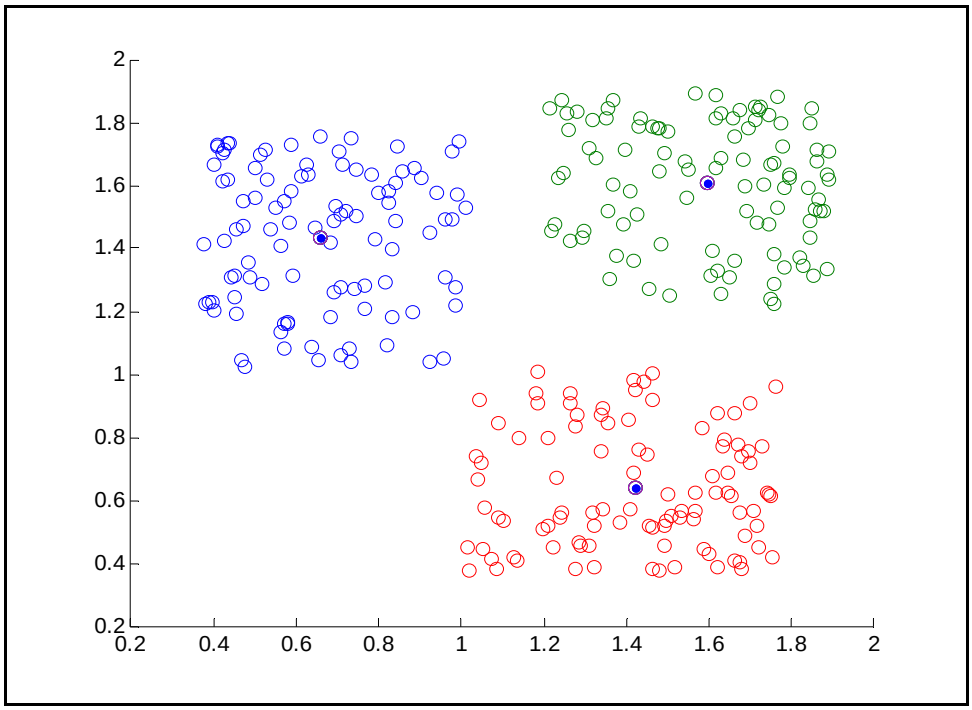


Figure 14: HCM results of data set 1

### 4.2.2 Applications on data set 2

When we apply HCM algorithm to data set 2 by changing the number of cluster centers between 2 to 10, the objective function values which were calculated for data set1 are shown in the following table:

Table 7: HCM-Changes in objective function according to the number of clusters for data set 2

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 7.6763e+003 |
| 3 | 1.7454e+004 |
| 4 | 1.4088e+004 |
| 5 | 1.2130e+004 |
| 6 | 1.2395e+004 |
| 7 | 1.3740e+004 |
| 8 | 1.2378e+004 |
| 9 | 1.0858e+004 |
| 10 | 9.2537e+003 |

According to the table 7, the graph of objective function vs. number of clusters is shown in the following figure.

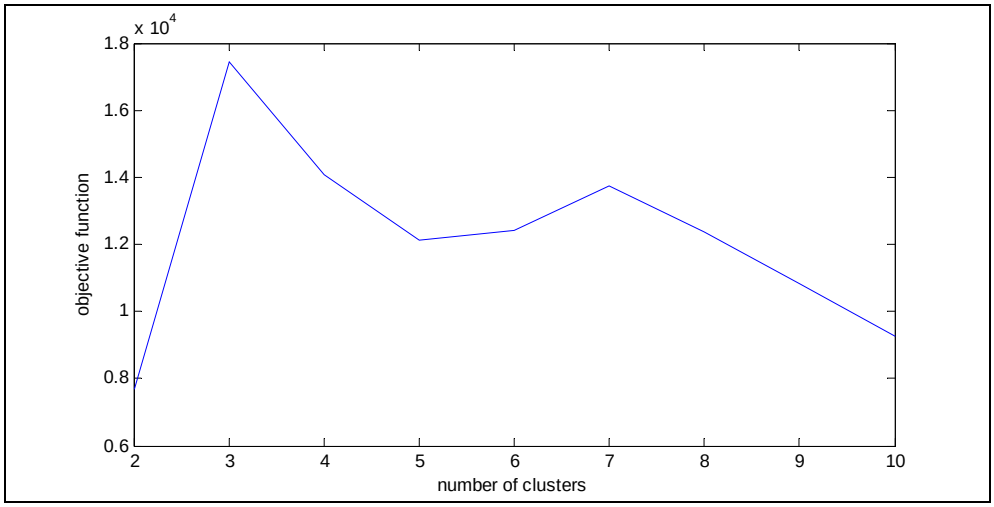


Figure 15: HCM - Objective function vs. the number of clusters for data set2

According to figure 15, it can be easily seen that two clusters is the ideal case for data set2 since, the objective function is minimum for two clusters.

The result of the HCM clustering for data set2 is shown in the following figure.

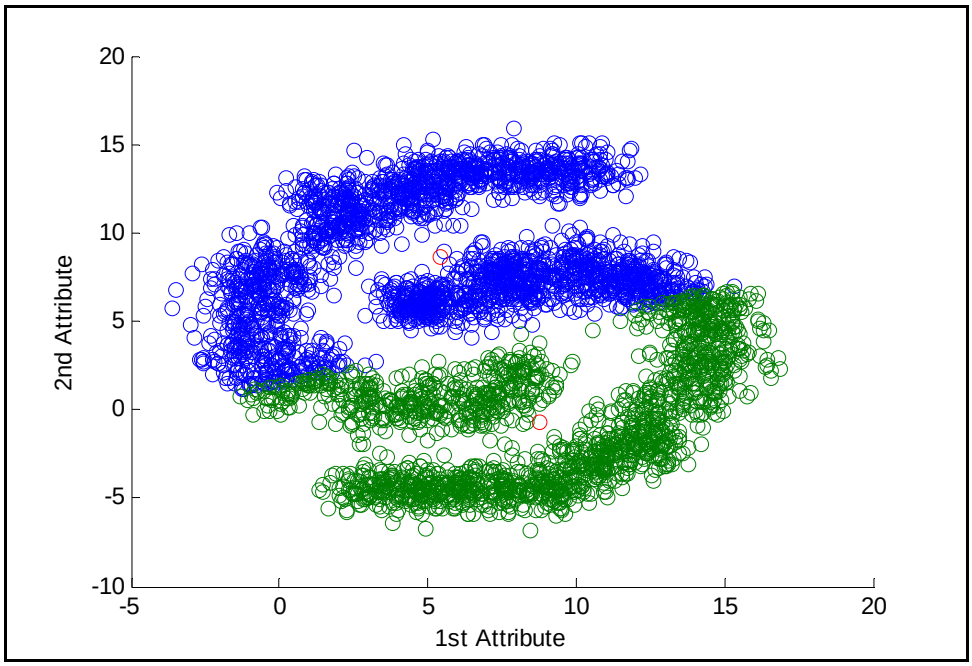


Figure 16: HCM results of data set 2

### 4.2.3 Applications on data set 3

When we apply HCM algorithm to data set3 by changing the number of cluster centers between 2 to 10, the objective function values which were calculated for data set3 are shown in the following table:

Table 8: HCM-Changes in objective function according to the number of clusters for data set 3

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 319.0631 |
| 3 | 271.5517 |
| 4 | 236.5913 |
| 5 | 281.7972 |
| 6 | 363.0127 |
| 7 | 339.9260 |
| 8 | 325.9909 |
| 9 | 413.9736 |
| 10 | 521.2896 |

According to the table 8, the graph of objective function vs. number of clusters for data set 3 is shown in the following figure.

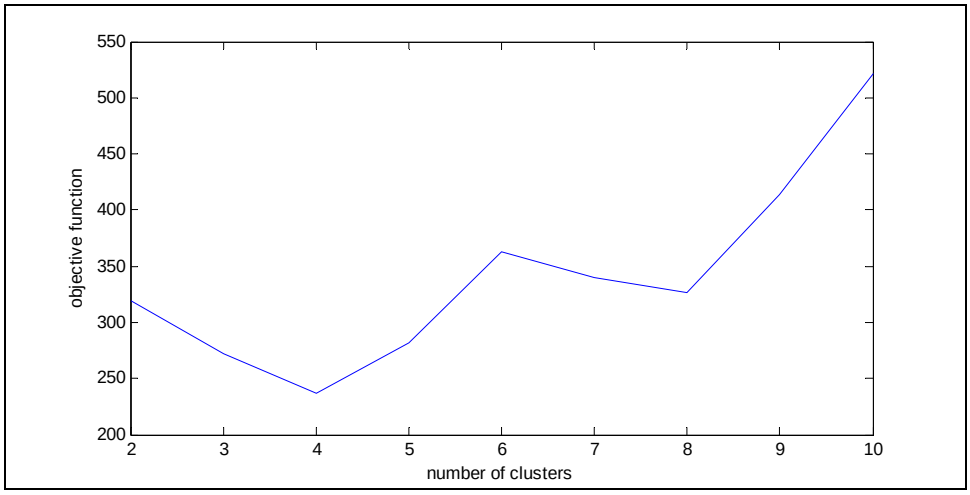


Figure 17: HCM - Objective function vs. the number of clusters for data set3

According to figure 17, it can be easily seen that four clusters is the ideal case for data set3 since, the objective function is minimum for four clusters.

The result of the HCM clustering for data set3 is shown in the following figure.

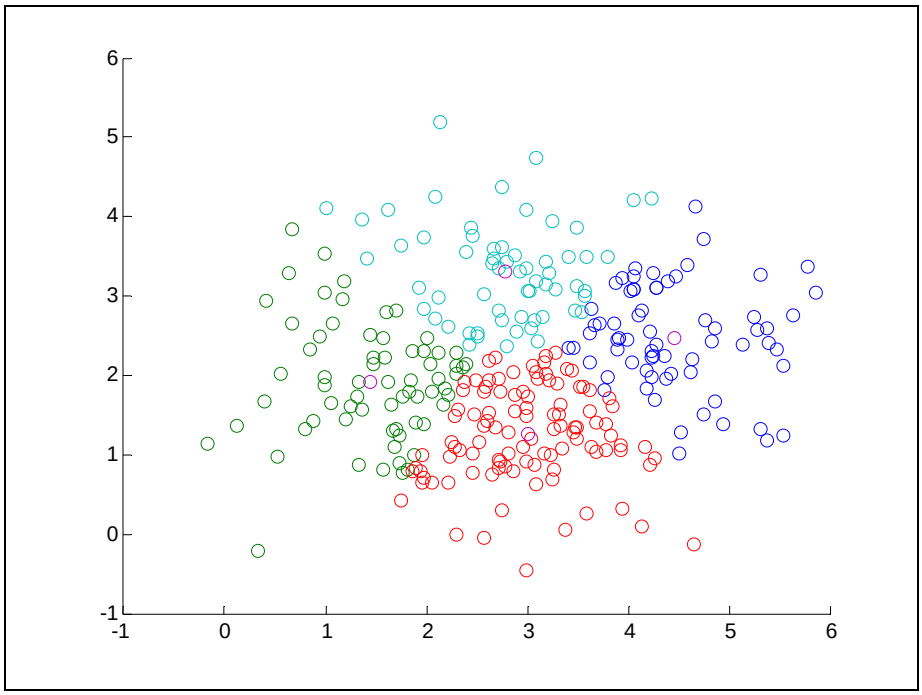


Figure 18: HCM results of data set 3

### 4.2.4 Applications on data set 4

When we apply HCM algorithm to data set4 by changing the number of cluster centers between 2 to 10, the objective function values which were calculated for data set4 are shown in the following table:

Table 9: HCM-Changes in objective function according to the number of clusters for data set 4

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 438.3406 |
| 3 | 114.4488 |
| 4 | 876.6811 |
| 5 | 1095.9 |
| 6 | 1315 |
| 7 | 1534.2 |
| 8 | 1753.4 |
| 9 | 1972.5 |
| 10 | 2191.7 |

According to the table 9, the graph of objective function vs. number of clusters for data set 4 is shown in the following figure.

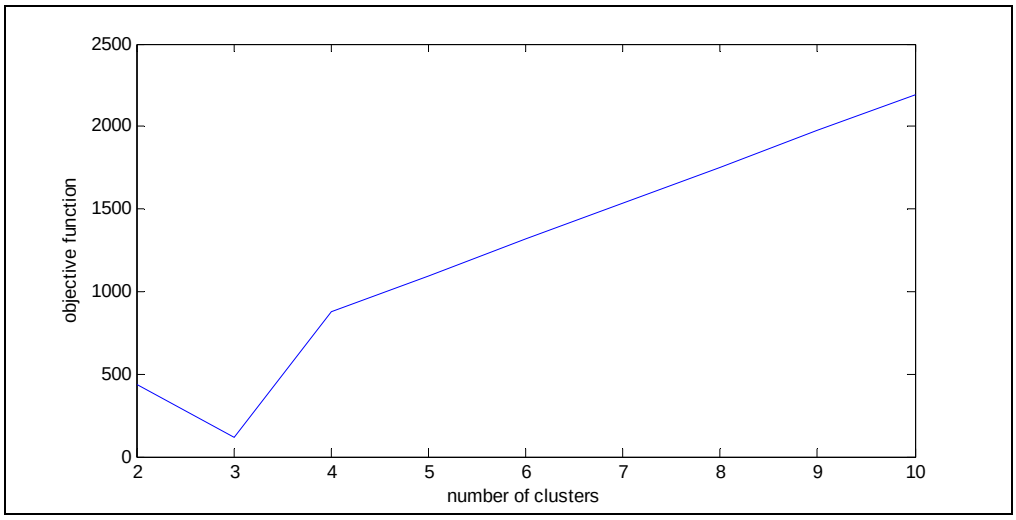


Figure 19: HCM - Objective function vs. the number of clusters for data set4

According to figure 19, it can be easily seen that three clusters is the ideal case for data set4 since, the objective function is minimum for three clusters.

The result of the HCM clustering for data set4 is shown in the following figure.

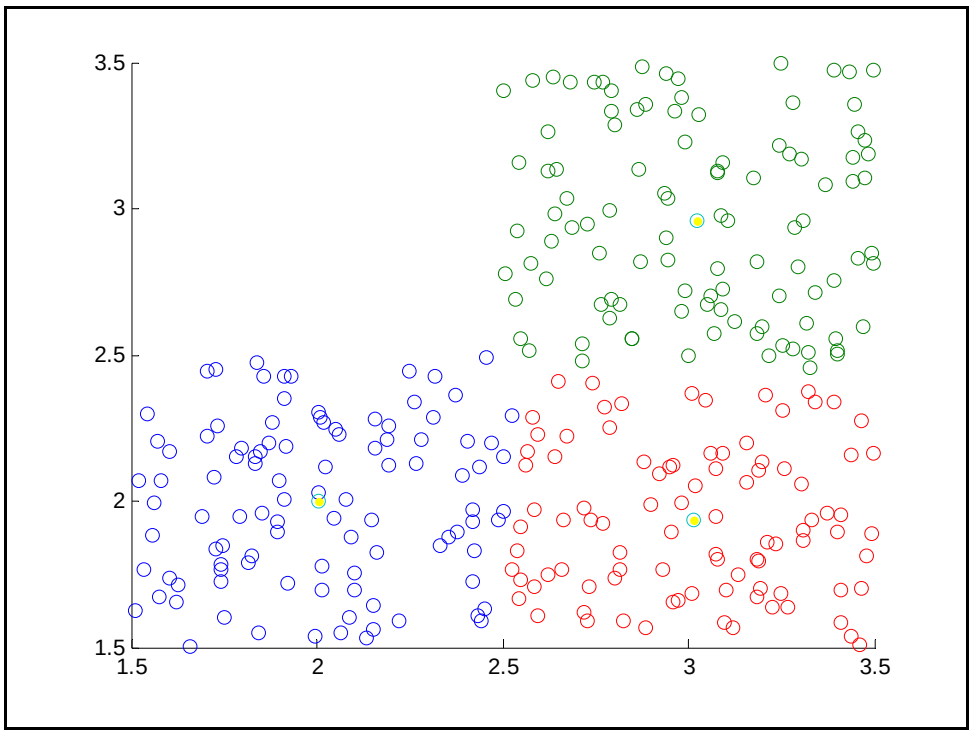


Figure 20: HCM results of data set 4

### 4.3 K-means (KM) Clustering

KM algorithm was applied to the four artificial data sets described in chapter 3. Fuzzy ants initialization was used as an initialization technique for the algorithm.

KM algorithm is sensitive to the selected number of cluster centers. To find good clusters, the aim is to minimize the objective function shown in the equation (2.8).

The following chapters describe the selecting of ideal cluster centers by using objective function calculations and clustering applications on the data sets.

### 4.3.1 Applications on data set 1

When we apply KM algorithm to data set 1 by changing the number of cluster centers between 2 to 10, the objective function values which were calculated for data set1 are shown in the following table:

Table 10: KM-Changes in objective function according to the number of clusters for data set 1

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 340.9845 |
| 3 | 80.8584 |
| 4 | 112.6390 |
| 5 | 124.2412 |
| 6 | 168.8760 |
| 7 | 201.2408 |
| 8 | 220.0654 |
| 9 | 251.1772 |
| 10 | 240.6178 |

According to the table 10, the graph of objective function vs. number of clusters is shown in the following figure.

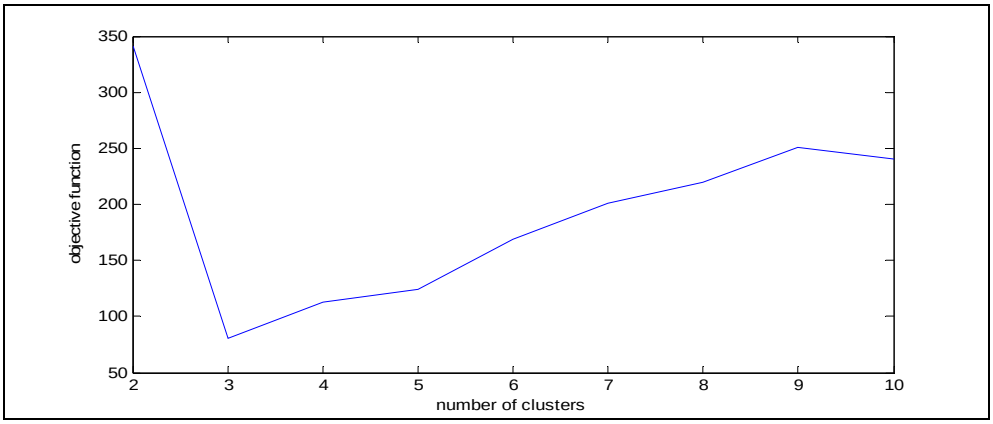


Figure 21: KM - Objective function vs. the number of clusters for data set1

According to figure 21, it can be easily seen that three clusters is the ideal case for data set1 since, the objective function is minimum for three clusters.

The result of the KM clustering for data set1 is shown in the following figure.

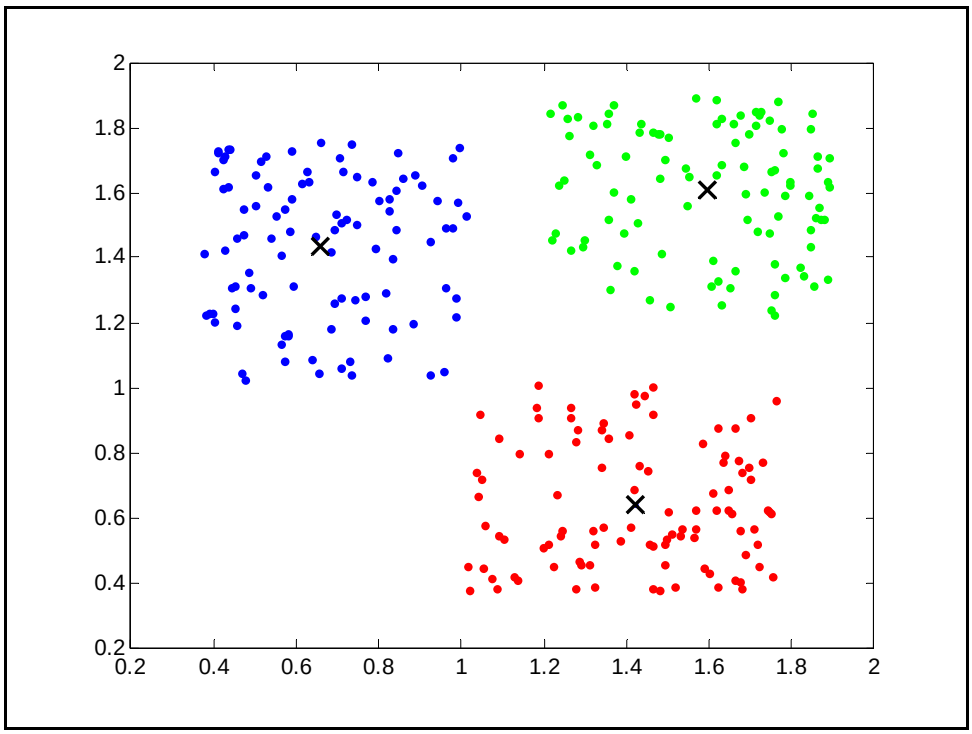


Figure 22: K-means results of data set 1

### 4.3.2 Applications on data set 2

When we apply KM algorithm to data set 2 by changing the number of cluster centers between 2 to 10, the objective function values which were calculated for data set1 are shown in the following table:

Table 11: KM-Changes in objective function according to the number of clusters for data set 2

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 7.9049e+003 |
| 3 | 1.7454e+004 |
| 4 | 1.4088e+004 |
| 5 | 1.2130e+004 |
| 6 | 1.2395e+004 |
| 7 | 9.6667e+003 |
| 8 | 8.6420e+003 |
| 9 | 9.9879e+003 |
| 10 | 9.2537e+003 |

According to the table 11, the graph of objective function vs. number of clusters is shown in the following figure.

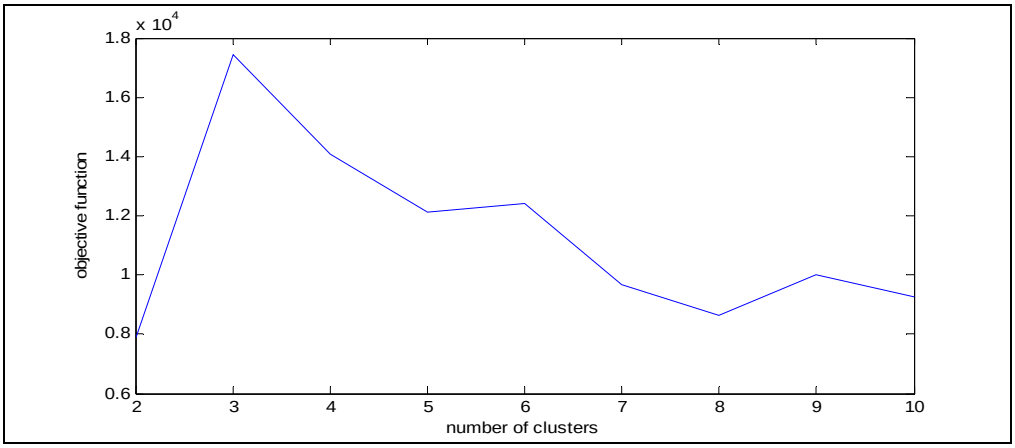


Figure 23: KM - Objective function vs. the number of clusters for data set2

According to figure 23, it can be easily seen that two clusters is the ideal case for data set2 since, the objective function is minimum for two clusters.

The result of the KM clustering for data set2 is shown in the following figure.

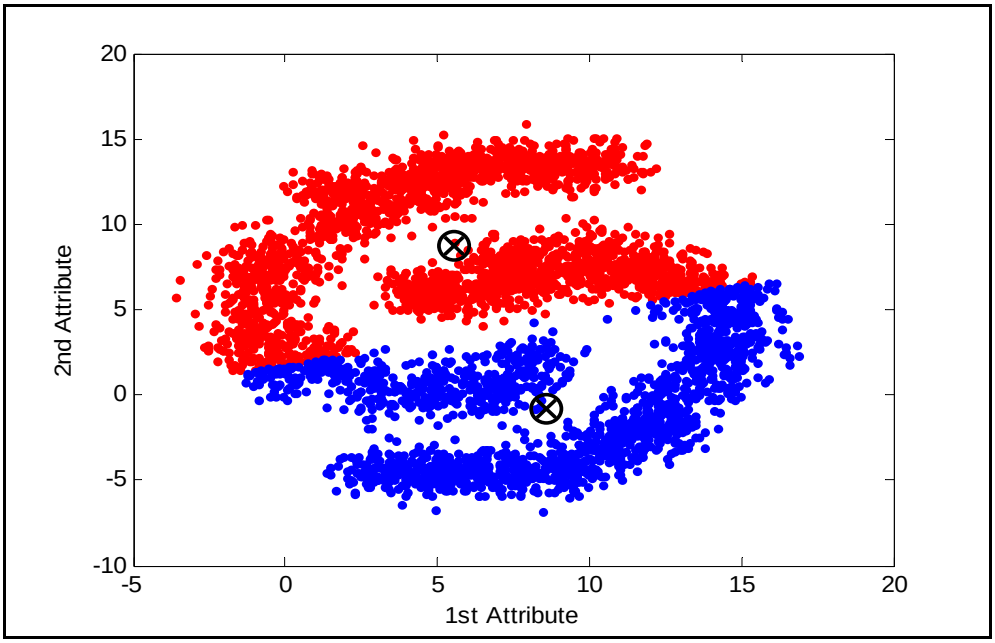


Figure 24: K-means results of data set 2

### 4.2.3 Applications on data set 3

When we apply KM algorithm to data set3 by changing the number of cluster centers between 2 to 10, the objective function values which were calculated for data set3 are shown in the following table:

Table 12: KM-Changes in objective function according to the number of clusters for data set 3

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 320.0041 |
| 3 | 288.9635 |
| 4 | 250.5391 |
| 5 | 300.0712 |
| 6 | 388.0127 |
| 7 | 342.6412 |
| 8 | 335.1682 |
| 9 | 433.6379 |
| 10 | 520.8732 |

According to the table 12, the graph of objective function vs. number of clusters for data set 3 is shown in the following figure.

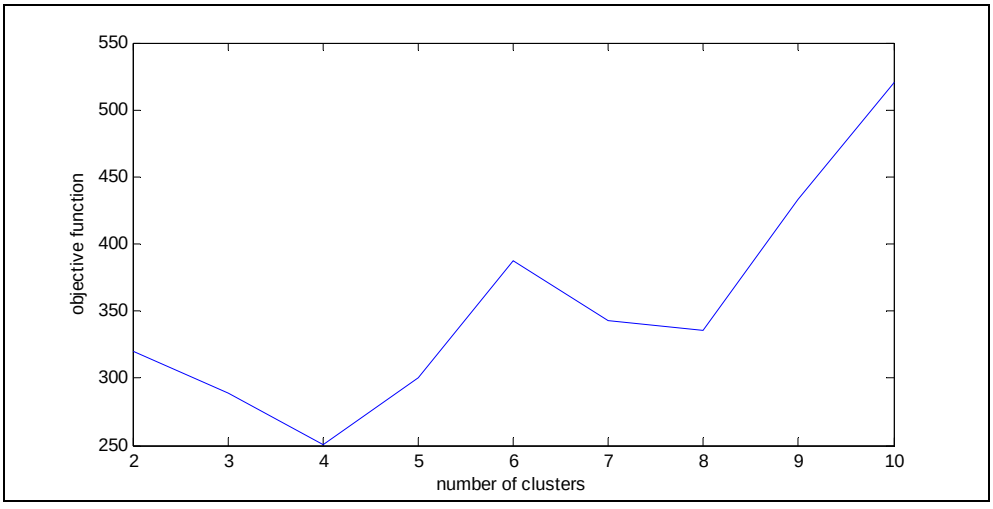


Figure 25: KM - Objective function vs. the number of clusters for data set3

According to figure 25, it can be easily seen that four clusters is the ideal case for data set3 since, the objective function is minimum for four clusters.

The result of the KM clustering for data set3 is shown in the following figure.

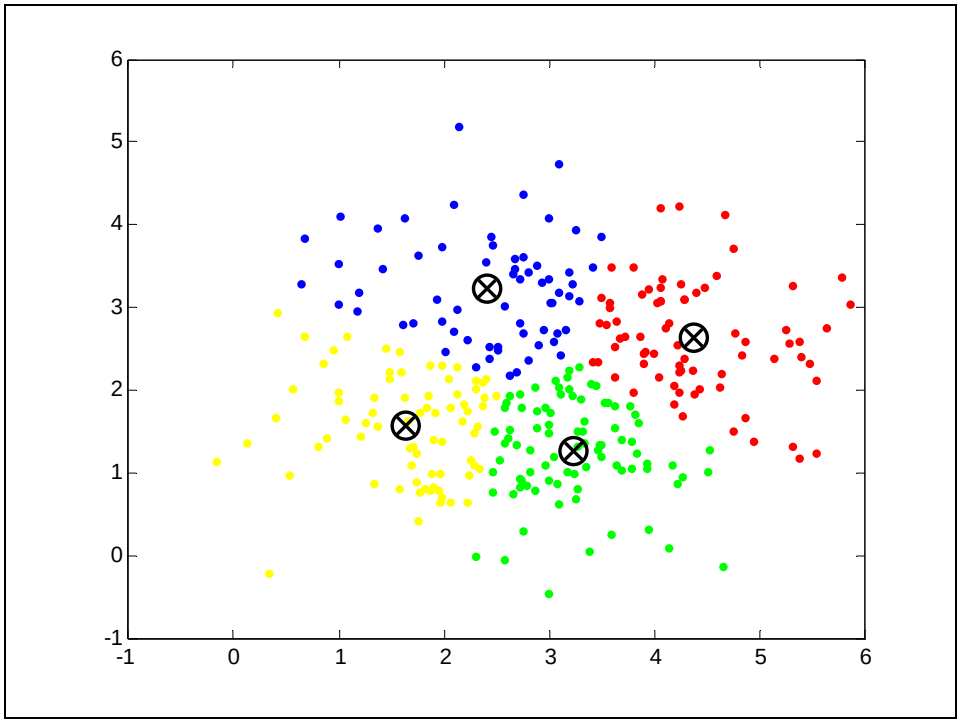


Figure 26: K-means results of data set 3

### 4.34 Applications on data set 4

When we apply KM algorithm to data set4 by changing the number of cluster centers between 2 to 10, the objective function values which were calculated for data set4 are shown in the following table:

Table 13: KM-Changes in objective function according to the number of clusters for data set 4

| Number of Clusters | Value of the Objective Function |
|---|---|
| 2 | 441.6549 |
| 3 | 117.4569 |
| 4 | 879.1168 |
| 5 | 1098.9 |
| 6 | 1318.3454 |
| 7 | 1537.2981 |
| 8 | 1758.4301 |
| 9 | 1975.5777 |
| 10 | 2194.7104 |

According to the table 13, the graph of objective function vs. number of clusters for data set 4 is shown in the following figure.

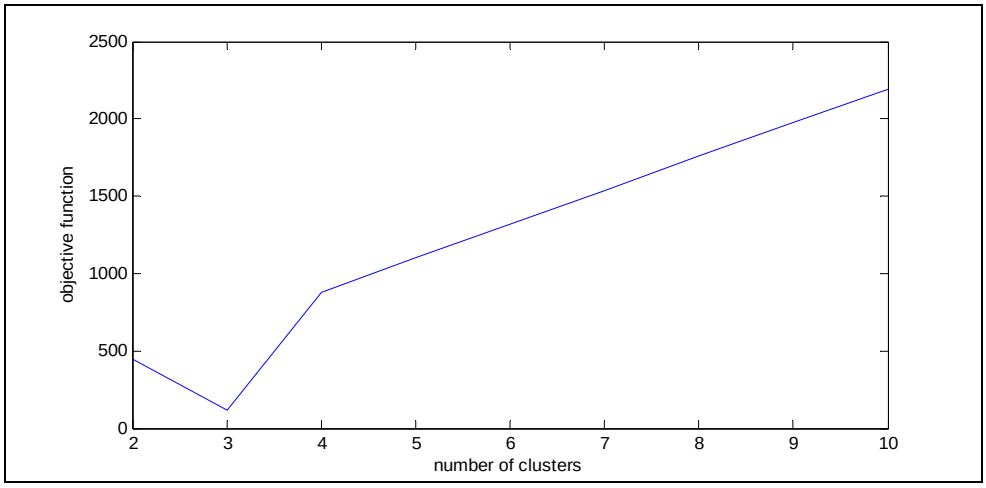


Figure 27: KM - Objective function vs. the number of clusters for data set4

According to figure 27, it can be easily seen that three clusters is the ideal case for data set4 since, the objective function is minimum for three clusters.

The result of the KM clustering for data set4 is shown in the following figure.

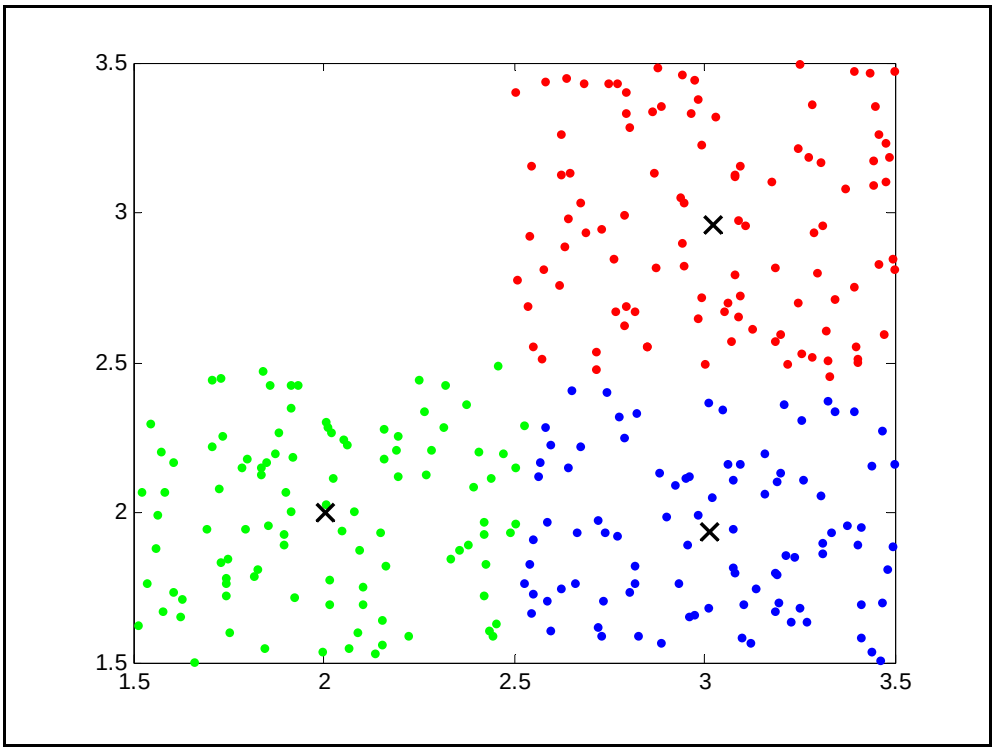


Figure 28: K-means results of data set 4

### 4.4. Similarity Based Clustering (SBC)

SBC algorithm was applied to the four artificial data sets described in chapter 3. The algorithm is not sensitive to initialization of cluster centers and number of clusters since, it assumes all members in the data set as an initial cluster centers.

The results of the algorithm for the data sets described in chapter 3 and hierarchical clustering trees for the data sets are given below.

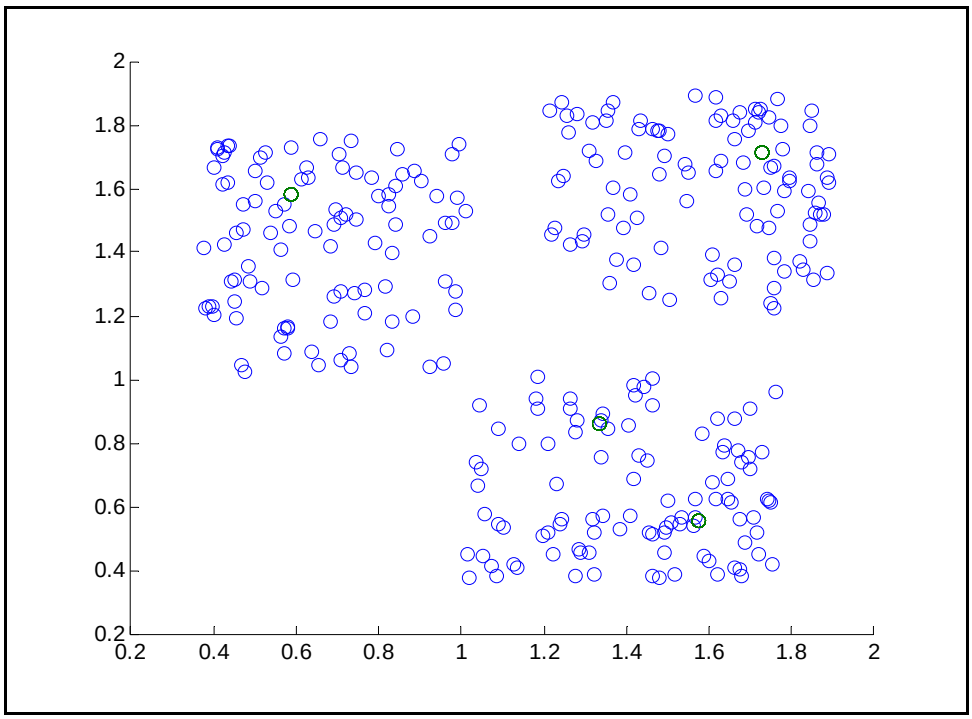


Figure 29: SBC results of data set1

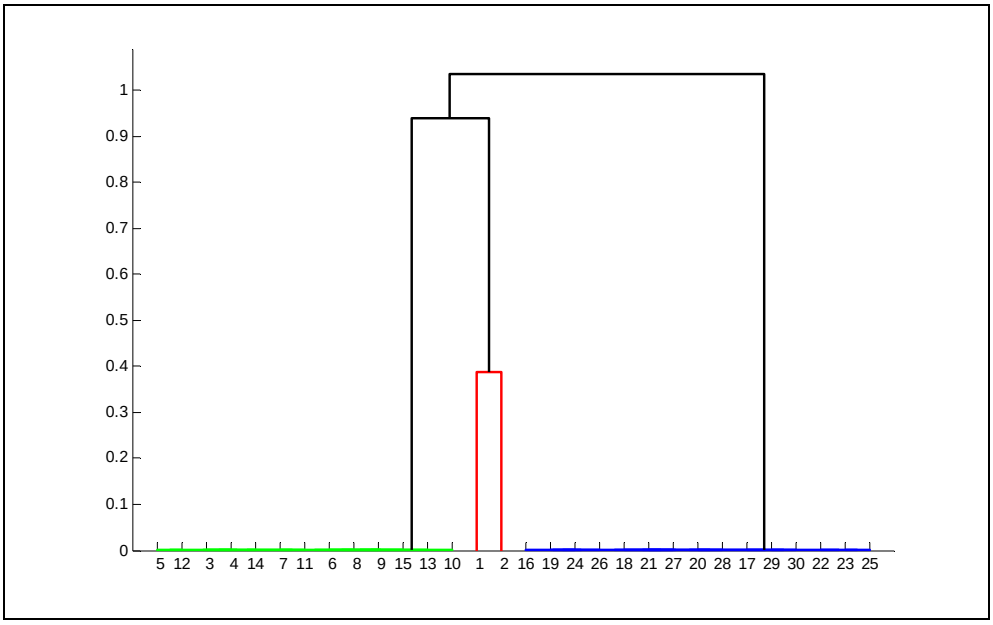


Figure 30: Hierarchical clustering tree of data set1

Cluster centers were computed by calculating the coordinates of the members in the cluster and divided by the total number of members in the clusters.

Cluster centers of data set 1: [(0.5891, 1.5804); (1.7287, 1.7151); (1.4550, 0.70995)].

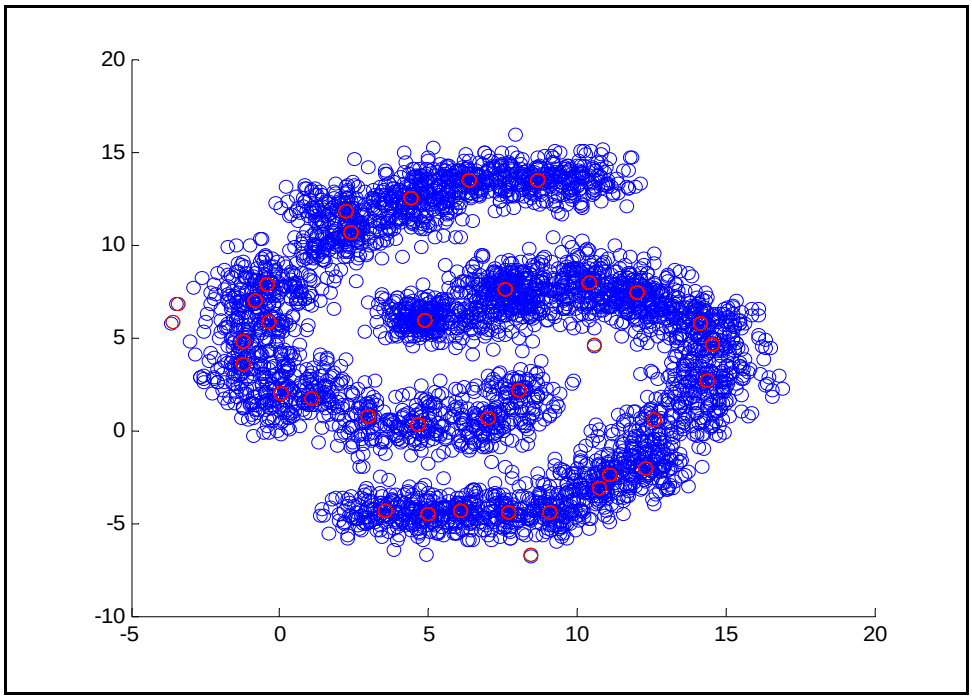


Figure 31: SBC results of data set 2

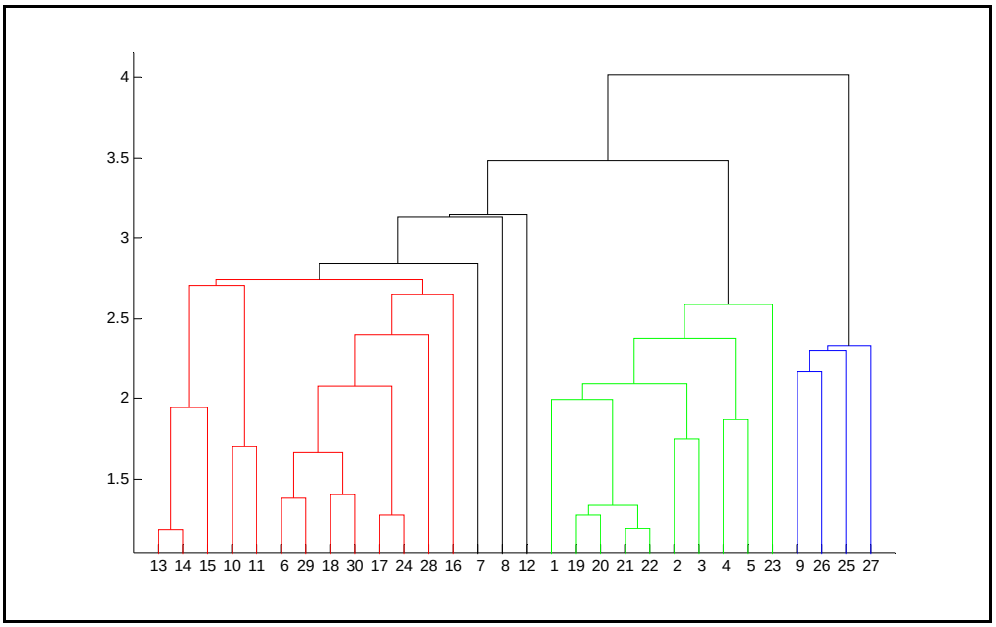


Figure 32: Hierarchical clustering tree of data set2

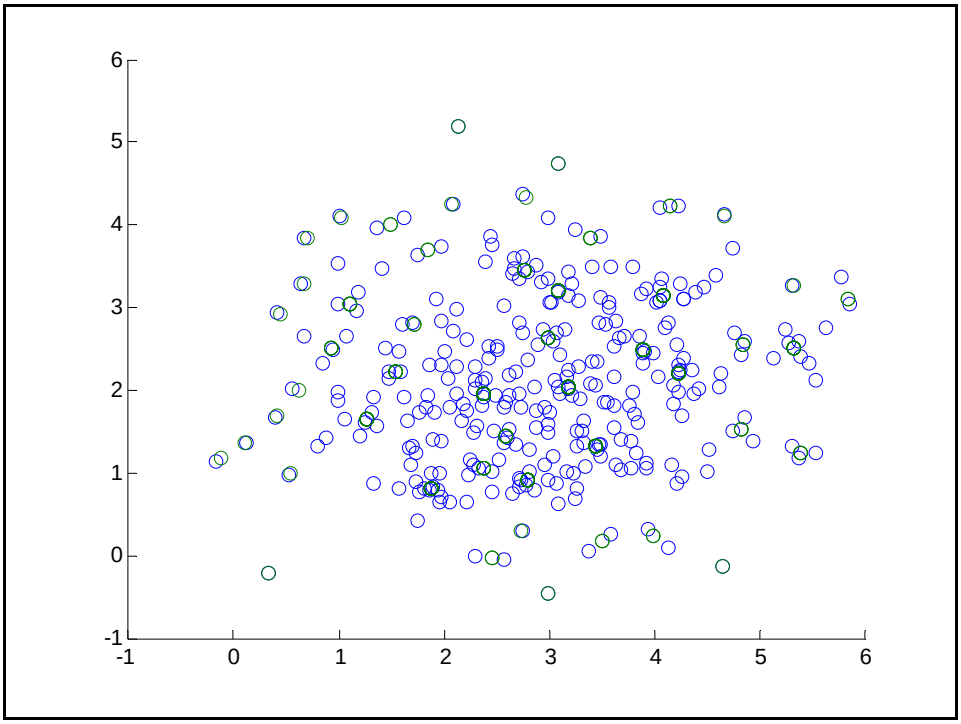


Figure 33: SBC results of data set 3

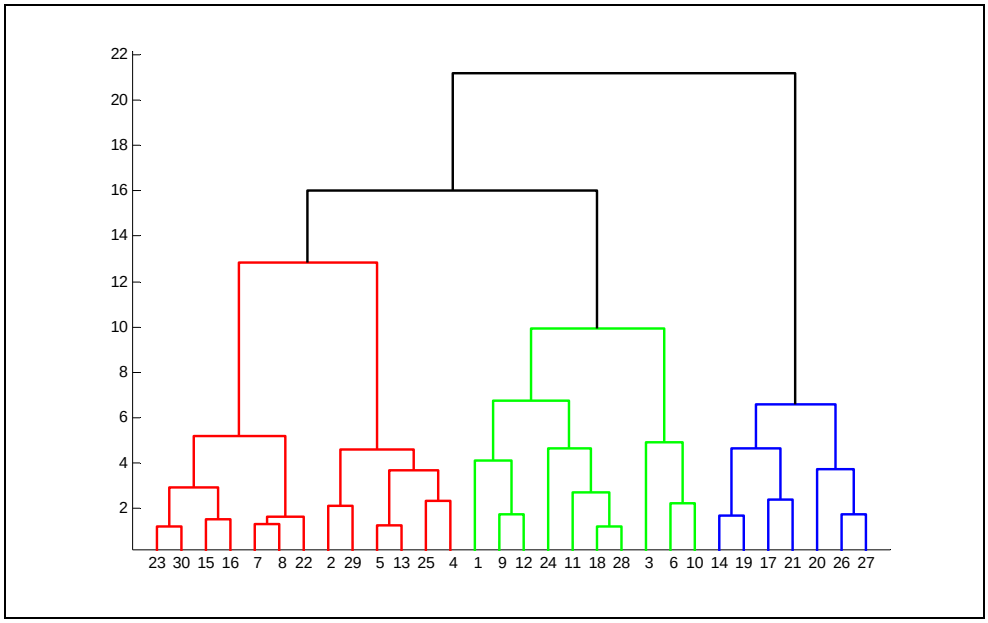


Figure 34: Hierarchical clustering tree of data set3

Cluster centers were computed by calculating the coordinates of the members in the cluster and divided by the total number of members in the clusters.

Cluster centers of data set 3: [(1.8184, 2.9193); (2.7777, 1.5270); (4.4319, 2.6395)].

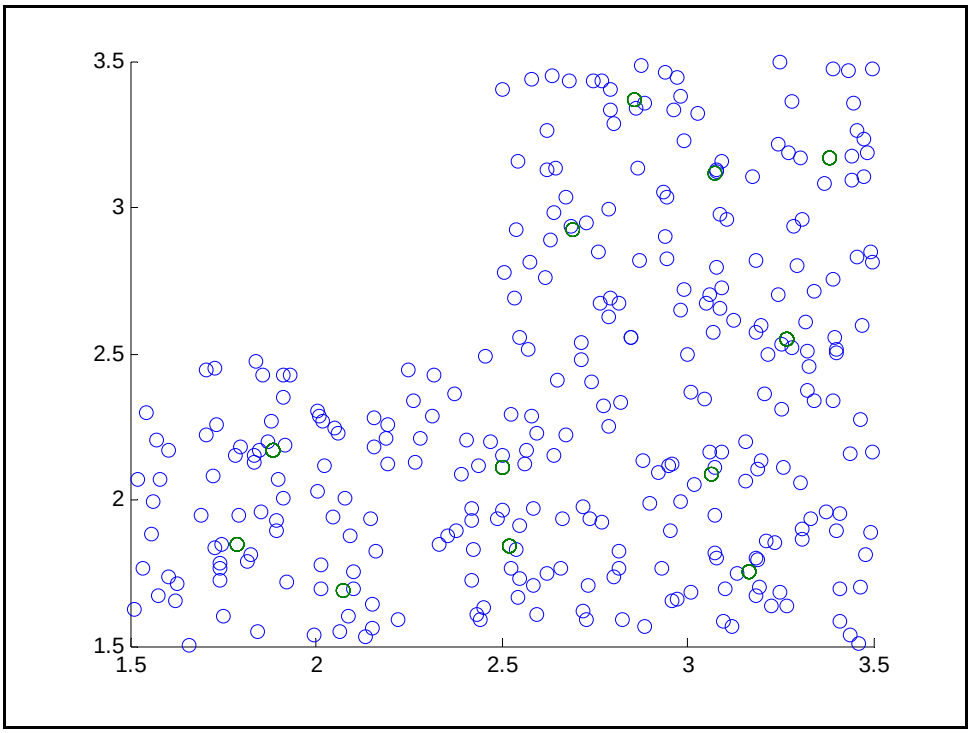


Figure 35: SBC results of data set 4

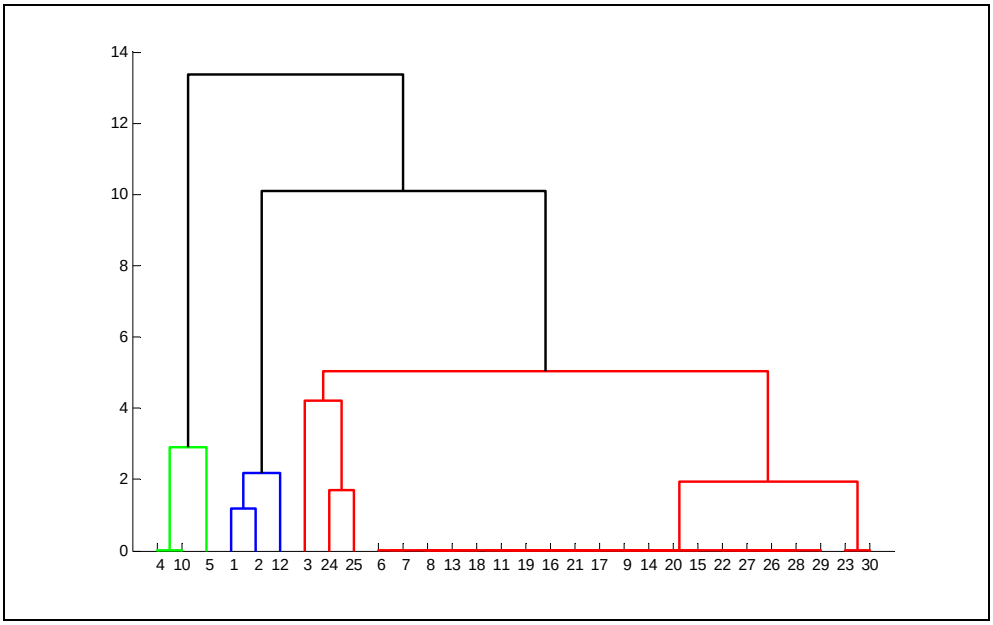


Figure 36: Hierarchical clustering tree of data set 4

Cluster centers were computed by calculating the coordinates of the members in the cluster and divided by the total number of members in the clusters.

Cluster centers of data set 4: [(2.9327, 3.1064); (1.8867, 1.8568); (2.8456, 2.1134)].

# CHAPTER 5

# FUZZY CLUSTERING BASED MULTIPLE CENTERED CLUSTERING METHOD

## 5.1 Multiple centered fuzzy clustering (MCFC)

The clustering algorithms discussed so far are all single-centered. Even if there are relatively dissimilar data or data groups in the same cluster, they are connected to only one center in the cluster. To specify dissimilar data groups in the same cluster, the proposed MCFC algorithm is composed to find out the different cluster centers in the same cluster group. The MCFC algorithm can find the sub-clusters located in the same cluster. Furthermore, and maybe more importantly, the proposed algorithm is able to distinguish clusters having a non-convex structure. This feature of the proposed algorithm makes it very useful in the identification of clusters which are usually impossible to be constructed by using classical clustering algorithms since they employ mostly Euclidean based distance measures.

In general, in FCM, HCM, KM and similarity based clustering algorithms, there is only one cluster center assigned to each cluster. A cluster that has only one cluster center usually forces all the data in the cluster to behave as similar. Similarly, if the Euclidean distance between a member and the cluster center is large enough, it means that the member is inside this cluster but it is not much more similar to the other members and the center. In reality, exactly opposite situations may hold. That is, even though the distance between two cluster members is relatively large, it may be true that these members are very similar to each other. Also, converse of this statement can also be valid as well. Therefore, multiple-centered clustering algorithm suggested in this study, will resolve this seemingly conflicting situation.

The suggested method has an ability to find dissimilar groups in clusters. Hence, it allows the existence of different groups which may have different behaviors in the same cluster.

In MCFC method, FCM clustering algorithm with a relatively high number of clusters is applied on the data set. The algorithm starts with selecting the initial cluster centers and number of clusters. The algorithm stops when the predefined threshold in the FCM algorithm is satisfied. Outputs of this first stage of the overall algorithm are the cluster centers (c) and membership function value of each data member. The details of the standard FCM algorithm are given in chapter 2.2.1. The membership function values can be defined in a "membership matrix" and every cluster has its own membership matrix. The number of columns of this matrix shows the number of data elements in the cluster and the number of rows of the matrix shows the total number of clusters. After performing the FCM algorithm, the MCFC is applied on the data set. At the beginning of the MCFC, the final number of clusters (clast) should be selected.

After selecting the final number of clusters, the similarity functions are calculated. At first, the data set D is described as the following:

$$D = \left( \bar{x}_1, \bar{x}_2, ..., \bar{x}_n \right) \tag{5.1}$$

where;

$\bar{x}_i$ : ith data vector (i = 1,2,…,n)

n: number of data vectors

According to FCM, each data member has membership function value vector, defined as:

$$Y = \left( \bar{y}_1, \bar{y}_2, ..., \bar{y}_c \right) \tag{5.2}$$

where;

$\bar{y}_i$ : Membership value vector of ith data member (i = 1,2,…,c)

c: Number of clusters

The membership function value vectors consist of the membership values of the data members to all clusters. This vector can be represented as:

$$\bar{y}_i = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_c \end{pmatrix}, \ \mu_j \geq 0, \ \sum \mu_j = 1 \tag{5.3}$$

where;

$\mu_j$ : j-th cluster membership function value of the i-th data vector $\overline{x_i}$ whose representative is $\overline{y}_i$ (j = 1,2,….,c ).

From expression (5.3), it can be seen that sum of membership function values of each data member is equal to 1.

MCFC algorithm depends on merging processes between clusters. The merging process is according to the similarity ratios between clusters. This ratio depends on the total similarity of cluster and inter-cluster similarities between clusters.

In MCFC, there is no definite assignment to the clusters since, this method is fuzzy. So, the clusters are formed according to the following statement:

$$\bar{x}_i \in C_k \quad \text{if} \quad \left(\bar{y}_i\right)_k \geq \left(\bar{y}_i\right)_j, \; j = 1, 2, ..., c \tag{5.4}$$

where;

c: total number of clusters

$C_k$ : k-th cluster

The total similarity (this is the generalization of the number of elements in a given crisp set) of the k-th cluster is the sum of k-th cluster membership values of the data members which belongs to k-th cluster. The expression is:

$$S(k) = \sum_{i \in \lambda_k} \left(\bar{y}_i\right)_k \tag{5.5}$$

where;

$\lambda_k$ : the indices of elements belonging to $C_k$

The total similarity of the jth cluster in the ith cluster is described as:

$$S_i(j) = \sum_{k \in \lambda_j} \left(\bar{y}_k\right)_i \tag{5.6}$$

The similarity ratio to be defined below is an indicator of measuring the similarities of the clusters.

**Definition (5.1):** The similarity of the j-th cluster to the i-th cluster is given by SR(i,j) defined by using the following formula:

$$SR(i,j) = S_i(j) / S(i) \quad (\text{i, j} = 1,\ldots,\text{c}). \tag{5.7}$$

**Corollary (5.1):** *If* $i = j, SR(i,j) = 1$ *and* $S_i(i) = S(i)$.

**Corollary (5.2):** Similarity measure between two clusters i and j is not symmetrical, that is $SR(i,j)$ *is not equal to* $SR(j,i)$.

**Corollary (5.3):** $\sum_{i=1}^{M} SR(i,j) =$ *number of elements in* $C_i$.

After finding SR(j,i) values, i and j indices which correspond to the maximum value of SR(j,i) is selected and the algorithm starts merging the i-th and j-th clusters accordingly. This step of the algorithm must be performed until the similarity values get under a similarity threshold. During the iterations, a threshold ϵ to stop merging processes between the clusters is selected. This threshold should be higher than the minima of similarity ratios of the clusters. If maxima of the similarity ratios are smaller than ϵ, the clusters can not be merged. The iterations continue till the number of cluster groups reaches the value “clast”. When the clusters are merged, the membership of the merged clusters should be evaluated according to the new cluster assignments. During the merging process, the data in the clusters are merged but the coordinates of the original cluster centers do not change. While the data in two similar clusters are merging and creating a new group, they are clustered in a new group but this group has two more cluster centers. As a result of these processes, the clusters including multiple cluster centers are constructed. The table below shows the steps of the algorithm in general.

Table 14: Fuzzy Clustering Based Multiple-Centered Clustering Algorithm (MCFC)

| |
|---|
| 1. Select the number of clusters (c) and initial cluster centers. |
| 2. Apply fuzzy c-means algorithm and find the clusters, cluster centers and membership values of each data. |
| 3. Define the number of groups (clast) that are desired to be reached at the end of the algorithm and define the threshold $\epsilon$. |
| 4. Calculate the total of each cluster membership and find the similarity ratios for the clusters. |
| 5. Start the merging process from the cluster that has maximum similarity ratio to the other cluster. |
| 6. After merging, decrease c. |
| 7. If c is equal to clast, stop the processes and find the new clusters with multiple-centers. Else, continue from step 3. |

### 5.2 Experiments and Results

The proposed MCFC algorithm was applied on the artificial data sets defined in chapter 3. The following figures are the results of these applications of the MCFC algorithm on artificial data sets which are the mixtures of Gaussian distributions.

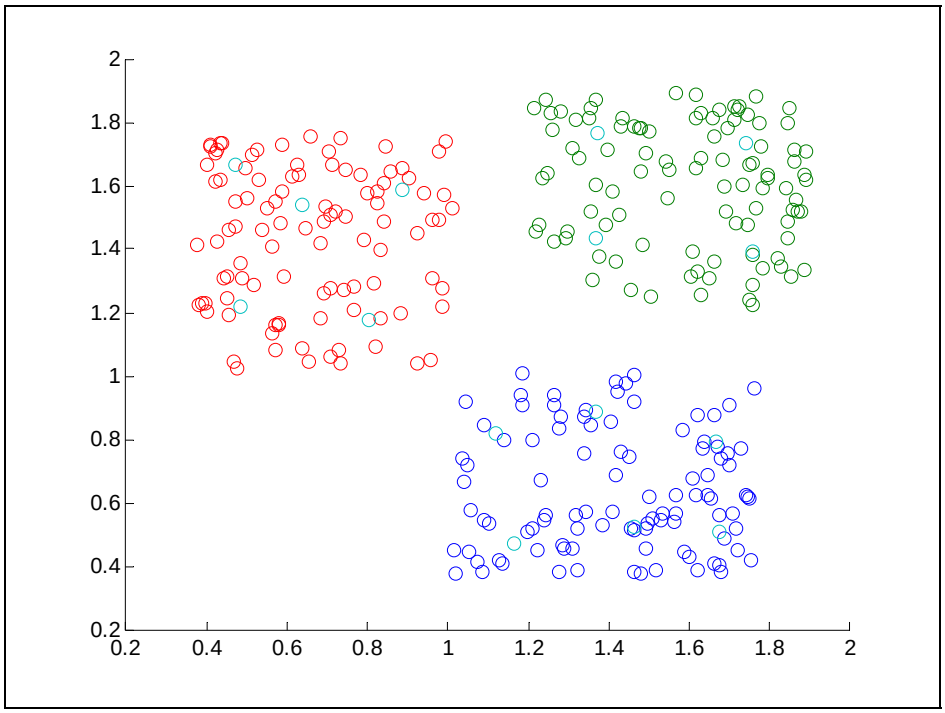


Figure 37: MCFC results of data set 1

In figure 37, different colors shows three different clusters and blue “o” signs shows the cluster centers in these clusters. It can be easily seen that each cluster has more than one cluster center. The total number of cluster centers for this application is fifteen. The result of the FCM, HCM and KM algorithms are similar to the MCFC algorithm but the difference is that these algorithms have only one cluster center for each group.

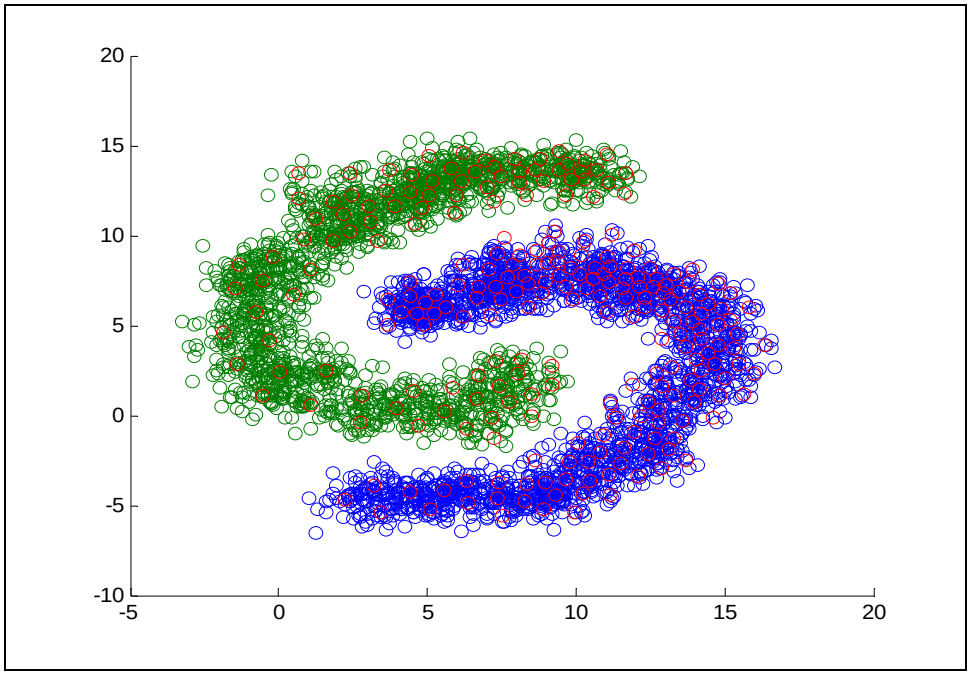


Figure 38: MCFC results of data set2

In figure 38, the application of MCFC algorithm on the data set 2 with 250 cluster centers can be seen. The blue and green colors show the two different clusters and the red circles show the cluster centers of the defined groups. The execution time is too high to find the 250 different cluster centers and to merge them into two different groups. Fuzzy ants initialized cluster centers were used in this algorithm. The data set is sensitive to the selection of the similarity threshold value for producing a reasonably successful clustering.

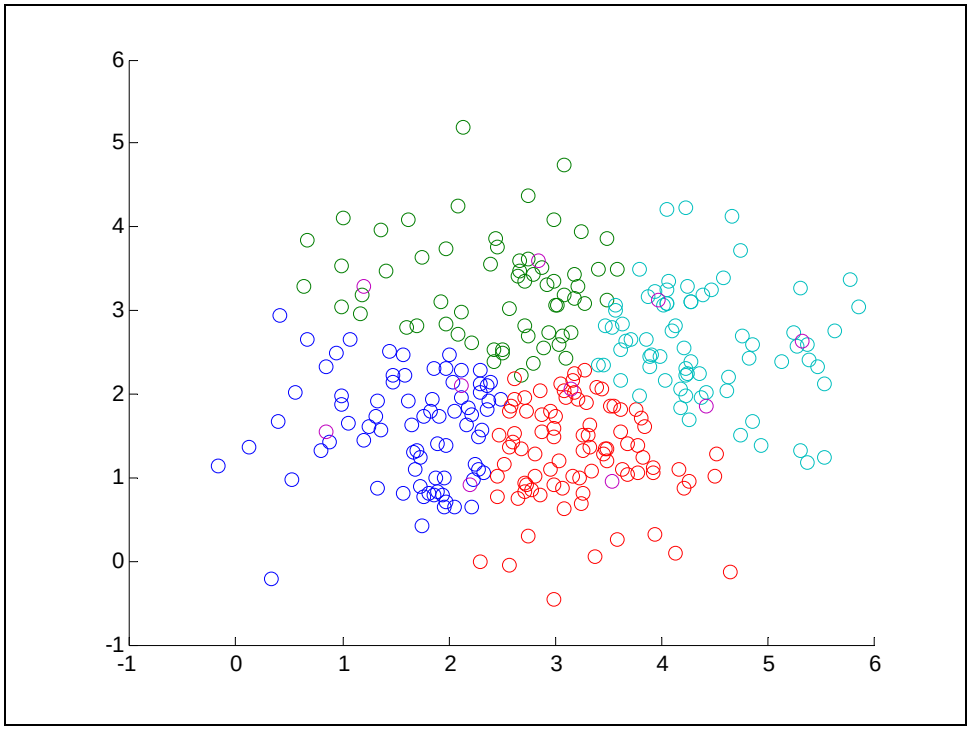


Figure 39: MCFC results of data set 3

In figure 39, the MCFC result of data set 3 can be seen. The different colors show the different clusters and the purple “o” signs show the cluster centers of the clusters. The total number of cluster centers for this application is ten. The shape of the clusters is similar to the FCM, HCM and KM applications defined in chapter 4.

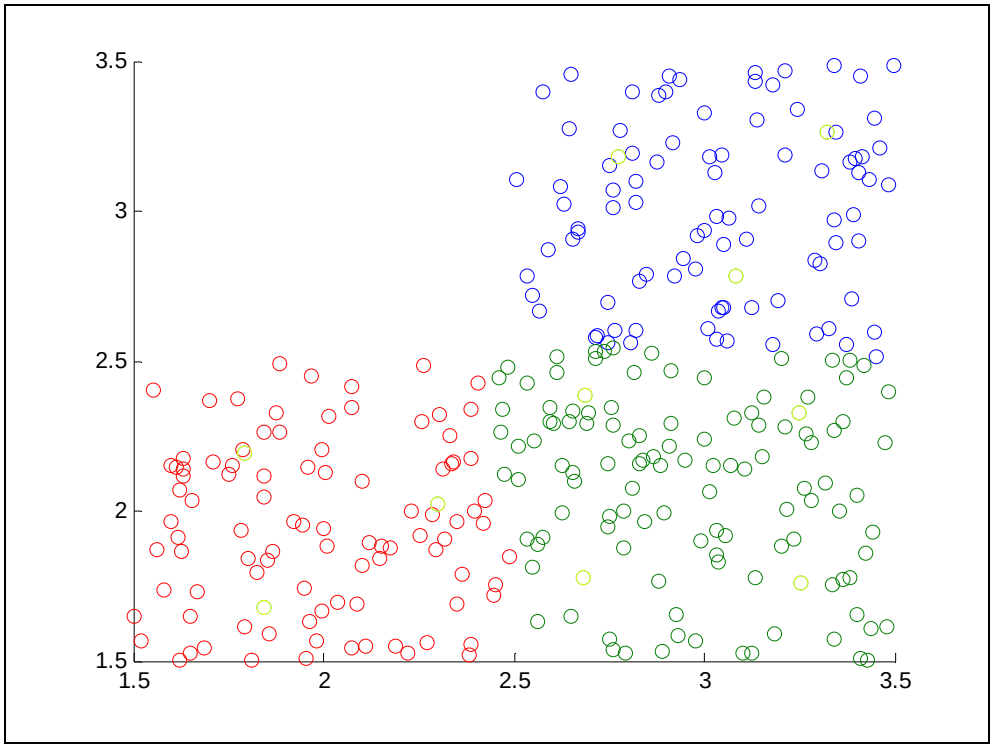


Figure 40: MCFC results of data set 4

In figure 40, the results of the MCFC by using ten cluster centers are shown. The shape of the clusters is similar to the other clustering applications defined in chapter 3.

High number of clusters with an intelligent selection of the threshold value can decrease the sensitivity of the algorithm. Therefore, defining the best threshold value according to the distances between cluster centers is an important case for this method and it helps to find the most reasonable solution for clustering processes.

The number of clusters which is selected at the beginning stage of the algorithm do not change the results of the algorithm. Therefore, the algorithm is robust to initial number of clusters selected at the first stage.

# CHAPTER 6

# DEVELOPMENT OF DECISION MAKING SYSTEM ALGORITHM ON MENOPAUSE DATA SET

In this section, development of the decision making system is described. The development of decision making system depends on the clustering results. Therefore, the results of the clustering algorithms are described in the following.

## 6.1 Single-Centered Clustering

In this part, the results of the single-centered clustering algorithms which were defined in chapter two were compared. The real medical data set is used in the clustering analysis which includes menopause data members. Menopause data members were collected from the laboratory test results of the women in menopause duration at the Hacettepe University, Faculty of Medicine Hospital. In the data matrix, some of the members were missing for some dimensions. The missing values were filled by using “imputation by mean” method which was defined in chapter 2 for FCM, HCM, KM, SBC algorithms. For FCM with missing values algorithm, there is no need to assign the mean imputation for the missing values initially. After completing the data matrix processes, fuzzy ant initialization was applied to the data set. Then, main clustering algorithms were executed for the real data set.

At first, the number of cluster centers was selected as three, four, five, six, seven and eight. The medical investigation showed that results of five clusters are much better than the medical results of other clustering analysis when compared by the doctors. So, number of five clusters is selected in order to construct the decision making system. The FCM, HCM, KM and FCM without imputation algorithms are sensitive to the selection of the number of cluster centers but, SBC algorithm can find the ideal number of clusters by itself. The number of clusters which was found by SBC was computed as four. So, the results of the SBC were different from the other algorithms. The results are shown in the following tables.

Table 15: FCM Results

| | Age | Weight | Height | Mens Duration | FSH | LH | Estradiol |
|---|---|---|---|---|---|---|---|
| Center1 | 53.1486 | 64.3600 | 161.2819 | 66.9066 | 55.6719 | 23.5095 | 47.3417 |
| Center2 | 50.8623 | 66.2802 | 160.0161 | 41.9383 | 99.7066 | 40.4597 | 48.1106 |
| Center3 | 47.4117 | 61.7364 | 160.4082 | 45.8723 | 58.4312 | 24.4635 | 60.1050 |
| Center4 | 51.4516 | 64.0653 | 159.3962 | 58.1185 | 74.8273 | 34.5873 | 88.1688 |
| Center5 | 53.4833 | 62.1927 | 159.8406 | 79.9777 | 53.0494 | 23.0000 | 48.4950 |

| | T3 | T4 | TSH | Glucose | Cholesterol | Triglyceride | HDL |
|---|---|---|---|---|---|---|---|
| Center1 | 0.6083 | 4.9182 | 2.3321 | 58.1994 | 139.2785 | 66.0289 | 41.1250 |
| Center2 | 1.0216 | 8.1214 | 2.4136 | 89.3879 | 228.2073 | 92.2364 | 65.9211 |
| Center3 | 0.7834 | 6.6909 | 6.4811 | 75.8494 | 158.8233 | 81.3673 | 48.4664 |
| Center4 | 0.9708 | 8.7056 | 2.7500 | 91.3671 | 220.2330 | 109.5055 | 67.5453 |
| Center5 | 0.6280 | 5.2840 | 2.7704 | 60.2501 | 143.1874 | 66.8512 | 41.2766 |

| | LDL | VLDL | Hemoglobin | Hematocrite | Men. Type | HRT | HRT Duration |
|---|---|---|---|---|---|---|---|
| Center1 | 82.3368 | 14.2486 | 8.5095 | 24.9059 | 1.0556 | 1.9880 | 0.5018 |
| Center2 | 143.0941 | 15.5970 | 13.6812 | 40.0552 | 1.1480 | 1.9921 | 0.3672 |
| Center3 | 99.1547 | 16.9564 | 10.4884 | 31.1015 | 3.2827 | 1.9851 | 0.6732 |
| Center4 | 130.1173 | 23.3308 | 13.4899 | 39.6504 | 1.6329 | 1.0150 | 34.8350 |
| Center5 | 90.1229 | 15.0223 | 8.7854 | 25.8184 | 1.5222 | 1.0085 | 43.7213 |

Table 16: KM Results

| | Age | Weight | Height | Mens Duration | FSH | LH | Estradiol |
|---|---|---|---|---|---|---|---|
| Center1 | 54.2706 | 63.7326 | 160.9048 | 74.4574 | 54.8738 | 22.5091 | 40.3450 |
| Center2 | 50.4758 | 66.5084 | 160.3745 | 40.3837 | 96.6965 | 40.0243 | 57.5677 |
| Center3 | 53.4597 | 61.0900 | 159.9041 | 81.2468 | 51.3619 | 22.0738 | 41.6521 |
| Center4 | 47.4305 | 61.5272 | 160.1916 | 45.3329 | 59.2294 | 24.2597 | 56.5682 |
| Center5 | 51.3903 | 64.6547 | 159.3178 | 59.5462 | 75.0554 | 34.7077 | 94.2429 |

| | T3 | T4 | TSH | Glucose | Cholesterol | Triglyceride | HDL |
|---|---|---|---|---|---|---|---|
| Center1 | 0.5579 | 4.6288 | 2.1486 | 55.7713 | 136.4640 | 64.1339 | 40.8760 |
| Center2 | 1.0619 | 8.1906 | 2.1690 | 89.2042 | 224.8847 | 94.0241 | 65.1415 |
| Center3 | 0.6087 | 5.1202 | 2.0467 | 56.4344 | 130.6787 | 61.3809 | 39.7282 |
| Center4 | 0.7760 | 6.7073 | 7.6738 | 76.4514 | 155.4225 | 80.5449 | 48.0766 |
| Center5 | 0.9498 | 8.5254 | 4.1364 | 92.6322 | 225.2365 | 110.2206 | 65.8237 |

| | LDL | VLDL | Hemoglobin | Hematocrite | Men. Type | HRT | HRT Duration |
|---|---|---|---|---|---|---|---|
| Center1 | 79.3781 | 13.7272 | 8.1196 | 23.7807 | 1.0432 | 1.9893 | 0.4039 |
| Center2 | 140.5274 | 19.7147 | 13.7363 | 40.1307 | 1.1455 | 2.0000 | 0 |
| Center3 | 85.5291 | 12.9704 | 8.2189 | 24.1100 | 1.5383 | 1.0000 | 43.5744 |
| Center4 | 97.3899 | 16.3510 | 10.4700 | 31.0740 | 3.3809 | 2.0000 | 0 |
| Center5 | 132.5570 | 25.6181 | 13.5671 | 39.3523 | 1.6586 | 1.0000 | 36.5976 |

Table 17: HCM Results

| | Age | Weight | Height | Mens Duration | FSH | LH | Estradiol |
|---|---|---|---|---|---|---|---|
| Center1 | 53.9924 | 63.9323 | 160.9309 | 73.7912 | 55.4667 | 23.9077 | 48.7257 |
| Center2 | 50.5936 | 66.4228 | 160.3409 | 40.2251 | 97.0385 | 39.3945 | 52.1087 |
| Center3 | 47.4305 | 61.5272 | 160.1916 | 45.3329 | 59.2294 | 24.2597 | 56.5682 |
| Center4 | 51.3903 | 64.6547 | 159.3178 | 59.5462 | 75.0554 | 34.7077 | 94.2429 |
| Center5 | 53.4597 | 61.0900 | 159.9041 | 81.2468 | 51.3619 | 22.0738 | 41.6521 |

| | T3 | T4 | TSH | Glucose | Cholesterol | Triglyceride | HDL |
|---|---|---|---|---|---|---|---|
| Center1 | 0.5710 | 4.7398 | 2.1180 | 55.7611 | 136.3467 | 64.0624 | 40.8218 |
| Center2 | 1.0623 | 8.1796 | 2.1996 | 89.8368 | 226.6046 | 94.6319 | 65.6288 |
| Center3 | 0.7760 | 6.7073 | 7.6738 | 76.4514 | 155.4225 | 80.5449 | 48.0766 |
| Center4 | 0.9498 | 8.5254 | 4.1364 | 92.6322 | 225.2365 | 110.2206 | 65.8237 |
| Center5 | 0.6087 | 5.1202 | 2.0467 | 56.4344 | 130.6787 | 61.3809 | 39.7282 |

| | LDL | VLDL | Hemoglobin | Hematocrite | Men. Type | HRT | HRT Duration |
|---|---|---|---|---|---|---|---|
| Center1 | 79.4040 | 13.7104 | 8.2557 | 24.1374 | 1.0420 | 1.9895 | 0.3917 |
| Center2 | 141.6407 | 19.8493 | 13.7474 | 40.1856 | 1.1480 | 2.0000 | 0 |
| Center3 | 97.3899 | 16.3510 | 10.4700 | 31.0740 | 3.3809 | 2.0000 | 0 |
| Center4 | 132.5570 | 25.6181 | 13.5671 | 39.9523 | 1.6586 | 1.0000 | 36.5976 |
| Center5 | 85.5291 | 12.9704 | 8.2189 | 24.1100 | 1.5383 | 1.0000 | 43.5744 |

Table 18: FCM Missing Results

| | Age | Weight | Height | Mens Duration | FSH | LH | Estradiol |
|---|---|---|---|---|---|---|---|
| Center1 | 58.7394 | 65.5673 | 159.9153 | 128.0109 | 105.4988 | 42.9698 | 49.1101 |
| Center2 | 49.2737 | 65.4152 | 161.0504 | 26.3291 | 100.6644 | 40.6171 | 56.2606 |
| Center3 | 47.2703 | 61.6603 | 160.2513 | 12.6869 | 65.4777 | 26.9829 | 77.0971 |
| Center4 | 49.5094 | 61.8314 | 158.9818 | 34.2606 | 88.3730 | 41.3488 | 76.7126 |
| Center5 | 56.3306 | 64.5882 | 160.4567 | 122.1416 | 54.2353 | 25.5194 | 125.5360 |

| | T3 | T4 | TSH | Glucose | Cholesterol | Triglyceride | HDL |
|---|---|---|---|---|---|---|---|
| Center1 | 1.2668 | 9.4937 | 3.1578 | 89.9898 | 229.5363 | 106.2877 | 70.0466 |
| Center2 | 1.0938 | 8.5819 | 2.3015 | 90.1633 | 228.4418 | 91.2711 | 64.9250 |
| Center3 | 1.0247 | 9.0296 | 11.5680 | 95.9276 | 186.4596 | 100.7458 | 59.8683 |
| Center4 | 1.1345 | 9.7194 | 2.9947 | 91.1121 | 207.6462 | 103.6419 | 62.8351 |
| Center5 | 1.0317 | 9.9705 | 3.8408 | 91.3773 | 239.3478 | 120.4462 | 71.9307 |

| | LDL | VLDL | Hemoglobin | Hematocrite | Men. Type | HRT | HRT Duration |
|---|---|---|---|---|---|---|---|
| Center1 | 130.5067 | 22.9776 | 14.4426 | 41.5990 | 1.1387 | 1.9916 | 0.4284 |
| Center2 | 144.8590 | 19.3447 | 13.7805 | 40.4258 | 1.0609 | 1.9968 | 0.1591 |
| Center3 | 121.1374 | 21.1443 | 13.3207 | 39.9729 | 3.2768 | 1.9897 | 0.4284 |
| Center4 | 128.8456 | 21.9516 | 13.8724 | 40.6796 | 1.4729 | 1.0055 | 17.2584 |
| Center5 | 140.0316 | 26.9132 | 13.8393 | 40.9746 | 1.7187 | 1.0035 | 69.2050 |

Table 19: SBC Results

| | Age | Weight | Height | Mens. Duration | FSH | LH | Estradiol |
|---|---|---|---|---|---|---|---|
| Center1 | 51.4186 | 64.6262 | 159.6502 | 58.2454 | 74.6221 | 34.5132 | 91.7650 |
| Center2 | 53.5823 | 60.8523 | 159.3552 | 85.4029 | 50.1989 | 21.3698 | 41.8059 |
| Center3 | 50.4852 | 66.3943 | 160.2961 | 40.5423 | 98.3840 | 40.2466 | 56.8757 |
| Center4 | 51.7627 | 62.9722 | 160.7218 | 63.9878 | 54.0528 | 22.6295 | 46.7266 |

| | T3 | T4 | TSH | Glucose | Cholesterol | Triglyceride | HDL |
|---|---|---|---|---|---|---|---|
| Center1 | 0.9428 | 8.4617 | 4.0141 | 91.6732 | 219.9985 | 108.3256 | 65.0874 |
| Center2 | 0.5934 | 4.9491 | 2.0772 | 55.1387 | 132.1641 | 60.6658 | 38.8728 |
| Center3 | 1.0588 | 8.2233 | 2.1588 | 89.3573 | 223.7902 | 93.4520 | 65.0874 |
| Center4 | 0.6341 | 5.3113 | 4.2384 | 62.7701 | 143.1092 | 70.3193 | 43.2040 |

| | LDL | VLDL | Hemoglobin | Hematocrite | Men. Type | HRT | HRT Duration |
|---|---|---|---|---|---|---|---|
| Center1 | 133.1799 | 24.9622 | 13.4421 | 39.5818 | 1.6511 | 1.0000 | 36.3650 |
| Center2 | 80.5200 | 13.0881 | 7.9909 | 23.4513 | 1.5417 | 1.0000 | 44.5781 |
| Center3 | 139.5904 | 19.5970 | 13.7290 | 40.1170 | 1.1787 | 2.0000 | 0 |
| Center4 | 85.9703 | 14.7364 | 8.9104 | 26.2369 | 1.8849 | 1.9931 | 0.2570 |

When the results of the single-centered clustering methods are compared, it is seen that the results of FCM algorithm are quite meaningful with respect to the other algorithms according to the medical experts. The results of the algorithms were evaluated by the doctors and instructors of Hacettepe University, Faculty of Medicine, Gynecology Department. Therefore, it was decided that results of FCM algorithm can be used during the development of decision making system for single-centered clustering.

The medical meanings of the cluster centers are described in the following table.

Table 20: Medical meanings of cluster centers found by FCM method

| FCM Method | |
|---|---|
| Center1 | The patient is suitable for spontaneous menopause. The support treatment should be given. ERT should not be given. |
| Center2 | Metabolism is not stabilized. ERT should be given. |
| Center3 | The patient is in the transition phase. ERT should be decided after 9-12 months. |
| Center4 | ERT can be progressed. |
| Center5 | Medical assessment should be required. The support treatment should be given. |

If the data members are assigned to the any of the clusters defined above, the treatment of the cluster should be applied to the patient also. The decision could be given according to this condition.

### 6.2 Multiple Centered Clustering (MCFC)

MCFC, which was the method described briefly in Chapter 5, is used to find the clusters and their cluster centers using the menopause data set.

Algorithm is started using the fuzzy ant initialization in order to find the best initial cluster centers so as to decrease the sensitivity of the FCM algorithm. After finding the best initial cluster centers from fuzzy ant initialization process, FCM part could start. There were sixteen clusters defined for FCM process initially.

For MCFC process, the final number of clusters was selected as an output of the algorithm. Because of the medical meaning of five clusters defined in chapter 6.1, five final clusters were selected for the algorithm. According to the steps defined in chapter 5, the results which are presented in the tables below, were found by MCFC algorithm.

Table 21: Results of MCFC-1st cluster

| | Age | Weight | Height | Mens. Duration | FSH | LH | Estradiol |
|---|---|---|---|---|---|---|---|
| Center1 | 49.8320 | 62.4486 | 160.8014 | 40.7473 | 48.5256 | 23.9321 | 113.8151 |
| Center2 | 48.6767 | 59.5607 | 158.9037 | 34.3864 | 52.5264 | 22.7226 | 43.9949 |
| Center3 | 50.6996 | 63.1603 | 157.7126 | 38.7098 | 98.5254 | 45.5608 | 63.2934 |
| Center4 | 54.2820 | 65.7326 | 160.2086 | 104.3357 | 56.7857 | 26.013 | 87.6512 |

| | T3 | T4 | TSH | Glucose | Cholesterol | Triglyceride | HDL |
|---|---|---|---|---|---|---|---|
| Center1 | 0.9638 | 8.8829 | 2.5063 | 92.6080 | 211.6226 | 101.3134 | 73.0027 |
| Center2 | 0.5931 | 4.9043 | 2.4268 | 58.3283 | 133.7203 | 64.4953 | 39.0830 |
| Center3 | 0.9399 | 7.8558 | 2.1847 | 89.9343 | 202.0765 | 100.3058 | 60.2226 |
| Center4 | 1.0309 | 9.9880 | 2.7782 | 90.2676 | 248.4374 | 130.9413 | 70.2954 |

| | LDL | VLDL | Hemoglobin | Hematocrit | Men. Type | HRT | HRT Duration |
|---|---|---|---|---|---|---|---|
| Center1 | 117.1002 | 21.5754 | 13.1731 | 38.8314 | 2.6257 | 1.0042 | 33.2116 |
| Center2 | 87.4521 | 14.0813 | 8.5732 | 25.1642 | 1.4484 | 1.0037 | 14.0733 |
| Center3 | 123.1116 | 20.2592 | 13.8507 | 40.5461 | 1.2159 | 1.0036 | 21.2441 |
| Center4 | 149.5605 | 27.7897 | 12.9955 | 38.4430 | 1.2879 | 1.0033 | 62.7697 |

Table 22: Results of MCFC-2nd cluster

| | Age | Weight | Height | Mens. Duration | FSH | LH | Estradiol |
|---|---|---|---|---|---|---|---|
| Center1 | 49.3413 | 64.5195 | 156.8573 | 38.5193 | 176.8233 | 60.8639 | 31.4490 |
| Center2 | 46.3830 | 60.8883 | 161.1947 | 23.3060 | 117.8063 | 47.7256 | 47.5728 |
| Center3 | 48.7893 | 66.6091 | 159.0758 | 24.4442 | 43.5638 | 24.1038 | 93.1084 |
| Center4 | 52.4321 | 63.9444 | 159.3775 | 65.9194 | 141.5257 | 56.3916 | 39.7124 |
| Center5 | 53.2474 | 80.0463 | 166.9530 | 43.2647 | 69.7371 | 28.3457 | 47.0656 |
| Center6 | 50.2762 | 65.7022 | 157.3163 | 41.8003 | 93.9655 | 36.3508 | 48.4192 |
| Center7 | 52.8960 | 64.7000 | 159.5003 | 45.6722 | 96.2918 | 39.6161 | 45.5916 |

| | T3 | T4 | TSH | Glucose | Cholesterol | Triglyceride | HDL |
|---|---|---|---|---|---|---|---|
| Center1 | 0.9541 | 9.1523 | 2.7446 | 96.5151 | 203.8642 | 77.5820 | 74.7589 |
| Center2 | 0.9404 | 6.6733 | 2.4780 | 81.7936 | 194.0663 | 70.6019 | 64.0972 |
| Center3 | 1.0624 | 8.1294 | 2.6637 | 96.6536 | 201.5161 | 91.1969 | 50.1722 |
| Center4 | 1.1334 | 9.8039 | 3.0331 | 83.0533 | 230.3238 | 106.5845 | 63.4189 |
| Center5 | 0.8743 | 6.5822 | 3.3617 | 94.8091 | 225.4344 | 105.3762 | 58.7171 |
| Center6 | 0.9888 | 8.3468 | 2.3540 | 87.5155 | 271.5685 | 102.1203 | 67.1275 |
| Center7 | 1.0582 | 8.2988 | 2.4828 | 90.7424 | 221.8907 | 75.8353 | 81.8569 |

| | LDL | VLDL | Hemoglobin | Hematocrite | Men. Type | HRT | HRT Duratn. |
|---|---|---|---|---|---|---|---|
| Center1 | 113.3241 | 16.0622 | 13.3541 | 39.6126 | 2.6588 | 1.9583 | 1.5235 |
| Center2 | 114.3595 | 15.1845 | 13.7153 | 40.1390 | 1.0568 | 1.9899 | 0.4424 |
| Center3 | 129.2225 | 22.3191 | 13.9047 | 40.6719 | 1.1141 | 1.9877 | 0.5382 |
| Center4 | 144.8165 | 21.6994 | 13.5396 | 39.4052 | 1.2551 | 1.9834 | 0.7725 |
| Center5 | 142.6656 | 23.5655 | 13.1081 | 38.4530 | 1.0455 | 1.9872 | 0.5865 |
| Center6 | 185.7190 | 20.8908 | 13.6893 | 39.6136 | 1.0603 | 1.9899 | 0.4874 |
| Center7 | 123.8053 | 15.6280 | 13.0816 | 39.2221 | 1.1018 | 1.9887 | 0.5158 |

Table 23: Results of MCFC-3rd Cluster

| | Age | Weight | Height | Mens. Duration | FSH | LH | Estradiol |
|---|---|---|---|---|---|---|---|
| Center1 | 46.7659 | 62.6700 | 162.7698 | 45.4652 | 51.2465 | 21.1124 | 42.0676 |
| Center2 | 47.6238 | 59.0242 | 158.0277 | 44.6216 | 55.7260 | 25.6254 | 80.5563 |

| | T3 | T4 | TSH | Glucose | Cholesterol | Triglyceride | HDL |
|---|---|---|---|---|---|---|---|
| Center1 | 0.5689 | 4.6799 | 3.3446 | 58.0327 | 129.2508 | 62.9580 | 39.4788 |
| Center2 | 1.0404 | 9.2851 | 7.7468 | 95.8845 | 194.4213 | 102.6322 | 55.8865 |

| | LDL | VLDL | Hemoglobin | Hematocrite | Men. Type | HRT | HRT Duration |
|---|---|---|---|---|---|---|---|
| Center1 | 81.6143 | 13.4431 | 8.1384 | 23.8746 | 3.3719 | 1.9956 | 0.1698 |
| Center2 | 120.4233 | 20.7875 | 13.1368 | 39.3563 | 3.3450 | 1.9912 | 0.4594 |

Table 24: Results of MCFC-4th cluster

| | Age | Weight | Height | Mens. Duration | FSH | LH | Estradiol |
|---|---|---|---|---|---|---|---|
| Center1 | 67.0856 | 67.5672 | 160.4557 | 194.3475 | 52.6146 | 22.4299 | 42.3389 |
| Center2 | 50.0379 | 63.2425 | 161.3454 | 37.8576 | 53.3344 | 22.2073 | 42.5224 |

| | T3 | T4 | TSH | Glucose | Cholesterol | Triglyceride | HDL |
|---|---|---|---|---|---|---|---|
| Center1 | 0.5695 | 4.6830 | 2.3611 | 57.2360 | 141.3032 | 66.4465 | 43.2696 |
| Center2 | 0.5672 | 4.6838 | 2.2254 | 55.5639 | 133.1004 | 61.6784 | 39.1679 |

| | LDL | VLDL | Hemoglobin | Hematocrite | Men. Type | HRT | HRT Duration |
|---|---|---|---|---|---|---|---|
| Center1 | 81.2758 | 14.2996 | 8.2094 | 24.0175 | 1.0372 | 1.9788 | 0.8598 |
| Center2 | 80.5248 | 13.3252 | 8.0820 | 23.7100 | 1.0390 | 1.9962 | 0.1485 |

Table 25: Results of MCFC-5th cluster

| | Age | Weight | Height | Mens. Duration | FSH | LH | Estradiol |
|---|---|---|---|---|---|---|---|
| Center1 | 59.8786 | 64.2288 | 161.1444 | 143.3209 | 51.4776 | 21.9075 | 43.4352 |

| | T3 | T4 | TSH | Glucose | Cholesterol | Triglyceride | HDL |
|---|---|---|---|---|---|---|---|
| Center1 | 0.5727 | 4.7804 | 2.5214 | 56.3900 | 136.4126 | 61.8369 | 39.4854 |

| | LDL | VLDL | Hemoglobin | Hematocrite | Men. Type | HRT | HRT Duration |
|---|---|---|---|---|---|---|---|
| Center1 | 83.3929 | 13.8009 | 8.1543 | 23.9503 | 1.4876 | 1.0018 | 81.0239 |

According to the tables shown above, it can be easily understood that nearly all clusters have more than one cluster center. This condition allows locating similar and dissimilar data members in the same group. For the medical data set, the cluster centers show different treatments for the patients in the same cluster. The following table shows the medical meanings of the cluster centers which were discussed with the doctors and instructors of Hacettepe University.

Table 26: Meanings of cluster centers found by MCFC method

| 1st Cluster | |
|---|---|
| Center1 | Estrogen Replacement Therapy (ERT) can be progressed. |
| Center2 | ERT may not be appropriate. Medical assessment should be required. |
| Center3 | Estrogen Replacement Therapy (ERT) can be progressed. |
| Center4 | Estrogen Replacement Therapy (ERT) can be progressed. (Note: Duration of the ERT should be considered) |
| 2nd Cluster | |
| Center1 | The patient may be benefited from ERT. |
| Center2 | ERT may be appropriate. |
| Center3 | If the doctor desires, ERT can be given to the patient. |
| Center4 | The patient may be benefited from ERT. |
| Center5 | ERT may be appropriate. |
| Center6 | The patient may be benefited from ERT. |
| Center7 | ERT may be appropriate. |
| 3rd Cluster | |
| Center1 | ERT may not be appropriate. Medical assessment should be required. |
| Center2 | ERT may not be appropriate. |
| 4th Cluster | |
| Center1 | ERT is not suitable for the patient. |
| Center2 | ERT may not be appropriate. Medical assessment should be required. |
| 5th Cluster | |
| Center1 | ERT should be aborted. Medical assessment should be required. |

In high dimensional data sets, it is always difficult to visualize the behavior of the clusters and cluster centers. To see the behavior, three most important components, HRT, HRT duration and menopause type, are taken from the original data set and these three components create a new data set. For this data set, the classical clustering algorithms and the new proposed approach, MCFC, are applied on it. The results of the clustering in three dimensional menopause data set are given in the following figures.

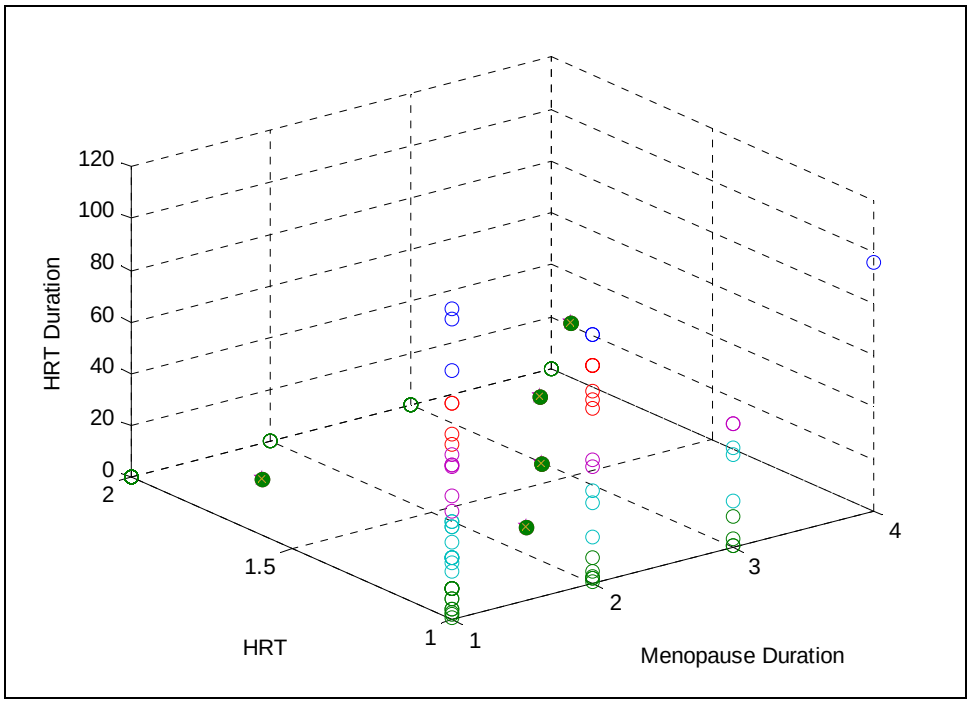


Figure 41: FCM results of three dimensional menopause data set

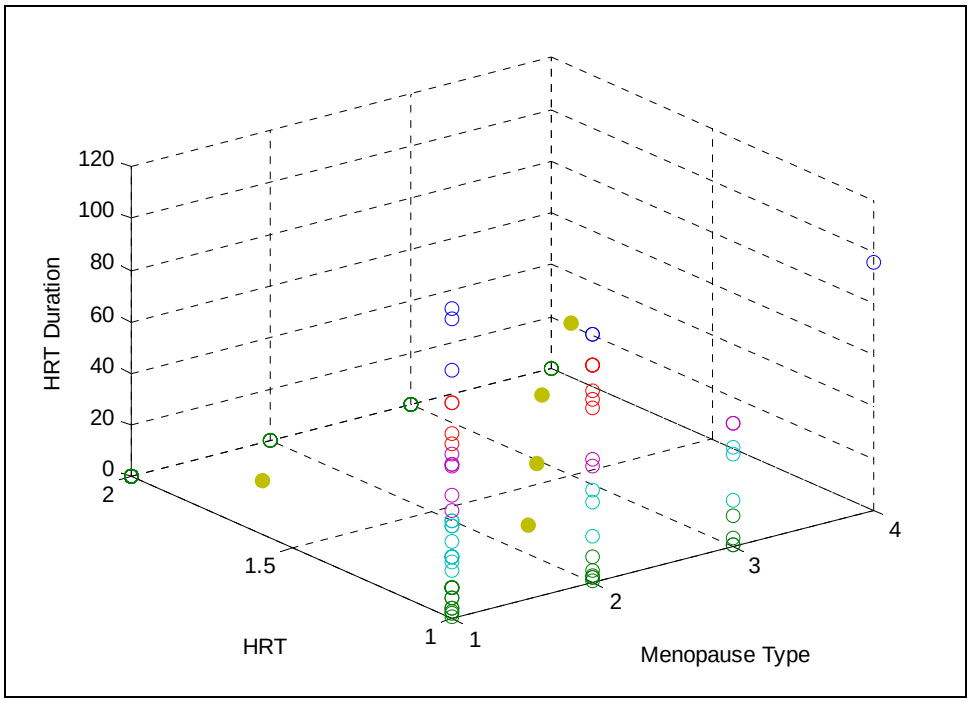


Figure 42: HCM results of three dimensional menopause data set

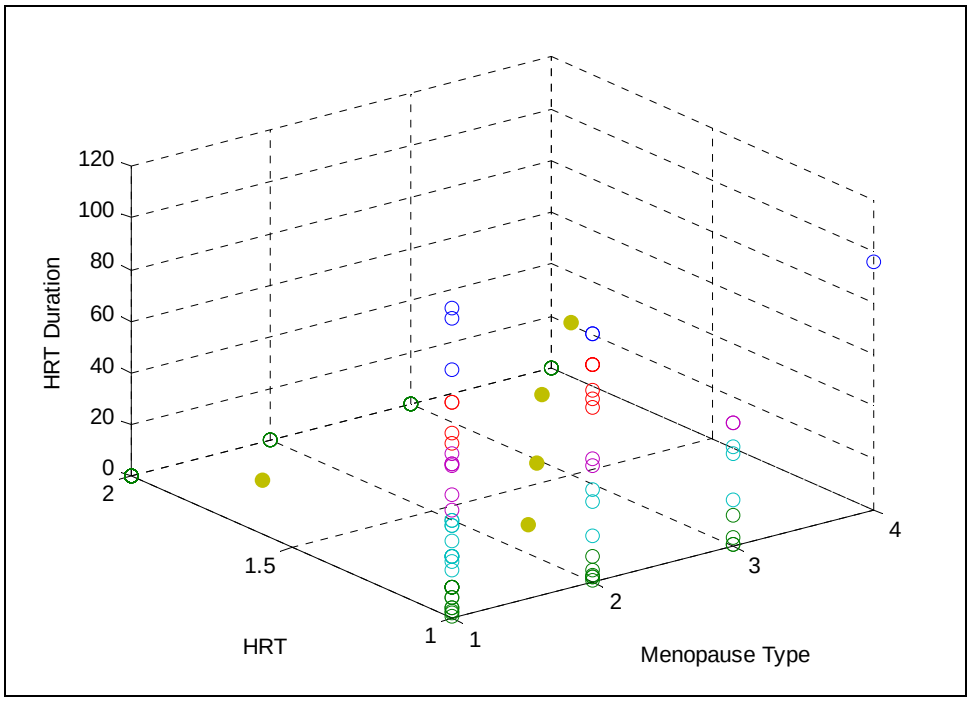


Figure 43: KM results of three dimensional menopause data set

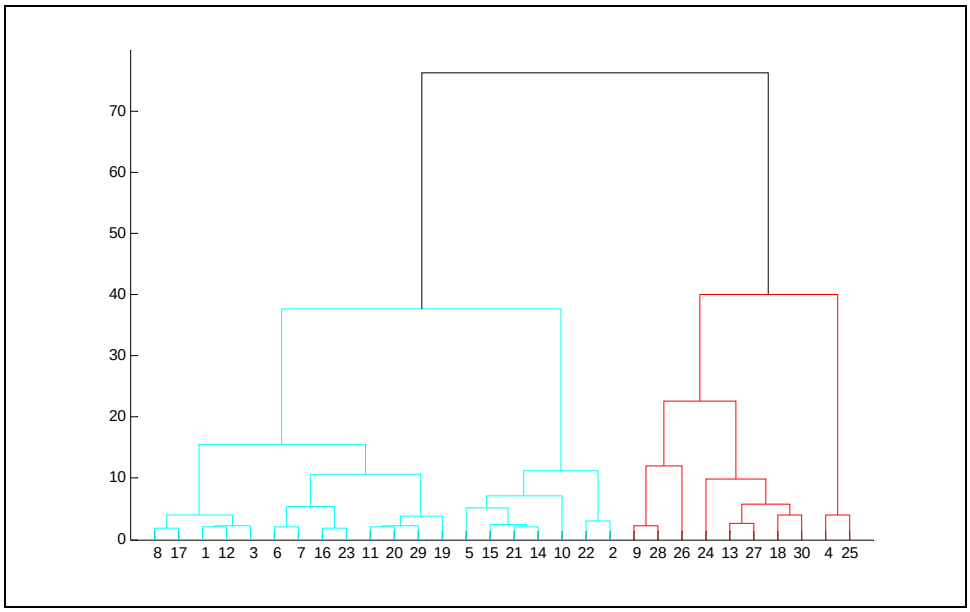


Figure 44: Hierarchical Clustering Tree of SBC

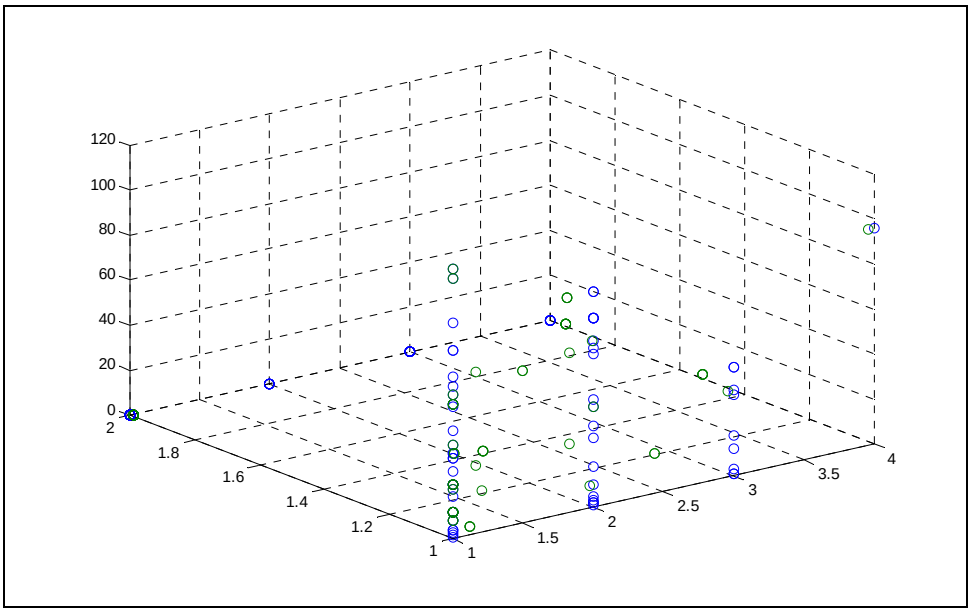


Figure 45: SBC results of three dimensional menopause data set

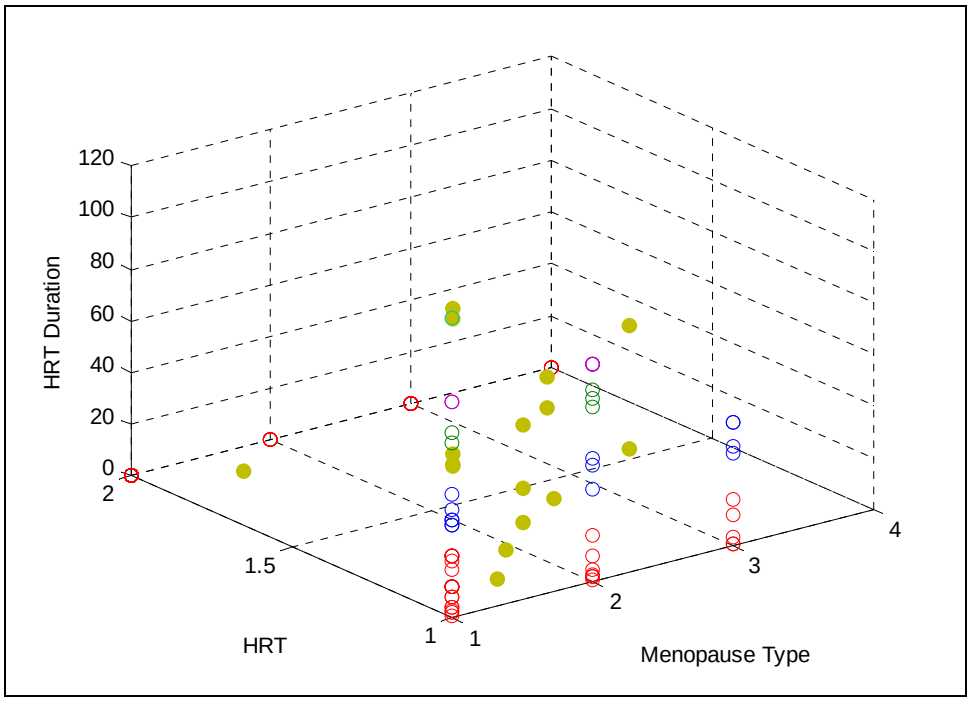


Figure 46: MCFC results of three dimensional menopause data set

When the results in figure 41 – 46, it can be easily understood that the well known clustering algorithms defined in chapter two can find the clusters, that are taken estrogen hormone or not. The clusters are assigned to only one result: Taking ERT or not. The results of MCFC in figure 46 show that the different treatments should be assigned to the same cluster by using multiple-centered tasking. In MCFC in figure 46, initial number of cluster centers was selected as sixteen and the final number of clusters was selected as five. In one of the clusters, the HRT value is assigned as 1 and 1.9 for the cluster centers. The meaning of ERT equals to 1 is "The patient should take ERT" and the meaning of ERT equals to 1.9 is "The patient should not take ERT". Therefore, MCFC can find different sub-clusters located in the same cluster.

According to the medical and mathematical results defined in the tables above, the decision making system was designed. This system consists of two parts. In the first part, there is a questionnaire section for the user. In this part, the user starts using the program by filling the general personal information about the patient. This information is name, surname, age, blood group, dossier number of the patient. After filling this part, the doctor should fill the obstetrics part. In this part, the obstetrics information of the patient is needed by the program. These are gravida, para, abortus/fetal loss, number of living children.

After filling the information part, the medical questionnaire part should be filled. If one of the answers of the defined question is "YES", the 2nd stage of the program is not appeared. For this patient, Estrogen Replacement Therapy is not suitable. Therefore, the program does not continue. In this condition, the program shows warning on the screen "ERT is not suitable for the patient!". The first stage of the program can be seen in the following figures.

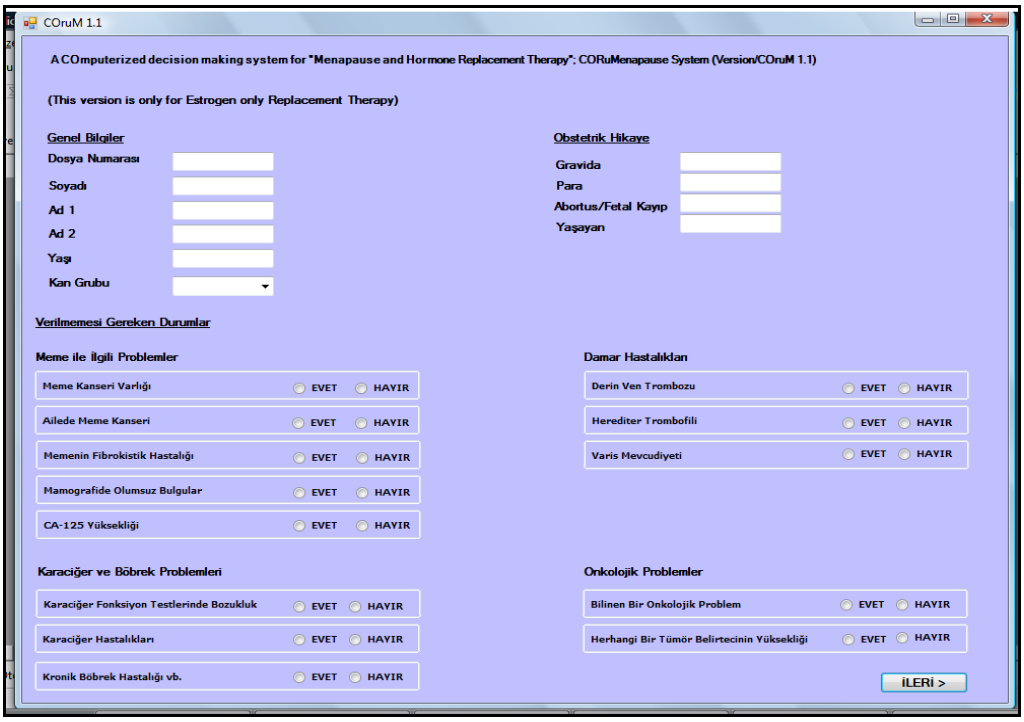


Figure 47: First stage of the program

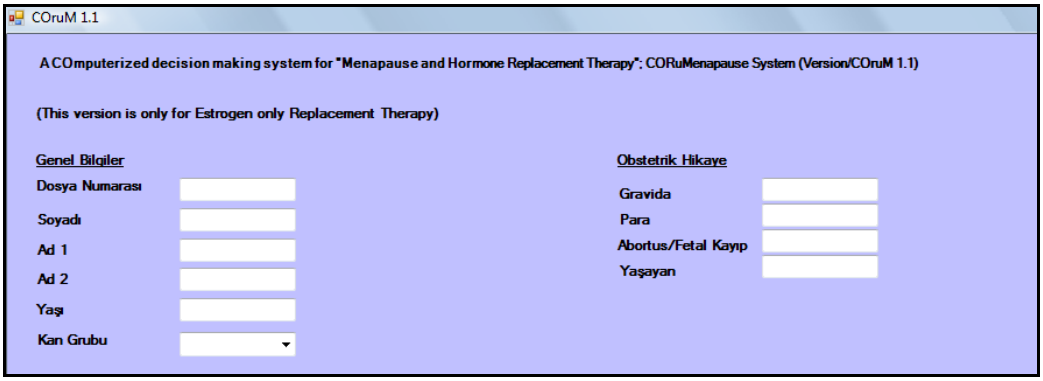


Figure 48: General information part of the first stage

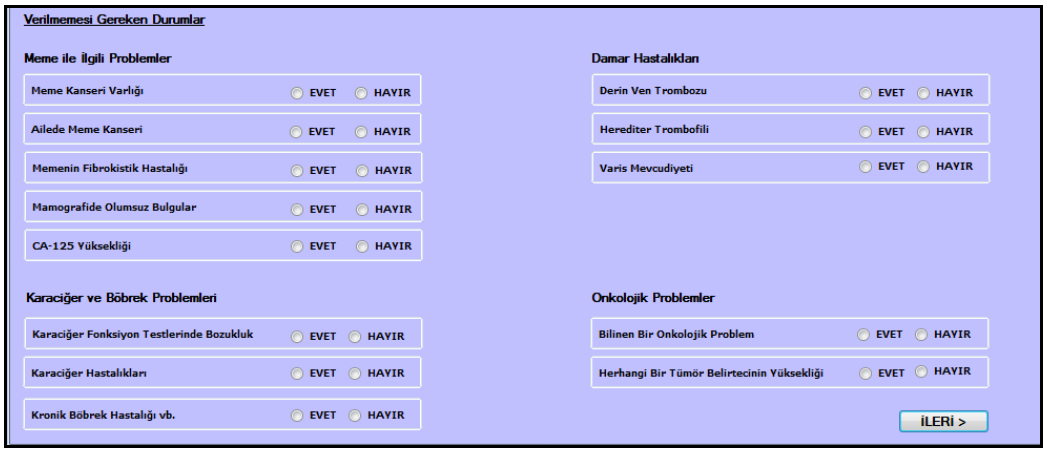


Figure 49: The questionnaire part of the first stage

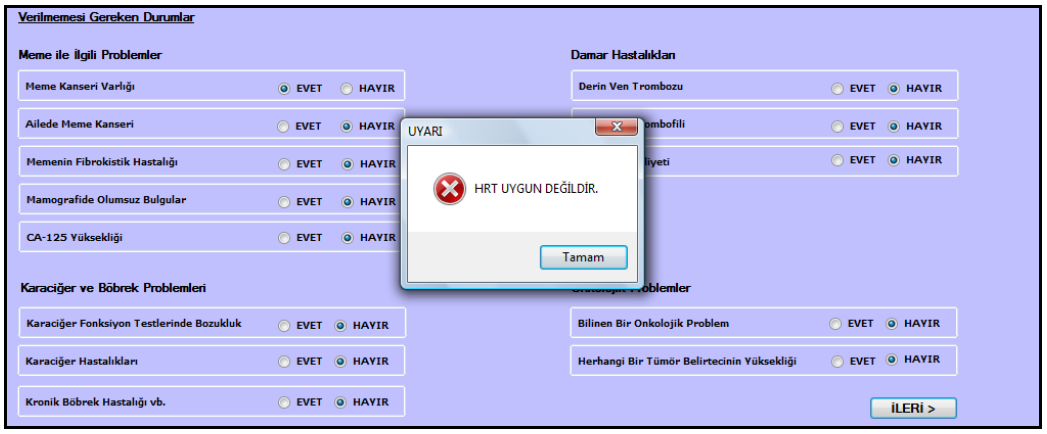


Figure 50: An example of the filled questionnaire stage

After completing the first stage by using the rules defined below, the second stage continues. For processing the second stage, the laboratory test results of the patients is required by the program. For processing, the second stage of the program is designed to fill the laboratory results of the patient. The design of the second stage can be seen in figure 51.

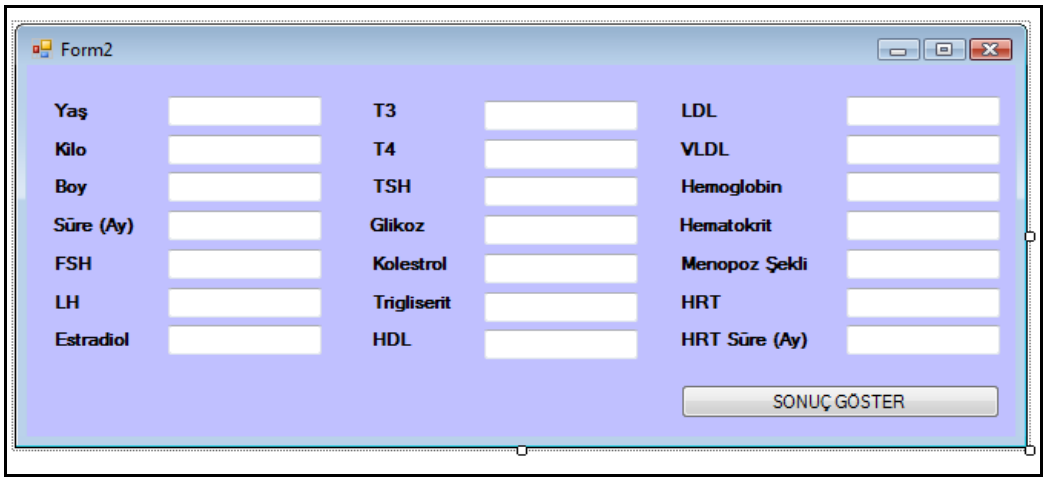


Figure 51: The design of the second stage

At the end of the second stage, the program will make a decision for the patient. The decision depends on the clustering results of MCFC defined in table 14.

The system decides a result and gives a recommendation as a result of using Euclidean distance measure. Each cluster center defined in table 26 is accepted as a row vector. Each cluster center is indicated as $c_i$ where i = 1,2,...,c. When the doctor fills the form 2, the information from the form 2 creates a data vector $x_k$. After creating $x_k$, the Euclidean distance between $x_k$ and cluster centers are calculated by using the equation below:

$$d(x_k - c_i) = \sqrt{\sum_{j=1}^{n} ((x_k)_j - (c_i)_j)^2} \tag{6.1}$$

Euclidean distance measure should be calculated for all the clusters. The patient should be assigned to the cluster center which the Euclidean distance measure is minimum with this center. According to the medical meanings of the cluster centers defined in table 14, the treatment of the patient should be the same treatment of the cluster that it will be assigned to.

# CHAPTER 7

# CONCLUSIONS

## 7.1 Results

The aim of this thesis study is to develop a decision making system algorithm on menopause data set. The data set members were collected from the laboratory test results of the women in Hacettepe University, Faculty of Medicine, Department of Gynaecology by Prof. M. Sinan BEKSAÇ.

This study consists of mainly two stages. The first stage is the MCFC clustering part, and the second stage is the development of the decision making system. In the clustering part, many well-known clustering methods were applied on the data set. According to the theoretical background of these methods and applications of these methods on the artificial data set, the MCFC method was proposed. When we compare MCFC and other well-known clustering algorithms, it could be understood that the proposed multiple-centered clustering method gives much more reasonable results than the other clustering methods. On the other hand, because of the merging process between clusters and total number of cluster centers, it takes too much time to converge than the other well known clustering methods. As a point in favor of MCFC, the presented clustering approach seems clearly advantageous for the data sets where dissimilar data groups are located in the same clusters. The outputs of MCFC are the most reasonable results when compared with the well-known and most frequently used clustering algorithms defined in chapter two. When we compare the proposed MCFC and density based clustering algorithms, it can be easily seen that shape of clusters found by MCFC algorithm are similar to clusters found by density-based clustering algorithm. Although shape of clusters is similar, the proposed

MCFC algorithm can find multiple cluster centers in the same cluster. Density based clustering algorithm finds only one cluster center for all clusters. Because of using single-centered clusters, density based clustering algorithm can not find the sub-clusters located in the same cluster. The main difference between MCFC and density based clustering algorithm is to find the different sub-clusters located in the same cluster by using multiple-centered clustering method. When the hierarchical clustering algorithms and MCFC are compared, it can be understood that there is no approach called as multiple-centered in hierarchical clustering. In general for hierarchical clustering, the trees, or dendrograms, are defined for hierarchical clustering. The leaves in the tree correspond to the individual observations. At the top of tree, which is called root, is a single cluster which includes all the observations in the leaves. Therefore, in hierarchical clustering, there is no approach called as a single cluster includes multiple cluster centers. So, the algorithm is different from the MCFC algorithm.

At the second stage, the decision making systems was developed to recommend the doctors to make a decision about the women that needs ERT or not. Sixteen cluster centers for five clusters were used for the clustering processes. All clusters have different meanings from each other in general. For cluster centers, the medical meanings of all clusters were defined by the doctors of Hacettepe University, Faculty of Medicine, and Gynecology Department.

After defining the meaning of the clusters, the decision making system was designed. The aim of the system is to find the new patient's group which she will be assigned to and give the treatment of this group to her. The cluster of the patient is found by using the Euclidean distance measure between the patient's test results and cluster centers. All distances are calculated by using the Euclidean distance. Then, the cluster, which is the closest to the patient's test results, is the cluster that she will be assigned to. After finding the cluster she assigned, the treatment of this cluster can be applied to the patient.

According to the stages defined above, the new system was designed to give a recommendation to the doctors. The name of the system, which will be used, is "A COmputerized decision making system for "Menopause and Hormone Replacement Therapy"; COruMenapause System (Version/COruM 1.1)". This program was created by using Microsoft Visual Basic programming language. Because of its successful visual packages, this program is used for creating a human-machine interface. The COruM version 1.1 is only for Estrogen Replacement Therapy. This system is very useful for treatment recommendation.

This program includes two parts. In the first part, the user (doctor) fills the questionnaire section at the first page. According to the answers of the questionnaire, the second part can start. In the second part, the new page will open and the doctor will be imputing the laboratory test results of the woman. After imputing, the program gives a treatment recommendation to the doctor. As a result, the recommendation system for the doctors was designed and it can be used in the hospitals after visual arrangements.

## 7.2 Recommendations

As an extension of this thesis study, the future work will be recommended as:

- The other programming languages, which are much faster than the Matlab R2007b can be used for decreasing the time complexity of the algorithm. Simulations in Matlab take too much time during processing. For example, in Chapter 4, the simulation of SBC for data set 2 took 40 hours and the simulation of MCFC with 250 clusters for data set 2 using fuzzy ants took 56 hours during whole process.

- The total number of cluster centers can be selected as total members of data set to increase the robustness of the program. The MCFC simulations of data set 2 with 250 clusters took 56 hours. So, the number of clusters which is equal to the number of members in the data set could not be tried. Therefore, changing of the programming language shows the robustness results in small time duration.

- The concept of similarity of sets should be further investigated. There are some other plausible definitions of similarity.

## 7.3 Papers

As a result of this study, three papers were accepted in national and international conferences. These are listed below:

- H.Ö. Bacak and K. Leblebicioğlu, *"Fuzzy Clustering Based Multiple-Centered Clustering Method"*, IEEE The Ninth International Conference on Machine Learning and Applications , Washington DC, USA, 2010.

- H.Ö. Bacak and K. Leblebicioğlu, *"Bulanık Kümeleme Yöntemi Tabanlı Çok Merkezli Kümeleme Yöntemi"*, Bilimde Modern Yöntemler Sempozyumu 2010, Diyarbakır, Turkey, 2010.

- H.Ö. Bacak, K. Leblebicioğlu and S. Beksaç, *"Menopoz Datalarıyla İlgili Karar Verebilen Bir Sistem Geliştirilmesi"*, Bilimde Modern Yöntemler Sempozyumu 2010, Diyarbakır, Turkey, 2010.

## REFERENCES


[1] P.M. Kanade and L.O. Hall, "*Fuzzy Ants and Clustering"*, IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, vol. 37, no. 5, pp. 758 – 769, 2007.

[2] C. Bishop, *Pattern Recognition and Machine Learning*, Singapore: Springer, 2006.

[3] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification*, Wiley, 2006.

[4] M. Sarkar and T. Leong, *Fuzzy K-Means Clustering with Missing Values*, Department of Computer Science, School of Computing National University of Singapore, 2001.

[5] A. Likas, N. Vlassis and J. Verbeek, "*The Global K-means Clustering Algorithm"*, Pattern Recognition Letters, vol. 36, pp. 451 – 461, 2002.

[6] A.K. Jain, *Data Clustering: 50 Years Beyond K-means*, Journal of the Pattern Recognition Society, Pattern Recognition Letters vol. 31, pp 651–666, 2010.

[7] H. Park and C. Jun, *A Simple and Fast Algorithm for K-medoids Clustering*, Expert Systems with Applications, vol. 36, pp. 3336 – 3341, 2009.

[8] L. Tari, C. Baral and S. Kim, *Fuzzy c-means Clustering with Prior Biological Knowledge*, Journal of Biomedical Informatics, vol. 42, pp. 74 – 81, 2009.

[9] A. Flores-Sintas, J.M. Cadenas and F. Martin, *Membership Functions in the Fuzzy c-means Algorithm*, Fuzzy Sets and Systems, vol. 101, pp. 49 - 58, 1999.

[10] J. Fan, W. Zhen and W. Xie, *Suppressed Fuzzy c-means Clustering Algorithm*, Pattern Recognition Letters, vol. 24, pp. 1607 – 1612, 2003.

[11] A. Baraldi and P. Blonda, *A survey of fuzzy clustering algorithms for pattern recognition*, IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 29, no. 6, pp. 778 – 785, 1999.

[12] J.B. MacQueen, *Some Methods for classification and Analysis of Multivariate Observations*", 1. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press. pp. 281 – 297, 1967.

[13] S.P. Lloyd, *Least squares quantization in PCM*, IEEE Transactions on Information Theory, vol. 28, no. 2, 1982.

[14] K.R. Zalik, *An efficient k-means clustering algorithm*, Pattern Recognition Letters, vol. 29, pp. 1385 – 1391, 2007.

[15] R.J. Kuo, H.S. Wang, Tung-Lai Hu and S.H. Chou, *Application of Ant K-Means on Clustering Analysis*, An International Journal of Computer & Mathematics with Applications, vol. 50, pp. 1709 - 1724, 2005.

[16] S. Das, A. Abraham and A. Konar, *Automatic kernel clustering with a Multi-Elitist Particle Swarm Optimization Algorithm*, Pattern Recognition Letters, vol. 29, pp. 688 – 699, 2007.

[17] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman, *Missing Value Estimation Methods for DNA Micro arrays*, Bioinformatics, vol. 17, no 6, pp. 520 - 525, 2001.

[18] A. Ghanbari, S.F. Ghaderi, M. A. Azadeh*, A Clustering based Genetic Fuzzy Expert System for Electrical Energy Demand Prediction*, The 2nd International Conference on Computer and Automation Engineering, pp. 407 - 411, 2010.

[19] J. Kobayashi, H. Asaka, H. Mitsui, M. Sone, I. Terayama and Y. Kenmotsu, *Expert System with Fuzzy Clustering Method for Diagnosis on Life of Transformers,* Conference on Electrical Insulation and Dielectric Phenomena, pp. 409 - 414, 1992.

[20] Ghahramani, Z. and M. I. Jordan, *Learning from incomplete data*, Technical report, AI memo no. 1509, MIT, 1994.

[21] D.F. Heitjan, *Annotation: What can be done about missing data? Approaches to imputation*, American Journal of Public Health, vol. 87, no. 4, pp. 548 – 550, 1997.

[22] R.J.A Little and D.B. Rubin, *Statistical Analysis with Missing Data*, Wiley, New York, 1987.

[23] T. Schneider, *Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values*, Journal of Climate, vol. 14, no. 5, pp. 853 – 871, 2001.

[24] A.K. Jain and R.C. Dubes, "*Algorithms for Clustering Data*", Englewood Cliffs, NJ: Prentice-Hall, 1988.

[25] M.S. Yang and K.L Wu, *A Similarity Based Robust Clustering Method*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 4, pp. 434 – 448, 2004.

[26] http://en.wikipedia.org/wiki/K-means_clustering, September 2010.

[27] http://en.wikipedia.org/wiki/Fuzzy_clustering, September 2010.

## APPENDIX A

## MEDICAL DATA SET BY HACETTEPE UNIVERSITY, FACULTY OF MEDICINE, DEPARTMENT OF GYNAECOLOGY

(179 Rows-number of people, 21 columns-personal information and laboratory test results of people)

| No | Age | Weight | Height | Mens. Duration | FSH | LH | Estradiol |
|---|---|---|---|---|---|---|---|
| 1 | 50 | 66 | 160 | 5 | missing | missing | missing |
| 2 | 54 | 67 | 162 | 56 | 86 | 30.3 | 20 |
| 3 | 46 | 64 | 157 | missing | 77.6 | 35.5 | 135.2 |
| 4 | 48 | 65 | 158 | 1 | 164.6 | 37.9 | 20.7 |
| 5 | 50 | 57 | 150 | 41 | 162.6 | 49.5 | 20 |
| 6 | 51 | 63 | 158 | 24 | 78.1 | 44.9 | 121.5 |
| 7 | 52 | 59 | 158 | 12 | 40.4 | 29.6 | 140.3 |
| 8 | 46 | 58 | 162 | 2 | 200 | 81.9 | 20 |
| 9 | 54 | 65 | 165 | 120 | 65 | 36.7 | 55.2 |
| 10 | 42 | 62 | 153 | 4 | 103 | 40.8 | missing |
| 11 | 51 | 65 | 166 | 41 | 65.7 | 35.9 | missing |
| 12 | 39 | 65 | 162 | 2 | 74.7 | 44.5 | 47.3 |
| 13 | 49 | 63 | 150 | 5 | 25.3 | 46 | 334.5 |
| 14 | 51 | 52 | 157 | missing | 11.2 | 5.1 | 122.5 |
| 15 | 49 | 55 | 157 | 132 | missing | missing | missing |
| 16 | 51 | 54 | 163 | 12 | 113 | 47.9 | missing |
| 17 | 52 | 60 | 167 | 31 | 125 | 74 | 5 |
| 18 | 38 | 58 | 164 | 32 | 117.9 | 57.8 | 20 |
| 19 | 50 | 65 | 150 | 14 | 183.3 | 66.2 | 20.5 |
| 20 | 51 | 69 | 165 | 6 | 97.4 | 59.7 | 20.7 |
| 21 | 54 | 64 | 156 | 264 | 62.3 | 33.9 | 20 |
| 22 | 61 | 66 | 162 | 96 | missing | missing | missing |
| 23 | 47 | 55 | 160 | 2 | 37.9 | 23.9 | 66 |
| 24 | 55 | 60 | 157 | 84 | 11.9 | 16.9 | 185.5 |
| 25 | 45 | 60 | 157 | 12 | 68.6 | 41.5 | 99.1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 26 | 56 | 68 | 150 | 192 | missing | missing | missing |
| 27 | 54.07 | 67 | 152 | 156 | 83.1 | 18.8 | 20 |
| 28 | 49 | 65 | 154 | 24 | 89.2 | 24.2 | 20 |
| 29 | 50 | 65 | 160 | 13 | 84.7 | 37.3 | 79 |
| 30 | 43 | 75 | 174 | 5 | 98.4 | 24.4 | 24 |
| 31 | 53 | 68 | 158 | 168 | missing | missing | missing |
| 32 | 46 | 58 | 158 | 41 | missing | missing | missing |
| 33 | 61 | 54 | 155 | 46 | 68.2 | 26.8 | 129.2 |
| 34 | 69 | 58 | 163 | 240 | missing | missing | missing |
| 35 | 62 | 50 | 153 | 216 | missing | missing | missing |
| 36 | 55 | 64 | 153 | 72 | missing | missing | missing |
| 37 | 54 | 75 | 162 | 10 | 147.9 | 47.3 | 157.2 |
| 38 | 41 | 69 | 163 | 65 | missing | missing | missing |
| 39 | 57 | 57 | 160 | 72 | 98.2 | 43.1 | 20 |
| 40 | 50 | 60 | 168 | 132 | missing | missing | missing |
| 41 | 48 | 70 | 164 | missing | missing | missing | missing |
| 42 | 47 | 56 | 158 | 4 | 110 | 52 | 26 |
| 43 | 48 | 59 | 160 | 36 | 60.1 | 24.6 | 56.4 |
| 44 | 60 | 71 | 155 | 72 | 41.9 | 26.1 | 23.5 |
| 45 | 50 | 70 | 160 | 72 | 218.2 | 91 | 758.6 |
| 46 | 43 | 67 | 160 | 2 | missing | missing | missing |
| 47 | 44 | 58 | 150 | 5 | missing | missing | missing |
| 48 | 67 | 71 | 166 | 180 | missing | missing | missing |
| 49 | 51 | 63 | 158 | 24 | 78.1 | 44.9 | 121.5 |
| 50 | 51 | 73 | 158 | 12 | missing | missing | missing |
| 51 | 54 | 98 | 164 | 3 | missing | missing | missing |
| 52 | 31 | 53 | 165 | missing | missing | missing | missing |
| 53 | 57 | 65 | 160 | 108 | missing | missing | missing |
| 54 | 56 | 65 | 156 | 17 | 79.2 | 24.4 | 130.2 |
| 55 | 60 | 72 | 160 | 180 | missing | missing | missing |
| 56 | 49 | 60 | 160 | 3 | 113.4 | 34.6 | 20 |
| 57 | 50 | 56 | 155 | missing | 53.3 | 34.34 | 20 |
| 58 | 58 | 70 | 163 | 36 | 91.4 | 25.1 | 20 |
| 59 | 52 | 65 | 158 | 10 | 15.4 | 8 | 506.3 |
| 60 | 54 | 67 | 165 | 60 | 126.9 | 65.3 | 20 |
| 61 | 53 | 57 | 163 | missing | 67.1 | 38 | 91.9 |
| 62 | 42 | 51 | 168 | missing | missing | missing | missing |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 63 | 54 | 75 | 164 | 84 | 106.2 | 61.6 | 31.8 |
| 64 | 63 | 68 | 161 | 156 | 69 | 23.2 | 20 |
| 65 | 47 | 65 | 163 | 8 | 42 | 23.8 | 27.8 |
| 66 | 38 | 59 | 162 | 6 | 43.9 | 32.4 | 20 |
| 67 | 35 | 65 | 169 | missing | 85.4 | 26.7 | 108.2 |
| 68 | 53 | 57 | 164 | missing | 1.2 | 0.2 | 70 |
| 69 | 57 | 65 | 158 | 48 | missing | missing | missing |
| 70 | 48.8 | 66 | 163 | 84 | missing | missing | missing |
| 71 | 52 | 50 | 150 | 30 | 121.1 | 50.9 | 20 |
| 72 | 55 | 92 | 155 | 156 | 177.5 | 53 | 23.2 |
| 73 | 47 | 76 | 164 | 18 | 29.9 | 14.5 | 200.1 |
| 74 | 49 | 75 | 167 | 36 | missing | missing | missing |
| 75 | 46 | 65 | 171 | 14 | 121.4 | 28.8 | 20 |
| 76 | 53 | 54 | 163 | missing | 55.9 | 31 | 144.6 |
| 77 | 64 | 72 | 168 | 180 | missing | missing | missing |
| 78 | 51 | 60 | 156 | missing | 192.2 | 52.6 | 20 |
| 79 | 51 | 74 | 157 | 72 | 81.8 | 39.2 | 24.9 |
| 80 | 49 | 53 | 156 | 12 | 85.4 | 32.7 | 20 |
| 81 | 50 | 59 | 162 | 3 | 80.1 | 39.4 | 38.6 |
| 82 | 50 | 49 | 160 | missing | 43.4 | 21.7 | 127.7 |
| 83 | 59 | 64 | 160 | 84 | missing | missing | missing |
| 84 | 48 | 83 | 156 | missing | missing | missing | missing |
| 85 | 53 | 62 | 158 | 9 | missing | missing | missing |
| 86 | 49 | 58 | 163 | missing | missing | missing | missing |
| 87 | 43 | 56 | 159 | missing | 14.9 | 8.3 | 111.2 |
| 88 | 49 | 62 | 168 | 13 | missing | missing | missing |
| 89 | 57 | 59 | 162 | 77 | missing | missing | missing |
| 90 | 42 | 78 | 155 | 7 | 71.8 | 22.1 | 36.1 |
| 91 | 53 | 60 | 158 | 2 | missing | missing | missing |
| 92 | 59 | 58 | 158 | 84 | missing | missing | missing |
| 93 | 49 | 60 | 165 | 13 | missing | missing | missing |
| 94 | 51 | 47 | 143 | missing | 83.6 | 36.7 | 20 |
| 95 | 51 | 49 | 157 | 24 | 96 | 29.4 | 20 |
| 96 | 51 | 61 | 154 | 71 | missing | missing | missing |
| 97 | 52 | 55 | 160 | 15 | 167.8 | 78.7 | 20 |
| 98 | 58 | 76 | 169 | 120 | 65.3 | 24.9 | 20 |
| 99 | 74 | 65 | 159 | 228 | missing | missing | missing |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 100 | 48 | 62 | 155 | missing | 74.3 | 32.5 | 20.7 |
| 101 | 67 | 70 | 158 | 180 | missing | missing | missing |
| 102 | 43 | 80 | 170 | 72 | 68.1 | 45.1 | 48.1 |
| 103 | 51 | 65 | 162 | 6 | 89.1 | 41.9 | 20 |
| 104 | 52 | 64 | 167 | 4 | 42 | 13 | 10 |
| 105 | 48 | 60 | 150 | 2 | 70.8 | 55.1 | 48 |
| 106 | 50 | 59 | 157 | 8 | 200 | 63.3 | 27.4 |
| 107 | 49 | 61 | 160 | 24 | 162.3 | 34.6 | 20 |
| 108 | 53 | 74 | 160 | missing | missing | missing | missing |
| 109 | 52 | 66 | 159 | 11 | 112.8 | 47.8 | 20 |
| 110 | 47 | 63 | 160 | 7 | 11.8 | 6 | 110.6 |
| 111 | 48 | 73 | 165 | 72 | 51.9 | 19.9 | 149.4 |
| 112 | 51 | 64 | 162 | 73 | 133.8 | 34.8 | missing |
| 113 | 68 | 54 | 164 | 276 | missing | missing | missing |
| 114 | 46 | 75 | 155 | 2 | 173.3 | 64 | 20 |
| 115 | 52 | 70 | 170 | 24 | 67.7 | 31.5 | 209.7 |
| 116 | 55 | 60 | 165 | 84 | missing | missing | missing |
| 117 | 45 | 60 | 154 | 2 | missing | missing | missing |
| 118 | 43 | 70 | 158 | missing | 97.9 | 36 | 20 |
| 119 | 59 | 55 | 154 | 120 | 86.1 | 40.1 | 20 |
| 120 | 48 | 73 | 160 | 8 | 13.6 | 5.5 | 170.1 |
| 121 | 53 | 69 | 158 | 72 | 66.5 | 21.5 | 135.1 |
| 122 | 57 | 55 | 165 | 36 | missing | missing | missing |
| 123 | 46 | 89 | 155 | 74 | missing | missing | missing |
| 124 | 51 | 56 | 153 | missing | 36.6 | 15.2 | 48.4 |
| 125 | 57 | 70 | 160 | 72 | missing | missing | missing |
| 126 | 52 | 56 | 155 | 24 | missing | missing | missing |
| 127 | 57 | 70 | 165 | 23 | 136.4 | 45.9 | 20 |
| 128 | 53 | 64 | 160 | 16 | missing | missing | missing |
| 129 | 71 | 75 | 168 | 216 | missing | missing | missing |
| 130 | 52 | 70 | 169 | 96 | missing | missing | missing |
| 131 | 51 | 52 | 153 | missing | 30.4 | 25.3 | 278.8 |
| 132 | 50 | 60 | 162 | 9 | missing | missing | missing |
| 133 | 72 | 65 | 155 | 262 | missing | missing | missing |
| 134 | 47 | 63 | 163 | 36 | 118.6 | 57.8 | 38.3 |
| 135 | 52 | 65 | 158 | 78 | 119 | 38.3 | 31.1 |
| 136 | 49 | 59 | 163 | missing | missing | missing | missing |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 137 | 49 | 81 | 168 | 48 | 23.3 | 10.5 | 164.6 |
| 138 | 51 | 60 | 160 | missing | 53.7 | 15.4 | 23.7 |
| 139 | 53 | 72 | 162 | missing | 19.4 | 8.8 | 146.4 |
| 140 | 55 | 87 | 170 | 24 | 56.1 | 35.5 | 38.4 |
| 151 | 46 | 57 | 159 | 4.5 | missing | missing | missing |
| 152 | 47 | 78 | 169 | 60 | missing | missing | missing |
| 153 | 44 | missing | missing | 14 | missing | missing | missing |
| 154 | 48 | 79 | 174 | 24 | missing | missing | missing |
| 155 | 47 | 64 | 172 | 30 | missing | missing | missing |
| 156 | 46 | 66 | 161 | 4 | missing | missing | missing |
| 157 | 51 | 56 | 150 | 120 | 62.4 | 56.2 | 25.6 |
| 158 | 58 | 67 | 172 | 84 | missing | missing | missing |
| 159 | 68 | 66 | 153 | 252 | missing | missing | missing |
| 160 | 55 | 65 | 157 | 48 | 17.51 | 55.1 | 20 |
| 161 | 58 | 63 | 162 | 48 | 0.24 | 22.9 | 79.8 |
| 162 | 49 | 60 | 163 | missing | missing | missing | missing |
| 163 | 53 | 76 | missing | 40 | 129.5 | 39 | 25 |
| 164 | 55 | 94 | 170 | 72 | 88.8 | 32.3 | 25 |
| 165 | 55 | 60 | 155 | 72 | missing | missing | missing |
| 166 | 49 | 70 | 155 | 60 | missing | missing | missing |
| 167 | 46 | 68 | 162 | missing | missing | missing | missing |
| 168 | 54 | 56 | 165 | 72 | 129 | 35.4 | 20 |
| 169 | 45 | 75 | 164 | missing | missing | missing | missing |
| 170 | 54 | 69 | 171 | 60 | missing | missing | missing |
| 171 | 76 | missing | missing | missing | missing | missing | missing |
| 172 | 52 | 63 | 164 | missing | missing | missing | missing |
| 173 | 50 | 56 | 152 | 12 | missing | missing | missing |
| 174 | 53 | 55 | 158 | 84 | 185.5 | 70 | 20 |
| 175 | 49 | 79 | 167 | 12 | missing | missing | missing |
| 176 | 50 | 65 | 168 | 12 | missing | missing | missing |
| 177 | 67 | 61 | 156 | 312 | 223.8 | 58.5 | 20 |
| 178 | 49 | 69 | 165 | 36 | missing | missing | missing |
| 179 | 55 | 60 | 155 | 84 | missing | missing | missing |

| No | T3 | T4 | TSH | Glucose | Cholesterol | Triglyceride | HDL |
|---|---|---|---|---|---|---|---|
| 1 | 1.23 | 8.27 | 0.98 | 101 | 212 | missing | 43 |
| 2 | 1.19 | 9.38 | 3.59 | 98 | 224 | 53 | 102 |
| 3 | 0.83 | 8.52 | 1.21 | 93 | 177 | 80 | 63 |
| 4 | missing | missing | missing | 92 | 169 | 74 | 79 |
| 5 | 1 | 9.08 | 1.13 | 92 | 217 | 135 | 66 |
| 6 | 1.19 | 12.81 | 0.7 | 96 | 208 | 140 | 69 |
| 7 | missing | missing | missing | 90 | 215 | 60 | 43 |
| 8 | 0.75 | 9.72 | missing | 106 | 191 | 141 | 41 |
| 9 | 0.94 | 10.74 | 1.23 | 91 | 260 | 116 | 93 |
| 10 | missing | missing | missing | 86 | 246 | 75 | 52 |
| 11 | missing | missing | missing | 79 | 254 | 68 | 94 |
| 12 | 1.02 | 9.52 | 4.02 | 90 | 219 | 131 | 92 |
| 13 | 1.25 | 9.19 | 1.77 | 80 | 192 | 60 | 99 |
| 14 | 1.26 | 8.96 | 3.53 | 102 | 227 | 67 | 64 |
| 15 | 1.16 | 11.27 | 1.2 | 100 | 262 | 168 | 65 |
| 16 | missing | missing | missing | 77 | 209 | 58 | 62 |
| 17 | 4.6 | 18.9 | 0.34 | missing | 249 | 65 | 84 |
| 18 | 1.14 | 6.27 | 2.96 | 81 | 215 | 64 | 69 |
| 19 | 1.15 | 8.26 | 0.49 | 82 | 206 | 124 | 57 |
| 20 | 0.73 | 8.34 | 1.52 | 86 | 211 | 87 | 77 |
| 21 | 1.13 | 8.61 | 1.68 | 91 | 326 | 107 | 45 |
| 22 | 1.28 | 10.13 | 0.82 | 97 | 252 | 130 | 69 |
| 23 | missing | missing | missing | 108 | 216 | 98 | 84 |
| 24 | 1.16 | 10.03 | 1.35 | 76 | 254 | 172 | 62 |
| 25 | 0.85 | 8.1 | 1.63 | 97 | 241 | 57 | 67 |
| 26 | 1.15 | 15.12 | 1.1 | 153 | 200 | 114 | 80 |
| 27 | 1.07 | 10.19 | 0.38 | 95 | 243 | 122 | 71 |
| 28 | 0.95 | 9.43 | 2.27 | 83 | 301 | 153 | 65 |
| 29 | 1.05 | 6.88 | 2.51 | 87 | 275 | 63 | 71 |
| 30 | 1.29 | 13.91 | 1.19 | 82 | missing | missing | missing |
| 31 | missing | missing | missing | missing | missing | missing | missing |
| 32 | missing | missing | missing | missing | missing | missing | missing |
| 33 | 0.93 | 9.88 | 2.11 | 85 | 252 | 119 | 91 |
| 34 | missing | missing | missing | missing | missing | missing | missing |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 35 | missing | missing | missing | 78 | 217 | 94 | 90 |
| 36 | missing | missing | missing | missing | missing | missing | missing |
| 37 | 1.04 | 8.55 | 1.15 | 114 | 218 | 102 | 54 |
| 38 | missing | missing | missing | missing | missing | missing | missing |
| 39 | 0.81 | 9.16 | 1.16 | 97 | 202 | 63 | 83 |
| 40 | missing | missing | missing | missing | missing | missing | missing |
| 41 | missing | missing | missing | missing | missing | missing | missing |
| 42 | missing | missing | missing | 86 | missing | 107 | missing |
| 43 | missing | missing | missing | 93 | 202 | 55 | 76 |
| 44 | missing | missing | missing | 87 | 217 | 70 | 47 |
| 45 | 0.72 | 8.71 | 0.65 | 86 | 262 | 162 | 44 |
| 46 | missing | missing | missing | missing | missing | missing | missing |
| 47 | missing | missing | missing | missing | missing | missing | missing |
| 48 | missing | missing | missing | missing | missing | missing | missing |
| 49 | 1.19 | 12.81 | 0.7 | 96 | 208 | 140 | 69 |
| 50 | missing | missing | missing | missing | missing | missing | missing |
| 51 | missing | missing | missing | 112 | 236 | 125 | 66 |
| 52 | missing | missing | missing | missing | missing | missing | missing |
| 53 | missing | missing | missing | missing | missing | missing | missing |
| 54 | 1.18 | 12.65 | 1.8 | 92 | 272 | 117 | 93 |
| 55 | missing | missing | missing | missing | missing | missing | missing |
| 56 | 1.07 | 7.68 | 3.84 | 72 | 247 | 97 | 73 |
| 57 | 0.84 | 7.91 | 0.33 | 66 | 232 | 197 | 59 |
| 58 | 0.8 | 9.06 | missing | 92 | 250 | 130 | 62 |
| 59 | missing | missing | missing | 94 | 250 | 361 | 46 |
| 60 | 1.1 | 10.47 | 1.7 | 88 | 309 | 132 | 74 |
| 61 | missing | missing | missing | 94 | 228 | 88 | 75 |
| 62 | missing | missing | missing | missing | missing | missing | missing |
| 63 | 1.23 | 12.37 | 0.95 | 86 | 219 | 107 | 71 |
| 64 | 0.58 | 6.3 | 4.66 | 84 | 264 | 76 | 39 |
| 65 | 1.29 | 11.94 | 0.75 | 94 | 180 | 119 | 41 |
| 66 | 1.2 | 7.82 | 1.11 | 85 | missing | missing | missing |
| 67 | 0.92 | 8.68 | 24.09 | 94 | 154 | 73 | 45 |
| 68 | 0.6 | 8.01 | 0.06 | 81 | 214 | 91 | 76 |
| 69 | missing | missing | missing | missing | missing | missing | missing |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 70 | missing | missing | missing | missing | missing | missing | missing |
| 71 | 0.91 | 7.26 | 2.25 | 88 | 282 | 96 | 85 |
| 72 | 1.03 | 5.5 | 1.21 | 86 | 223 | 82 | 75 |
| 73 | 1.47 | 13.78 | 0.26 | 98 | 172 | 153 | 66 |
| 74 | missing | missing | missing | missing | missing | missing | missing |
| 75 | 0.99 | 6.66 | 1.66 | 88 | 250 | 119 | 49 |
| 76 | 0.99 | 9.66 | 1.62 | 90 | 194 | 89 | 81 |
| 77 | missing | missing | missing | missing | missing | missing | missing |
| 78 | 0.87 | 10.02 | 1.42 | 98 | 163 | 63 | 62 |
| 79 | 0.94 | 9.37 | 0.71 | 86 | 360 | 133 | 68 |
| 80 | missing | missing | missing | 91 | 216 | 61 | 86 |
| 81 | missing | missing | missing | 75 | 13 | 79 | 33 |
| 82 | 1.04 | 11.45 | 3.41 | 97 | 227 | 123 | 56 |
| 83 | missing | missing | missing | missing | missing | missing | missing |
| 84 | missing | missing | missing | 107 | 0.8 | 87 | missing |
| 85 | missing | missing | missing | missing | missing | missing | missing |
| 86 | missing | missing | missing | missing | missing | missing | missing |
| 87 | 0.78 | 9.04 | 4.01 | 91 | 259 | 118 | 59 |
| 88 | missing | missing | missing | missing | missing | missing | missing |
| 89 | missing | missing | missing | missing | missing | missing | missing |
| 90 | 1.13 | 9.72 | 0.67 | 106 | 134 | 54 | 39 |
| 91 | missing | missing | missing | missing | missing | missing | missing |
| 92 | missing | missing | missing | missing | missing | missing | missing |
| 93 | missing | missing | missing | missing | missing | missing | missing |
| 94 | 1.23 | 10.41 | 0.17 | 95 | 154 | 84 | 68 |
| 95 | 0.81 | 8.59 | 4.68 | 80 | 247 | 57 | 91 |
| 96 | missing | missing | missing | missing | missing | missing | missing |
| 97 | 1.11 | 9.15 | 3.2 | 88 | 259 | 82 | 83 |
| 98 | 1.01 | 8.12 | 6.57 | 100 | 247 | 10.9 | 54 |
| 99 | missing | missing | missing | missing | missing | missing | missing |
| 100 | 1.06 | 9.14 | 2.5 | 119 | 196 | 85 | 50 |
| 101 | missing | missing | missing | missing | missing | missing | missing |
| 102 | 1.02 | 7.21 | 1.5 | 84 | 212 | 56 | 52 |
| 103 | 1.04 | 8.46 | 2.26 | 100 | 235 | 61 | 75 |
| 104 | 1.2 | 8.2 | 4.4 | 91 | 221 | 149 | 4 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 105 | missing | missing | missing | 93 | 167 | 74 | 52 |
| 106 | 1.37 | 9.76 | 2.52 | 93 | 206 | 155 | 46 |
| 107 | missing | missing | missing | missing | missing | missing | missing |
| 108 | missing | missing | missing | missing | missing | missing | missing |
| 109 | 1.12 | 8.03 | 1 | 96 | 177 | 119 | 54 |
| 110 | 0.83 | 9.09 | 1.51 | 104 | 193 | 128 | 50 |
| 111 | 1.13 | 12.36 | 0.49 | 69 | 198 | 237 | 69 |
| 112 | 0.89 | 11.45 | 2.1 | 101 | 278 | 60 | 66 |
| 113 | missing | missing | missing | missing | missing | missing | missing |
| 114 | 0.9 | 7.41 | 1.92 | 95 | 243 | 47 | 110 |
| 115 | missing | missing | missing | 81 | 221 | 123 | 44 |
| 116 | missing | missing | missing | missing | missing | missing | missing |
| 117 | missing | missing | missing | missing | missing | missing | missing |
| 118 | 1.21 | 9.29 | 1.43 | 97 | 242 | 108 | 69 |
| 119 | missing | missing | missing | missing | 202 | 153 | 60 |
| 120 | 1.23 | 9.39 | 2.34 | 105 | 199 | 50 | 56 |
| 121 | 0.89 | 11.11 | 1.69 | 84 | 251 | 189 | 99 |
| 122 | missing | missing | missing | 94 | 193 | 64 | 65 |
| 123 | 0.32 | 1.81 | 100 | 135 | 384 | 86 | 46 |
| 124 | 1.4 | 12.31 | 0.42 | missing | missing | missing | missing |
| 125 | missing | missing | missing | missing | missing | missing | missing |
| 126 | missing | missing | missing | missing | missing | missing | missing |
| 127 | 0.89 | 10.94 | 1.09 | 90 | 214 | 84 | 65 |
| 128 | missing | missing | missing | missing | missing | missing | missing |
| 129 | missing | missing | missing | missing | missing | missing | missing |
| 130 | missing | missing | missing | missing | missing | missing | missing |
| 131 | 1.36 | 10.14 | 2.3 | 97 | 213 | 66 | 88 |
| 132 | missing | missing | missing | missing | missing | missing | missing |
| 133 | missing | missing | missing | missing | missing | missing | missing |
| 134 | 0.94 | 8.64 | 1.98 | 77 | 155 | 43 | 57 |
| 135 | 1.14 | 7.49 | 2.01 | 85 | 230 | 111 | 52 |
| 136 | missing | missing | missing | missing | missing | missing | missing |
| 137 | 1.31 | 10.69 | 1.36 | 92 | 226 | 54 | 87 |
| 138 | missing | missing | missing | missing | missing | missing | missing |
| 139 | 0.82 | 2.89 | 100 | 89 | 231 | 107 | 74 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 140 | missing | missing | missing | 87 | 182 | 86 | 69 |
| 151 | missing | missing | missing | missing | missing | missing | missing |
| 152 | 1.2 | 8.76 | 0.45 | 80 | 271 | 95 | 73 |
| 153 | missing | missing | missing | missing | missing | missing | missing |
| 154 | 1.04 | 7.01 | 0.7 | 85 | 206 | 170 | 44 |
| 155 | missing | missing | missing | missing | missing | missing | missing |
| 156 | 1.19 | 8.15 | 1.41 | 77 | 236 | 103 | 57 |
| 157 | 1.04 | 8.71 | 0.9 | missing | missing | missing | missing |
| 158 | 0.86 | 4.88 | 2.01 | 85 | 189 | 60 | 60 |
| 159 | 1.28 | 7.91 | 1.89 | 85 | 181 | 47 | 39 |
| 160 | missing | missing | missing | missing | missing | missing | missing |
| 161 | missing | missing | missing | missing | missing | missing | missing |
| 162 | missing | missing | missing | missing | missing | missing | missing |
| 163 | missing | missing | missing | missing | missing | missing | missing |
| 164 | missing | missing | missing | 104 | 218 | 129 | 45 |
| 165 | missing | missing | missing | missing | missing | missing | missing |
| 166 | missing | missing | missing | missing | missing | missing | missing |
| 167 | 1.05 | 8.52 | 0.59 | 84 | 182 | 54 | 62 |
| 168 | missing | missing | missing | missing | missing | missing | missing |
| 169 | missing | missing | missing | missing | missing | missing | missing |
| 170 | 1.04 | 10.66 | 1.37 | 97 | 214 | 137 | 60 |
| 171 | 1.34 | 7.85 | 5.77 | 117 | 248 | 219 | 52 |
| 172 | missing | missing | missing | missing | missing | missing | missing |
| 173 | 0.84 | 6.67 | 0.85 | 79 | 233 | 68 | 102 |
| 174 | 1.12 | 7.27 | 2.42 | 83 | 237 | 86 | 63 |
| 175 | missing | missing | missing | missing | missing | missing | missing |
| 176 | missing | missing | missing | missing | missing | missing | missing |
| 177 | missing | missing | missing | missing | missing | missing | missing |
| 178 | 3.67 | 1.1 | 5.64 | 94 | 190 | 97 | 82 |
| 179 | missing | missing | missing | missing | missing | missing | missing |

| No | LDL | VLDL | Hemoglobin | Hematocrite | Men. Type | HRT | HRT Duration |
|---|---|---|---|---|---|---|---|
| 1 | 134.6 | 34.4 | 12.8 | 38.6 | 1 | 2 | 0 |
| 2 | 111.4 | 10.6 | 12.7 | 36.8 | 1 | 2 | 0 |
| 3 | 98 | 16 | 13.6 | 40.1 | 4 | 2 | 0 |
| 4 | 75.2 | 14.8 | 15.1 | 44.4 | 1 | 2 | 0 |
| 5 | 124 | 27 | 14 | 40.6 | 1 | 1 | 42 |
| 6 | 111 | 28 | 13.3 | 39.5 | 1 | 1 | 24 |
| 7 | 160 | 12 | 13.1 | 38.1 | 1 | 2 | 0 |
| 8 | 121.8 | 28.2 | 12.9 | 38.5 | 2 | 2 | 0 |
| 9 | 143.8 | 23.2 | 13.3 | 39.2 | 1 | 1 | 96 |
| 10 | 209 | 15 | 14.6 | 42.1 | 1 | 2 | 0 |
| 11 | 146 | 14 | 13.1 | 39.7 | 1 | 1 | 12 |
| 12 | 100.8 | 26.2 | 13.9 | 41.4 | 2 | 1 | 2 |
| 13 | 81 | 12 | 12.5 | 37.2 | 1 | 2 | 0 |
| 14 | 149.6 | 13.4 | 13 | 38.1 | 4 | 2 | 0 |
| 15 | 163.4 | 33.6 | missing | missing | 1 | 1 | 36 |
| 16 | 130 | 17 | 13.7 | 40.6 | 1 | 2 | 0 |
| 17 | 152 | missing | missing | missing | 1 | 2 | 0 |
| 18 | 133.2 | 12.8 | 13.7 | 40.2 | 1 | 2 | 0 |
| 19 | 124.2 | 24.8 | 14.3 | 41.6 | 1 | 2 | 0 |
| 20 | 116.6 | 17.4 | 14.7 | 42.1 | 1 | 2 | 0 |
| 21 | 260 | 21 | 15.7 | 46.3 | 1 | 1 | 38.17 |
| 22 | 157 | 26 | 14.6 | 42.4 | 1 | 1 | 48 |
| 23 | 112 | 20 | 13.7 | 42.2 | 2 | 1 | 2 |
| 24 | 157.6 | 34.4 | 12.4 | 37.8 | 1 | 1 | 60 |
| 25 | 162.6 | 11.4 | 14.6 | 40.7 | 1 | 2 | 0 |
| 26 | 97.2 | 22.8 | 16.3 | 46.5 | 2 | 1 | 84 |
| 27 | 147.6 | 24.4 | 14.3 | 42.3 | 1 | 2 | 0 |
| 28 | 205.4 | 30.6 | 14.9 | 43.9 | 1 | 2 | 0 |
| 29 | 191.4 | 12.6 | 13.4 | 38.4 | 1 | 2 | 0 |
| 30 | missing | missing | 10.6 | 32.3 | 1 | 2 | 0 |
| 31 | missing | missing | missing | missing | 2 | 1 | 96 |
| 32 | missing | missing | missing | missing | 1 | 2 | 0 |
| 33 | 137.2 | 23.8 | 13.5 | 40.5 | 2 | 1 | 45 |
| 34 | missing | missing | missing | missing | 1 | 1 | 120 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 35 | 108.2 | 18.8 | missing | missing | 1 | 2 | 0 |
| 36 | missing | missing | missing | missing | 1 | 1 | 22 |
| 37 | 143.6 | 20.4 | 13.9 | 40.8 | 1 | 2 | 0 |
| 38 | missing | missing | missing | missing | 1 | 1 | 2 |
| 39 | 106.4 | 12.6 | 13.1 | 38.9 | 1 | 2 | 0 |
| 40 | missing | missing | missing | missing | 1 | 2 | 0 |
| 41 | missing | missing | missing | missing | 4 | 2 | 0 |
| 42 | missing | missing | 13.8 | 42.6 | 1 | 1 | 4 |
| 43 | 115 | 11 | 14 | 40.7 | 1 | 1 | 30 |
| 44 | 156 | 14 | 14.7 | 43.5 | 1 | 1 | 24 |
| 45 | 185.6 | 32.4 | 15.1 | 45.3 | 2 | 1 | 71 |
| 46 | missing | missing | missing | missing | 1 | 2 | 0 |
| 47 | missing | missing | missing | missing | 2 | 1 | 5 |
| 48 | missing | missing | missing | missing | 1 | 1 | 84 |
| 49 | 111 | 28 | 13.3 | 39.5 | 1 | 1 | 24 |
| 50 | missing | missing | missing | missing | 1 | 2 | 0 |
| 51 | 145 | 25 | 13.7 | 41.8 | 1 | 2 | 0 |
| 52 | missing | missing | missing | missing | 3 | 2 | 0 |
| 53 | missing | missing | missing | missing | 1 | 2 | 0 |
| 54 | 155.6 | 23.4 | 13.6 | 39.8 | 1 | 1 | 12 |
| 55 | missing | missing | missing | missing | 1 | 2 | 0 |
| 56 | 154.6 | 19.4 | 15.1 | 44.2 | 1 | 2 | 0 |
| 57 | 133.6 | 39.4 | 12.9 | 37.8 | 3 | 2 | 0 |
| 58 | 162 | 26 | 14.1 | 42.4 | 1 | 2 | 0 |
| 59 | 131.8 | 72.2 | 13.8 | 39.9 | 1 | 2 | 0 |
| 60 | 208.6 | 26.4 | 13 | 37.1 | 1 | 2 | 0 |
| 61 | 135.4 | 17.6 | 13.8 | 39.7 | 3 | 1 | 38.17 |
| 62 | missing | missing | missing | missing | 3 | 2 | 0 |
| 63 | 126.6 | 21.4 | 14.1 | 41.3 | 2 | 2 | 0 |
| 64 | 209.8 | 15.2 | 15.3 | 46.7 | 2 | 1 | 84 |
| 65 | 115.2 | 23.8 | 12.6 | 38.3 | 3 | 2 | 0 |
| 66 | missing | missing | 13.6 | 40.8 | 1 | 2 | 0 |
| 67 | 94.4 | 14.6 | 15 | 42.1 | 4 | 2 | 0 |
| 68 | 119.8 | 18.2 | 12.4 | 38.5 | 3 | 1 | 48 |
| 69 | missing | missing | missing | missing | 1 | 2 | 0 |
| 70 | missing | missing | missing | missing | 2 | 2 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 71 | 177.8 | 19.2 | 14.4 | 41.8 | 1 | 2 | 0 |
| 72 | 131.6 | 16.4 | 14.6 | 42.8 | 2 | 1 | 31 |
| 73 | 75.4 | 30.6 | 13.9 | 41.6 | 2 | 1 | 18 |
| 74 | missing | missing | missing | missing | 1 | 1 | 36 |
| 75 | 177.2 | 23.8 | 13.1 | 37.3 | 1 | 2 | 0 |
| 76 | 95.2 | 18 | 12.6 | 35.8 | 3 | 1 | 48 |
| 77 | missing | missing | missing | missing | 1 | 2 | 0 |
| 78 | 88.4 | 12.6 | 13.3 | 39.5 | 3 | 2 | 0 |
| 79 | 265.4 | 26.6 | 12.4 | 35.5 | 1 | 2 | 0 |
| 80 | 117.8 | 12.2 | 5.14 | 42.9 | 2 | 2 | 0 |
| 81 | 166.2 | 15.8 | 13.9 | 40.6 | 2 | 1 | 3 |
| 82 | 146.4 | 24.6 | 14.2 | 42.6 | 3 | 2 | 0 |
| 83 | missing | missing | missing | missing | 2 | 1 | 84 |
| 84 | missing | missing | missing | missing | 3 | 2 | 0 |
| 85 | missing | missing | missing | missing | 1 | 1 | 12 |
| 86 | missing | missing | missing | missing | 3 | 2 | 0 |
| 87 | 176.4 | 23.6 | 14.3 | 41.8 | 2 | 2 | 0 |
| 88 | missing | missing | missing | missing | 1 | 2 | 0 |
| 89 | missing | missing | missing | missing | 1 | 1 | 68 |
| 90 | 84.2 | 10.8 | 14.6 | 42.4 | 1 | 2 | 0 |
| 91 | missing | missing | missing | missing | 3 | 2 | 0 |
| 92 | missing | missing | missing | missing | 2 | 1 | 84 |
| 93 | missing | missing | missing | missing | 1 | 1 | 8 |
| 94 | 69.2 | 16.8 | 12.3 | 38.8 | 3 | 2 | 0 |
| 95 | 144.6 | 11.4 | 41 | 89.8 | 1 | 2 | 0 |
| 96 | missing | missing | missing | missing | 1 | 2 | 0 |
| 97 | 159.6 | 16.4 | 13.8 | 40.3 | 3 | 1 | 12 |
| 98 | 171.2 | 21.8 | 13.9 | 42.4 | 1 | 2 | 0 |
| 99 | missing | missing | missing | missing | 1 | 2 | 0 |
| 100 | 129 | 17 | 12.2 | 39.6 | 3 | 2 | 0 |
| 101 | missing | missing | missing | missing | 1 | 2 | 0 |
| 102 | 148.8 | 11.2 | 13.6 | 39.3 | 1 | 1 | 59 |
| 103 | 147.8 | 12.2 | 12.5 | 38.4 | 1 | 2 | 0 |
| 104 | 144 | 29 | 13.4 | 40.1 | 1 | 1 | 4 |
| 105 | 100.2 | 14.8 | 13.6 | 39.3 | 2 | 1 | 1 |
| 106 | 129 | 31 | 14.8 | 41.3 | 1 | 1 | 8 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 107 | missing | missing | missing | missing | 1 | 2 | 0 |
| 108 | missing | missing | missing | missing | 3 | 1 | 18 |
| 109 | 99.2 | 23.8 | 13.8 | 40.9 | 1 | 1 | 38.17 |
| 110 | 111.4 | 25.6 | 13.8 | 40 | 3 | 1 | 3 |
| 111 | 81.6 | 47.4 | 12.6 | 38.9 | 2 | 1 | 68 |
| 112 | 200 | 12 | 13.3 | 40.12 | 1 | 2 | 0 |
| 113 | missing | missing | missing | missing | 1 | 1 | 84 |
| 114 | 123.6 | 9.4 | 13.4 | 40 | 3 | 2 | 0 |
| 115 | 152 | 25 | 12.6 | 36.2 | 1 | 1 | 12 |
| 116 | missing | missing | missing | missing | 1 | 1 | 19 |
| 117 | missing | missing | missing | missing | 1 | 1 | 1 |
| 118 | 151.4 | 21.6 | 13.7 | 39.2 | 3 | 1 | 1 |
| 119 | 11.4 | 30.6 | 12.7 | 35.1 | 1 | 2 | 0 |
| 120 | 133 | 10 | 13.5 | 39.9 | 1 | 2 | 0 |
| 121 | 114.2 | 37.8 | 13.8 | 40.5 | 1 | 1 | 116 |
| 122 | 115.2 | 12.8 | 13.7 | 39.5 | 1 | 1 | 3 |
| 123 | 166 | 172 | missing | missing | 2 | 1 | 74 |
| 124 | missing | missing | missing | missing | 3 | 1 | 36 |
| 125 | missing | missing | missing | missing | 1 | 1 | 72 |
| 126 | missing | missing | missing | missing | 1 | 2 | 0 |
| 127 | 132.2 | 16.8 | 13.5 | 39 | 1 | 2 | 0 |
| 128 | missing | missing | missing | missing | 1 | 2 | 0 |
| 129 | missing | missing | missing | missing | 1 | 2 | 0 |
| 130 | missing | missing | missing | missing | 2 | 1 | 96 |
| 131 | 111.8 | 13.2 | 13.5 | 39.5 | 4 | 1 | 96 |
| 132 | missing | missing | missing | missing | 1 | 2 | 0 |
| 133 | missing | missing | missing | missing | 1 | 2 | 0 |
| 134 | 89.4 | 8.6 | 14.1 | 40.7 | 1 | 2 | 0 |
| 135 | 155.8 | 22.2 | 12.9 | 35.7 | 2 | 2 | 0 |
| 136 | missing | missing | missing | missing | 4 | 2 | 0 |
| 137 | 128.2 | 10.8 | 13.2 | 39.2 | 2 | 1 | 48 |
| 138 | missing | missing | missing | missing | 4 | 2 | 0 |
| 139 | 135.6 | 21.4 | 12.9 | 37.7 | 4 | 2 | 0 |
| 140 | 95.8 | 17.2 | 12.8 | 36.4 | 1 | 2 | 0 |
| 151 | missing | missing | missing | missing | 2 | 1 | 10 |
| 152 | 179 | 19 | missing | missing | 1 | 2 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 153 | missing | missing | missing | missing | 3 | 1 | 1 |
| 154 | 128 | 34 | 19.7 | 58.1 | 1 | 2 | 0 |
| 155 | missing | missing | missing | missing | 1 | 2 | 0 |
| 156 | 158.4 | 20.6 | 13 | 37.02 | 1 | 2 | 0 |
| 157 | missing | missing | 13.1 | 37.2 | 1 | 2 | 0 |
| 158 | 117 | 12 | 12.6 | 36.8 | 1 | 2 | 0 |
| 159 | 132.6 | 9.4 | 15 | 42.4 | 1 | 1 | 24 |
| 160 | missing | missing | missing | missing | 1 | 2 | 0 |
| 161 | missing | missing | missing | missing | 1 | 2 | 0 |
| 162 | missing | missing | missing | missing | 1 | 2 | 0 |
| 163 | missing | missing | missing | missing | 1 | 1 | 12 |
| 164 | 147 | 26 | missing | missing | 1 | 2 | 0 |
| 165 | missing | missing | missing | missing | 1 | 2 | 0 |
| 166 | missing | missing | missing | missing | 1 | 2 | 0 |
| 167 | 109.2 | 10.8 | 14.4 | 41.9 | 1 | 1 | 12 |
| 168 | missing | missing | missing | missing | 1 | 2 | 0 |
| 169 | missing | missing | missing | missing | 2 | 1 | 36 |
| 170 | 126.6 | 27.4 | 13.5 | 39.3 | 1 | 2 | 0 |
| 171 | 152.2 | 43.8 | 14.9 | 41.2 | 1 | 2 | 0 |
| 172 | missing | missing | missing | missing | 1 | 2 | 0 |
| 173 | 117.4 | 13.6 | 12.8 | 37.2 | 1 | 2 | 0 |
| 174 | 153.8 | 17.2 | 14.6 | 42.8 | 1 | 2 | 0 |
| 175 | missing | missing | missing | missing | 1 | 2 | 0 |
| 176 | missing | missing | missing | missing | 1 | 1 | 64 |
| 177 | missing | missing | missing | missing | 4 | 2 | 0 |
| 178 | 88.6 | 19.4 | 13.02 | 37.2 | 1 | 2 | 0 |
| 179 | missing | missing | missing | missing | 3 | 2 | 0 |