An Approach for Solving Missing Values in Data Set Using Clustering-Curve Fitting Technique

Dr. Kadhim AlJanabi
Dept of CS
Faculty of Math and CS
University of Kufa
kadhim.aljanabi@uokufa.edu.iq

Dr. Mansoor Habeebi
Dept of CS
Faculty of Math and CS
University of Kufa
Mansoor.Habeeb@uokufa.edu.iq

Nawras Riyadh Neamah
M.Sc. Student Dept of Mathematics
Faculty of Math and CS
University of Kufa
Nawras.8888@yahoo.com

Abstract:

Missing values in data sets represent one of the greatest challenge in analyzing data to extract knowledge from the data set. The work in this paper presents a new approach for solving the missing values problems by using and merging two different techniques; clustering (K-means and Expectation Maximization) and curve fitting. More than twenty thousand records of real health data set collected from different Iraqi hospitals were used to create and test the proposed approach that showed better results than the most popular techniques for estimation missing values such as most common values, overall overage, class average, and class most common values. Different software were used in the proposed work including WEKA (Waikato Environment for Knowledge Analysis), Matlab, Excel and C++.

Keywords: Data Mining, Missing Values, Clustering, Curve Fitting.

1. Introduction

Data Mining, or Knowledge Discovery in Databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as classification, association, prediction, clustering, data summarization, learning classification rules, analyzing changes, and detecting anomalies.

The analogy with the mining process is described as: Data mining refers to "using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful"

Basically data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It is the computer which is responsible for finding the patterns by identifying the underlying rules and features in the data. The idea is that it is possible to strike gold in unexpected places as the data mining software extracts patterns not previously discernable or so obvious that no-one has noticed them before.

The overall mining process and due to the data integration from different sources suffers from three main difficulties, they are:

1. Incompleteness
2. Inconsistent
3. Noisy data

Incompleteness referees to the existence of what is known as missing values in attributes or tuples that may occur due to many reasons like human errors, instrument bad functioning and others.
Data inconsistent refers to different naming and representation of data since they are collected from different sources each has its own data format and platforms. Noisy data may occur due to human errors, data conversion, machine restriction and others.

In On Line Transaction Processing (OLTP), the difficulties mentioned above do not represent big obstacle in data processing, whereas they represent a big challenge in On Line Analytical Processing due to the required nature of data that can be directed for analysis process other than the transactional approach.

2. Data Collection

In this research, the real data were collected from Al-Sader Medical City in Najaf at Iraq. More than 61500 records were collected from which only 20000 records have been used in the proposed solutions, Table (1) shows sample of the collected data. One hundred records from the selected data were used as missing values by deleting the age field. The data collected was stored in the FoxPro program. Then, the data was converted into Excel format by method of exporting the data into HTML format and then importing it into Excel.

The following characteristics were noticed in the collected data:

1. Some information collected was not consistent with the goal of the project and the whole idea of the patient data. They are not useful for analysis purposes.
2. Noisy data were deleted from the data set due to human errors but they were very rare and do not affect the overall analytical process.

Table (1). Patient Data Set (Sample of 50 Records with missing values).

number	Record#	Hospital	Medical Unit	Province	Elimination	Hand	Gender	Educational	Job	Status	Age	Patient province	Accomm-odation	Nationality	Input year	Input month	Input day	Out year	Out month	Out day	Disease	Lying
1	137	30	104	18	1	1	2	0	9	1		18	1	1	2011	1	3	2011	1	3	H04	0
2	204	30	101	18	1	1	2	2	13	1	20	18	1	1	2011	1	3	2011	1	3	L87	2
3	105	30	101	18	1	1	2	2	13	1	14	18	1	1	2011	1	3	2011	1	3	Q64	2
4	107	30	101	18	1	1	1	0	9	1	2	18	1	1	2011	1	2	2011	1	4	Q64	2
5	106	30	101	18	1	1	1	2	6	2	50	18	1	1	2011	1	2	2011	1	4	E14	2
6	221	30	101	18	1	1	2	4	8	1	7	18	1	1	2011	1	4	2011	1	4	Q36	2
7	158	30	101	18	1	1	2	0	9	1	1	6	1	1	2011	1	3	2011	1	5	Q36	2
8	97	30	101	18	1	1	2	2	13	1	16	18	1	1	2011	1	2	2011	1	5	L87	2
9	211	30	101	18	1	1	2	0	9	1	2	18	1	1	2011	1	3	2011	1	5	Q64	2
10	215	30	101	18	1	1	1	5	6	2	27	18	1	1	2011	1	4	2011	1	4	L02	2
11	201	30	101	18	1	1	1	5	6	2		6	1	1	2011	1	4	2011	1	4	Q38	3
12	241	30	105	18	1	1	1	2	6	2	45	18	1	1	2011	1	4	2011	1	4	H61	3
13	273	30	101	18	1	1	1	5	5	2	35	18	1	1	2011	1	5	2011	1	5	S67	0
14	25136	30	401	18	1	1	2	2	13	2	60	18	1	1	2010	12	29	2011	1	3	I20	0
15	25137	30	401	18	1	1	2	1	10	2	82	18	1	1	2010	12	30	2011	1	3	I20	0
16	25012	30	408	18	1	1	2	2	13	2	65	18	1	1	2010	12	28	2011	1	2	E14	0
17	25014	30	401	18	1	1	1	2	6	2	52	18	1	1	2010	12	28	2011	1	2	I21	0
18	12	30	401	18	1	1	2	2	13	2	60	18	1	1	2011	1	1	2011	1	3	I70	0
19	25080	30	401	18	1	1	2	2	13	2	55	18	1	1	2010	12	29	2011	1	3	I70	0
20	25102	30	401	18	1	1	1	2	10	2	60	18	1	1	2010	12	29	2011	1	3	I70	0
21	25158	30	401	18	1	1	1	1	10	2	60	18	1	1	2010	12	30	2011	1	4	I50	0
22	24645	30	401	18	1	1	1	4	7	2	60	18	1	1	2010	12	22	2011	1	4	I21	0
23	88	30	401	18	1	1	1	2	6	2	35	18	1	1	2011	1	2	2011	1	5	I20	0
24	9	30	408	18	1	1	2	1	10	2	65	18	1	1	2011	1	1	2011	1	4	I50	0
25	26	30	101	18	1	1	1	4	6	1	21	18	1	1	2011	1	2	2011	1	2	N30	3
26	25096	30	101	18	1	1	1	5	10	2	60	18	1	1	2010	12	29	2011	1	2	N41	3
27	24954	30	101	18	1	1	1	4	6	1	15	18	1	1	2010	12	27	2011	1	2	N35	3
28	27	30	101	18	1	1	1	5	10	2	70	13	1	1	2011	1	1	2011	1	2	N20	0
29	25075	30	101	18	1	1	1	5	10	2	65	13	1	1	2010	12	29	2011	1	2	N41	3
30	25097	30	101	18	1	1	1	4	8	1	12	18	1	1	2010	12	29	2011	1	1	K40	3
31	25117	30	101	18	1	1	1	5	10	2	65	18	1	1	2010	12	29	2011	1	1	N30	3
32	25065	30	101	18	1	1	1	5	10	2	60	18	1	1	2010	12	29	2011	1	2	N41	3
33	24951	30	101	18	1	1	1	4	10	2	60	18	1	1	2010	12	27	2011	1	2	N41	3
34	25115	30	101	18	1	1	1	5	6	2	48	16	1	1	2010	12	29	2011	1	2	N20	3
35	25094	30	101	18	1	1	1	5	6	2	27	18	1	1	2010	12	29	2011	1	1	N44	3
36	25093	30	101	18	1	1	1	5	6	2	25	18	1	1	2010	12	29	2011	1	1	N20	3
37	23922	30	412	18	1	1	1	5	10	2	70	18	1	1	2010	12	11	2011	1	2	C67	3
38	24883	30	412	18	1	1	1	5	10	2	60	18	1	1	2010	12	26	2011	1	2	C67	3
39	120	30	405	18	1	1	1	10	5	2	41	9	1	1	2011	1	3	2011	1	4	N18	0
40	25060	30	405	18	1	1	2	2	13	2	42	18	1	1	2010	12	29	2011	1	2	N18	0
41	142	30	405	18	1	1	1	2	10	2	66	18	1	1	2011	1	2	2011	1	4	N18	0
42	72	30	405	18	1	1	1	2	6	1	29	18	2	1	2011	1	2	2011	1	4	N18	0
43	24389	30	405	18	1	1	2	2	13	2	42	18	1	1	2010	12	19	2011	1	4	N18	0
44	25004	30	412	18	1	1	1	5	10	2	59	18	1	1	2010	12	28	2011	1	3	C67	0
45	1174	30	101	18	1	1	1	4	6	2	25	18	1	1	2011	1	2	2011	1	4	N44	3
46	25156	30	101	18	1	1	1	5	5	2	54	18	1	1	2010	12	30	2011	1	3	N35	0
47	180	30	101	18	1	1	2	1	10	2	65	18	1	1	2011	1	4	2011	1	4	N21	2
48	109	30	101	18	1	1	2	1	10	2	55	18	1	1	2011	1	2	2011	1	4	N21	2
49	101	30	105	18	1	1	1	6	8	1	17	18	1	1	2011	1	2	2011	1	3	J03	2
50	24865	30	101	18	1	1	1	0	9	1	6	18	1	1	2010	12	26	2011	1	3	N30	0

3. Data preprocessing

Real world data usually have the following drawbacks: Incompleteness, Noisy and Inconsistence. So, these data need to be preprocessed to get the data suitable for analysis purposes, and the preprocessing includes the following tasks [1, 2 ,3, 4]:

- Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Data integration: using multiple databases, data cubes, or files.
- Data transformation: normalization and aggregation.
- Data reduction: reducing the volume but producing the same or similar analytical results.
- Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

The process of extracting data from source systems and bringing it into the data warehouse is commonly called ETL, which stands for extraction, transformation, and loading.

During extraction, the desired data is identified and extracted from many different sources, including database systems and applications. After data is extracted, it has to be physically transported to the target system or to an intermediate system for further processing. The diagram in Figure (1) shows the preprocessing activities.

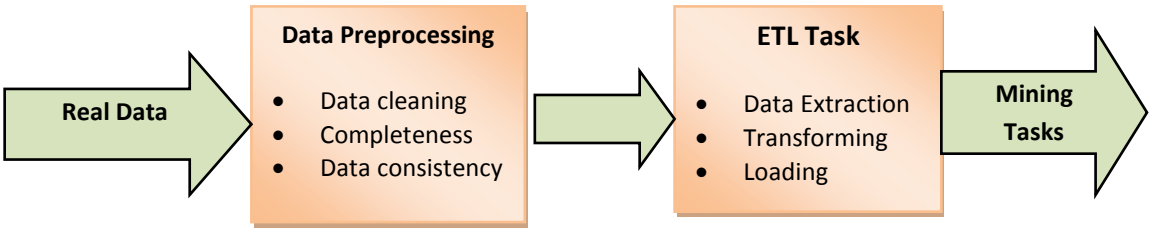

Figure (1): Data Preprocessing.

3.1. Clustering

Clustering techniques consider data tuples as objects. They partition the objects into groups or clusters, so that objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters. Similarity is commonly defined in terms of how "close" the objects are in space, based on a distance function. The "quality" of a cluster may be represented by its diameter, the maximum distance between any two objects in the cluster. Centroid distance is an alternative measure of cluster quality and is defined as the average distance of each cluster object from the cluster centroid (denoting the "average object," or average point in space for the cluster).[1,5]

3.1.1. Algorithms:

3.1.1.1. K-means Algorithm

The classic clustering technique is called *k-means*. First, you specify in advance how many clusters are being

sought: This is the parameter *k*. Then *k* points are chosen at random as cluster centers. All instances are assigned to their closest cluster center according to the ordinary Euclidean distance metric. Next the centroid, or mean, of the instances in each cluster is calculated—this is the "means" part. These centroids are taken to be new center values for their respective clusters. Finally, the whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive rounds, at which stage the cluster centers have stabilized and will remain the same forever.[6,7]

Error from the data points not in a cluster k can be calculated using the following equation typically, the square-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} |p - m_i|^2 \quad ..(1)$$

where E is the sum of the square error for all objects in the data set; p is the point in space representing a given object; and mi is the mean of cluster c_i (both p and m_i are multidimensional). In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible.

The diagram shown in Figure(2) represents the K-means algorithm.

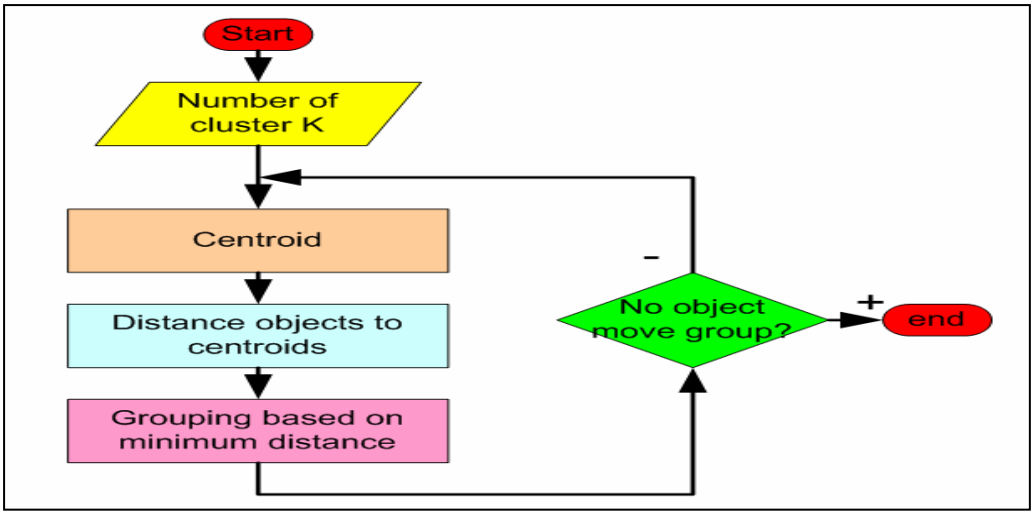

Figure (2): Steps of k-means Algorithm

3.1.1.2. The EM (Expectation Maximization) Algorithm

The EM (Expectation-Maximization) algorithm is a popular iterative refinement algorithm that can be used for finding the parameter estimates. It can be viewed as an extension of the k-means paradigm, which assigns an object to the cluster with which it is most similar, based on the cluster mean . Instead of assigning each object to a dedicated cluster, EM assigns each object to a cluster according to a weight representing the probability of membership. In other words, there are no strict boundaries between clusters. Therefore, new means are computed based on weighted measures.[1]

EM stands for Expectation Maximization is an iterative method for finding maximum likelihood of attributes in a given data set with many attributes.

Another way to present algorithm is given below

1. Initially, randomly assign k cluster centers
2. Iteratively refine the clusters based on two steps
   A. Expectation step: assign each data point X_i to cluster C_i with the following probability

$$P(X_i \in C_k) = P(C_k|X_i) = \frac{P(C_k)P(X_i|C_k)}{P(X_i)} \ldots(2)$$

   B. Maximization step: Estimation of model parameters.[1]

3.2. Curve Fitting Techniques

There are two general approaches for curve fitting that are distinguished from each other on the basis of the amount of error associated with the data. First, where the data exhibits a significant degree of error or "noise," the strategy is to derive a single curve that represents the general trend of the data. Because any individual data point may be incorrect, we make no effort to intersect every point. Rather, the curve is designed to follow the pattern of the points taken as a group. One approach of this nature is called least-squares regression .

Second, where the data is known to be very precise, the basic approach is to fit a curve or a series of curves that pass directly through each of the points. Such data usually originates from tables. For examples the values of the density of water or for the heat capacity of gases as a function of temperature. The estimation of values between well-known discrete points is called interpolation.

3.2.1. Least Squares Method

The method of least squares assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (*least square error*) from a given set of data.

Suppose that the data points are$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where x is the independent variable and y is the dependent variable. The fitting curve $f(x)$ has the deviation (error) d from each data point, i.e.

$$d_1 = y_1 - f(x_1),\ d_2 = y_2 - f(x_2), \ldots, d_n = y_n - f(x_n) \quad \ldots(3)$$

According to the method of least squares, the best fitting curve has the property that:

$$Q = d_1^2 + d_2^2 + \cdots + d_n^2 = \sum_{i=1}^{n}[y_i - f(x_i)]^2 = minimum \quad \ldots(4)$$

The least-squares line method uses a *straight line* $y = a + bx$ to approximate the given set of data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where $n \geq 2$.

The least-squares parabola method uses a *second degree curve* $y = a + bx + cx^2$ to approximate the given set of data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$,

where $n \geq 3$. The best fitting curve $f(x)$ has the least square error, i.e.,

$$Q = \sum_{i=1}^{n}[y_i - f(x_i)]^2 = \sum_{i=1}^{n}[y_i - (a + bx_i + cx_i{}^2)]^2 = minimum \quad ..(5)$$

a, b *and* c are unknown coefficients while all x_i and y_i are given. To obtain the least square error, the unknown coefficients a, b *and* c must yield zero first derivatives.

$$\frac{\partial Q}{\partial a} = -2\sum_{i=1}^{n}[y_i - (a + bx_i + cx_i{}^2)] = 0$$

$$\frac{\partial Q}{\partial b} = -2\sum_{i=1}^{n} x_i[y_i - (a + bx_i + cx_i{}^2)] = 0$$

$$\frac{\partial Q}{\partial c} = -2\sum_{i=1}^{n} x_i^2[y_i - (a + bx_i + cx_i{}^2)] = 0$$

Expanding the above equations, we have

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i{}^2 \quad ..(6)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i{}^2 + c\sum_{i=1}^{n} x_i{}^3 \quad ...(7)$$

$$\sum_{i=1}^{n} x_i^2\, y_i = a\sum_{i=1}^{n} x_i{}^2 + b\sum_{i=1}^{n} x_i{}^3 + c\sum_{i=1}^{n} x_i{}^4 \quad ...(8)$$

Put these into **matrix form**

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i{}^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i{}^2 & \sum_{i=1}^{n} x_i{}^3 \\ \sum_{i=1}^{n} x_i{}^2 & \sum_{i=1}^{n} x_i{}^3 & \sum_{i=1}^{n} x_i{}^4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i{}^2 y_i \end{bmatrix}$$

we have the data points(x_i, y_i) for i=1,2,…,n so we have all the summation terms in the matrix so unknows are *a*,*b* and *c*.

By using Gaussian elimination

$$Let\ A = \begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i{}^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i{}^2 & \sum_{i=1}^{n} x_i{}^3 \\ \sum_{i=1}^{n} x_i{}^2 & \sum_{i=1}^{n} x_i{}^3 & \sum_{i=1}^{n} x_i{}^4 \end{bmatrix},$$

$$B = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i{}^2 y_i \end{bmatrix} \ and\ X = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

so $AX = B$ using built in Mathcad matrix inversion, the coefficients *a*, *b* and *c* are solved

$$\Rightarrow X = A^{-1} * B$$

Then putting the values of *a*, *b* and *c* in equation $y = a + bx + cx^2$ we get the equation of the parabola of best fit. [8]

4. Distribution of Data

The distribution of the collected real data (20000 records) according to the six different attributes is shown in Figure(3)

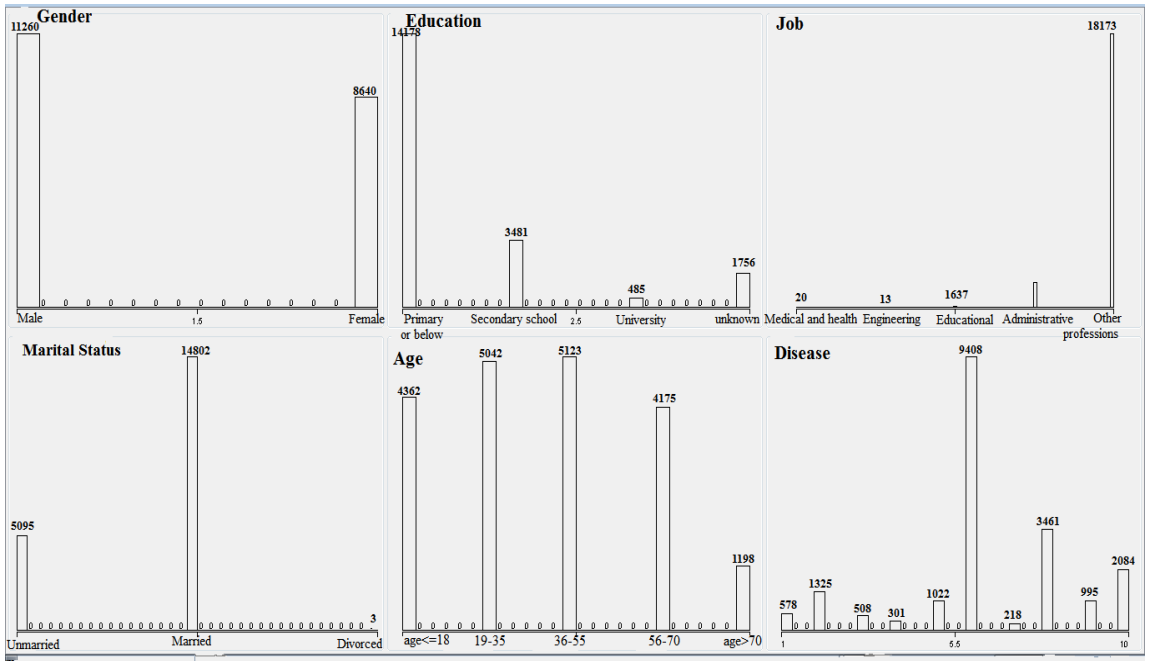

Figure (3): Distribution of the Original Data under Different Attributes

5. Dealing with Missing Values

Table (1) shows a sample of the data set that contain some missing values in patient age attribute. It is required to replace these missing values in order to prepare the data for the analysis process and since good decisions rely mainly on good data analysis reports (knowledge extracted from the data), that in turn require good quality data which means complete, correct and consistent, then missing values data are at the core of this research. Algorithms optimization using entropy and information gain is one of the research goals.

The following algorithms and techniques are the most common in replacing the missing values:

5.1. Most Common Values (Mode)

In some cases, missing values can be replaced by the most common value in the data set, especially when the frequency of the most common value represents high percentage of the given data. In our case for the patient data set given in Table (1), the most common value is shown in Table (2).

5.2. Overall Data set Average

Numeric attributes have more alternatives and algorithms for filling missing values in the data set, one of which is the overall average of the whole data set for the given attribute. This method used when the frequencies of the categories are close to each other . However, in some cases this may lead to results with a high standard deviation from the original. For the data set used in this research, the overall average for the age attribute shown in Table(2).

5.3. Class Most Common Values

As classification represents putting objects together according to predefined specifications, this means that the objects of a specific class have

a lot in common. This characteristic can be used to replace the missing values.

To get better estimation for the missing values in the data set, more efficient algorithms are available, one of which is to classify the data set according to some specific attributes and in our cases, data set is classified according to the patients gender (male and female) and the disease type. The most common values are then found for each class and these values are used to fill in the missing values in the objects under the same class. This method can be used for all data types of the attributes. As in Tables (3), (4),(5) and (6).

5.4. Class Average

When the frequencies of the occurrence of the objects within one class are close to each other, overall class average can be used to replace the missing values .For the numeric values attributes, the class average is an effective method for filling the missing values. Dividing the original table into smaller tables according to some predefined classes will lead to better estimation of the targeted missing values, as shown in Tables (3), (4),(5) and (6).

5.5. Expectation Maximization (EM) Clustering Technique and K-Means Clustering

When the attributes of the data to be classified are not defined prior to the classification process, this is known as clustering. Putting patients into groups: each group has some common characteristics (similarities) between group objects and has dissimilarities with other groups is known as clustering. There are many clustering techniques but in our research we used k-means and EM algorithms and the results in Tables (7) and (8).

Table (2) A Sample for Filling Ages Missing Values Using Most Common Value (Mode) and Data Set Average

Number	Missing value	Most Common age	Overall age average
1	40	60	39
2	27	60	39
3	65	60	39
4	35	60	39
5	39	60	39
6	85	60	39
7	68	60	39
8	41	60	39
9	3	60	39
10	68	60	39
11	64	60	39
12	31	60	39
13	25	60	39
14	60	60	39
15	80	60	39
16	45	60	39
17	45	60	39
18	31	60	39
19	59	60	39
20	63	60	39
21	32	60	39
22	40	60	39
23	40	60	39
24	20	60	39

Table (3) A Sample for Filling Ages Missing Values Using Most Common Value and Data Set Average for Gender Class (Female)

Number	Missing value	Female most common age	Female overall age average
1	25	60	40
2	60	60	40
3	80	60	40
4	45	60	40
5	45	60	40
6	31	60	40
7	59	60	40
8	63	60	40
9	32	60	40
10	40	60	40
11	40	60	40
12	20	60	40
13	45	60	40
14	63	60	40
15	70	60	40
16	70	60	40
17	6	60	40
18	43	60	40
19	26	60	40
20	60	60	40

Table (4) A Sample for Filling Ages Missing Values Using Most Common Value and Data Set Average for Gender Class (Male)

Number	Missing value	Male most common age	Male overall age average
1	2	60	38
2	55	60	38
3	10	60	38
4	1	60	38
5	39	60	38
6	27	60	38
7	43	60	38
8	4	60	38
9	40	60	38
10	19	60	38
11	65	60	38
12	35	60	38
13	39	60	38
14	85	60	38
15	68	60	38
16	41	60	38
17	3	60	38
18	68	60	38
19	64	60	38
20	31	60	38

Table (5) A Sample for Filling Ages Missing Values Using Most Common Value and Data Set Average for (Gender + Disease) Class

Number	Missing value	Gender	Disease	Most Common Age	Class age average
1	40	1	K62	35	35
2	27	1	S07	31	18
3	65	1	K55	45	43
4	35	1	N44	25	22
5	39	1	E10	65	56
6	85	1	I68	70	63
7	68	1	I34	60	59
8	41	1	N18	45	45
9	3	1	J03	6	16
10	68	1	I68	70	63
11	64	1	I68	70	63
12	31	1	S07	30	26
13	25	2	E14	50	51
14	60	2	I50	60	64
15	80	2	H25	60	66
16	45	2	B69	35	32
17	45	2	I83	30	38
18	31	2	J03	7	14
19	59	2	K81	35	41
20	63	2	H25	60	66
21	32	2	I84	35	41
22	40	2	I70	60	58
23	40	2	D36	25	32
24	20	2	V89	70	58

Table (6) A Sample for Filling Ages Missing Values Using Most Common Value and Data Set Average for(Gender + Disease+ Job) Class

Number	Missing value	Gender	Disease	Job	Most Common age for class Job	Age average for class Job
1	40	1	K62	5	40	37
2	27	1	S07	10	30	33
3	65	1	K55	13	65	68
4	35	1	N44	10	30	31
5	39	1	E10	10	50	41
6	85	1	I68	10	70	69
7	68	1	I34	5	60	51
8	41	1	N18	8	45	38
9	3	1	J03	13	5	5
10	68	1	I68	9	70	69
11	64	1	I68	5	55	49
12	31	1	S07	6	40	37
13	25	2	E14	8	50	42
14	60	2	I50	13	60	67
15	80	2	H25	6	60	66
16	45	2	B69	6	35	35
17	45	2	I83	13	30	34
18	31	2	J03	8	25	24
19	59	2	K81	9	50	60
20	63	2	H25	13	60	66
21	32	2	I84	10	35	38
22	40	2	I70	10	35	45
23	40	2	D36	8	25	32
24	20	2	V89	10	30	40

Table (7) A Sample for Filling Ages Missing Values Using Clustering K-means Algorithm

Number	Missing value	Age using 2cluster k-means	Age using 3cluster k-means	Age using 4cluster k-means	Age using 5cluster k-means	Age using 6cluster k-means	Age using 10cluster k-means
1	40	47	47	22	26	26	46
2	27	49	57	52	61	62	48
3	65	47	47	58	61	61	64
4	35	47	47	58	61	61	64
5	39	47	47	22	26	26	28
6	85	47	47	58	61	61	64
7	68	47	47	58	61	61	64
8	41	47	47	22	26	26	28
9	3	47	21	21	20	20	22
10	68	47	47	58	61	61	64
11	64	47	47	58	61	61	64
12	31	47	47	22	20	20	19
13	25	49	57	52	52	62	48
14	60	49	57	52	52	62	67
15	80	49	57	52	52	62	67
16	45	49	57	52	52	62	48
17	45	49	57	52	52	62	67
18	31	49	57	52	52	62	67
19	59	49	57	52	52	62	67
20	63	49	57	52	52	62	67
21	32	49	57	52	52	62	36
22	40	49	57	52	52	62	67
23	40	49	57	52	52	62	48
24	20	47	21	22	26	26	51

Table (8) A Sample for Filling Ages Missing Values Using Clustering EM Algorithm

Number	Missing value	Age using 2cluster EM	Age using 3cluster EM	Age using 4cluster EM	Age using 5cluster EM	Age using 6cluster EM	Age using 10cluster EM
1	40	35	22	22	47	23	18
2	27	55	56	51	53	59	45
3	65	55	58	57	57	50	34
4	35	55	58	57	53	59	59
5	39	35	22	22	47	50	47
6	85	55	58	57	53	59	59
7	68	55	58	57	57	50	34
8	41	35	22	22	57	39	34
9	3	35	22	19	19	18	18
10	68	55	58	57	53	59	59
11	64	55	58	51	47	50	50
12	31	35	22	19	47	18	18
13	25	55	56	51	53	59	45
14	60	55	56	57	53	59	59
15	80	55	56	57	53	59	59
16	45	55	56	51	19	59	45
17	45	55	56	57	53	59	59
18	31	55	56	57	53	59	59
19	59	55	56	57	53	59	59
20	63	55	56	57	53	59	59
21	32	55	56	57	57	39	59
22	40	55	56	57	53	59	59
23	40	55	56	51	19	59	45
24	20	35	22	22	25	23	18

Table (9) A Sample for Filling Ages Missing Values Using Modified Cluster K-means Algorithm by Adding a Constant

Number	Missing Values	Age using 10cluster k-means Algorithm	Age after the addition of constant
1	40	46	31
2	27	48	42
3	65	64	39
4	35	64	68
5	39	28	32
6	85	64	88
7	68	64	68
8	41	28	41
9	3	22	7
10	68	64	68
11	64	64	64
12	31	19	32
13	25	48	23
14	60	67	61
15	80	67	80
16	45	48	42
17	45	67	42
18	31	67	32
19	59	67	61
20	63	67	61
21	32	36	30
22	40	67	42
23	40	48	42
24	20	51	16

Table (10) A Sample for Filling Ages Missing Values Using Modified Cluster EM Algorithm by Adding a Constant

Number	Missing Values	Age using 10cluster EM Algorithm	Age after the addition of constant
1	40	18	22
2	27	45	40
3	65	34	38
4	35	59	63
5	39	47	32
6	85	59	83
7	68	34	66
8	41	34	38
9	3	18	3
10	68	59	63
11	64	50	64
12	31	18	32
13	25	45	30
14	60	59	63
15	80	59	83
16	45	45	45
17	45	59	44
18	31	59	33
19	59	59	59
20	63	59	63
21	32	59	33
22	40	59	44
23	40	45	40
24	20	18	22

5.6. Curve Fitting

The curve fitting is applied for each cluster after clustering, and instead of taking the average or mean of each cluster to find out the missing values, a curve fitting is applied. As in Table (11)

Table (11) A Sample for Filling Ages Missing Values Using Curve Fitting

Number	Missing value	Result EM for cluster 1 and 7				Result K-means for cluster 2 and 5			
		quadratic degree	cubic	4th degree	5th	quadratic	cubic	4th degree	5th degree
1	27	34.522	32.198	34.6487	43.2011871				
2	39	28.28	28.8725	28.4875	28.796875	50.775	52.125	51.2875	52.96875
3	40	27.6272	28.1387	27.03982	27.628405	62.631	62.163	61.9603	56.37861
4	65	27.6272	28.1387	27.03982	27.628405	62.631	62.163	61.9603	56.37861
5	35	27.3872	27.7988	26.42992	28.80592	66.156	65.832	63.7408	52.88064
6	85	27.8192	28.4033	27.57582	27.530295	59.391	59.737	60.0843	59.22039
7	68	27.8192	28.4033	27.57582	27.530295	59.391	59.737	60.0843	59.22039
8	41	27.3872	27.7988	26.42992	28.80592	66.156	65.832	63.7408	52.88064
9	3	27.72	748.7807	27.3	27.5	61.1	61	61.2	58
10	68	27.8192	28.4033	27.57582	27.530295	59.391	59.737	60.0843	59.22039
11	64	27.8192	28.4033	27.57582	27.530295	59.391	59.737	60.0843	59.22039
12	31					37.0998	35.7129	56.3986	85.53678
13	25	28.28	28.8725	28.4875	28.796875	50.775	52.125	51.2875	52.96875
14	60	27.8192	28.4033	27.57582	27.530295	59.391	59.737	60.0843	59.22039
15	80	27.9248	28.5344	27.85312	27.71584	57.504	58.296	58.5328	59.58048
16	45	28.6928	29.0396	28.21072	28.26736	42.444	41.544	42.5248	43.35552
17	45	27.8192	28.4033	27.57582	27.530295	59.391	59.737	60.0843	59.22039
18	31	27.72	748.7807	27.3	27.5	61.1	61	61.2	58
19	59	27.6272	28.1387	27.03982	27.628405	62.631	62.163	61.9603	56.37861
20	63	27.9248	28.5344	27.85312	27.71584	57.504	58.296	58.5328	59.58048
21	32	27.8192	28.4033	27.57582	27.530295	59.391	59.737	60.0843	59.22039
22	40	27.8192	28.4033	27.57582	27.530295	59.391	59.737	60.0843	59.22039
23	40	28.4112	28.9528	28.54272	29.02592	48.176	49.192	48.2928	48.79264
24	20	34.522	32.198	34.6487	40.9803929				

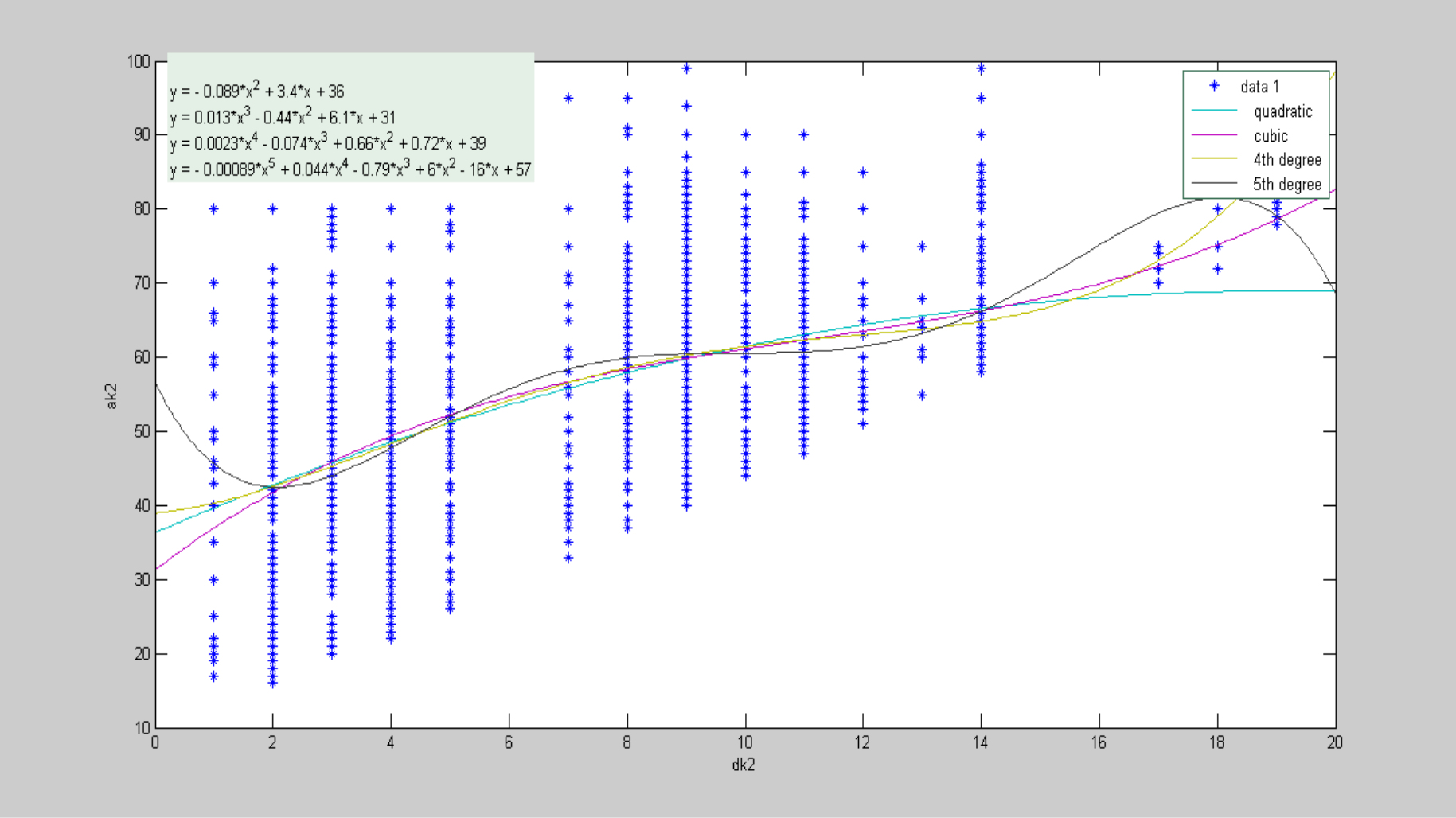

Figure (5): Curve Fitting Using Least Square Method to Cluster 2 of k-means when X-axis Represents the Disease and the Y-axis Represents Age Attribute

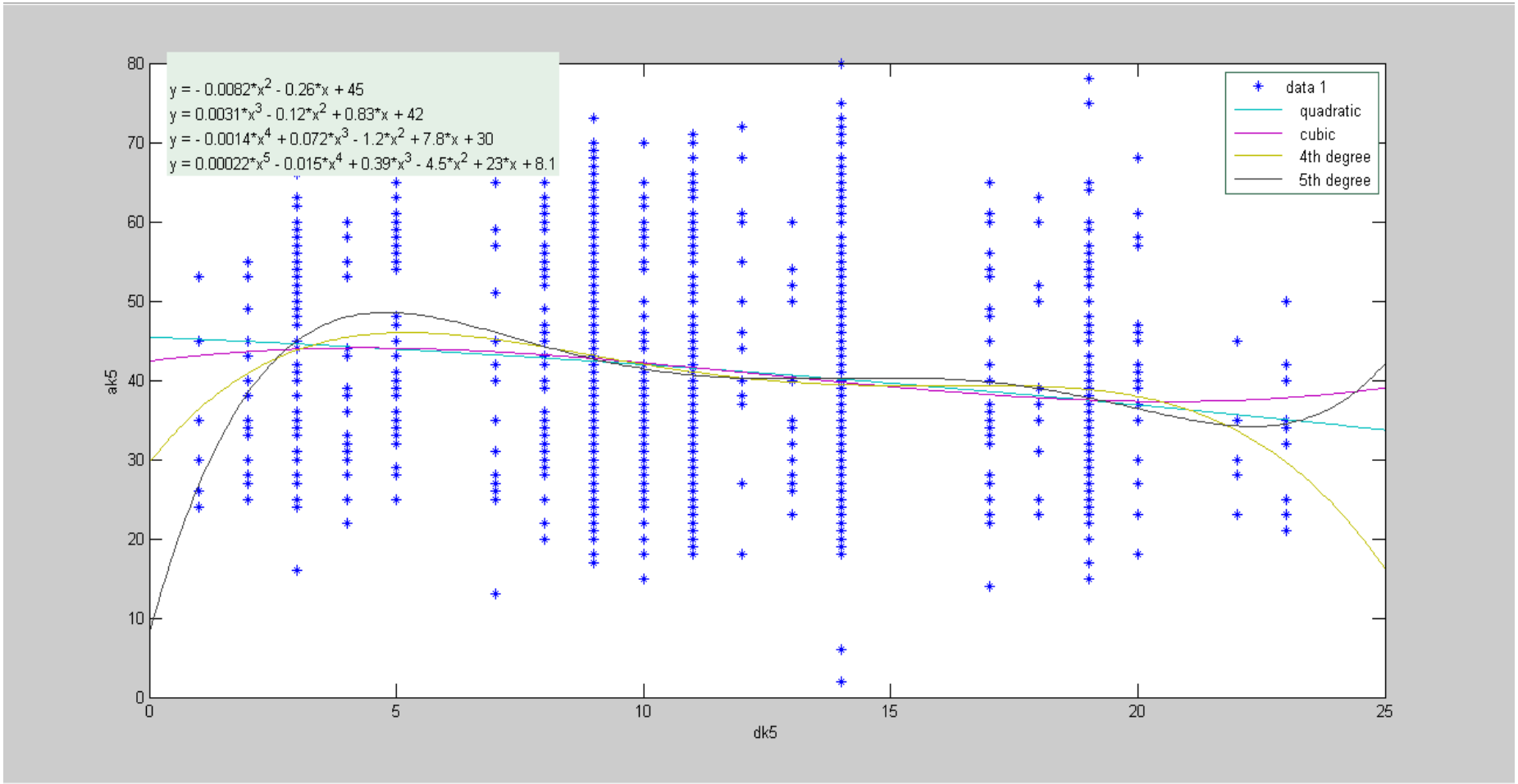

Figure (6): Curve Fitting Using Least Square Method to Cluster 5 of k-means when X-axis Represents the Disease and the Y-axis Represents Age Attribute

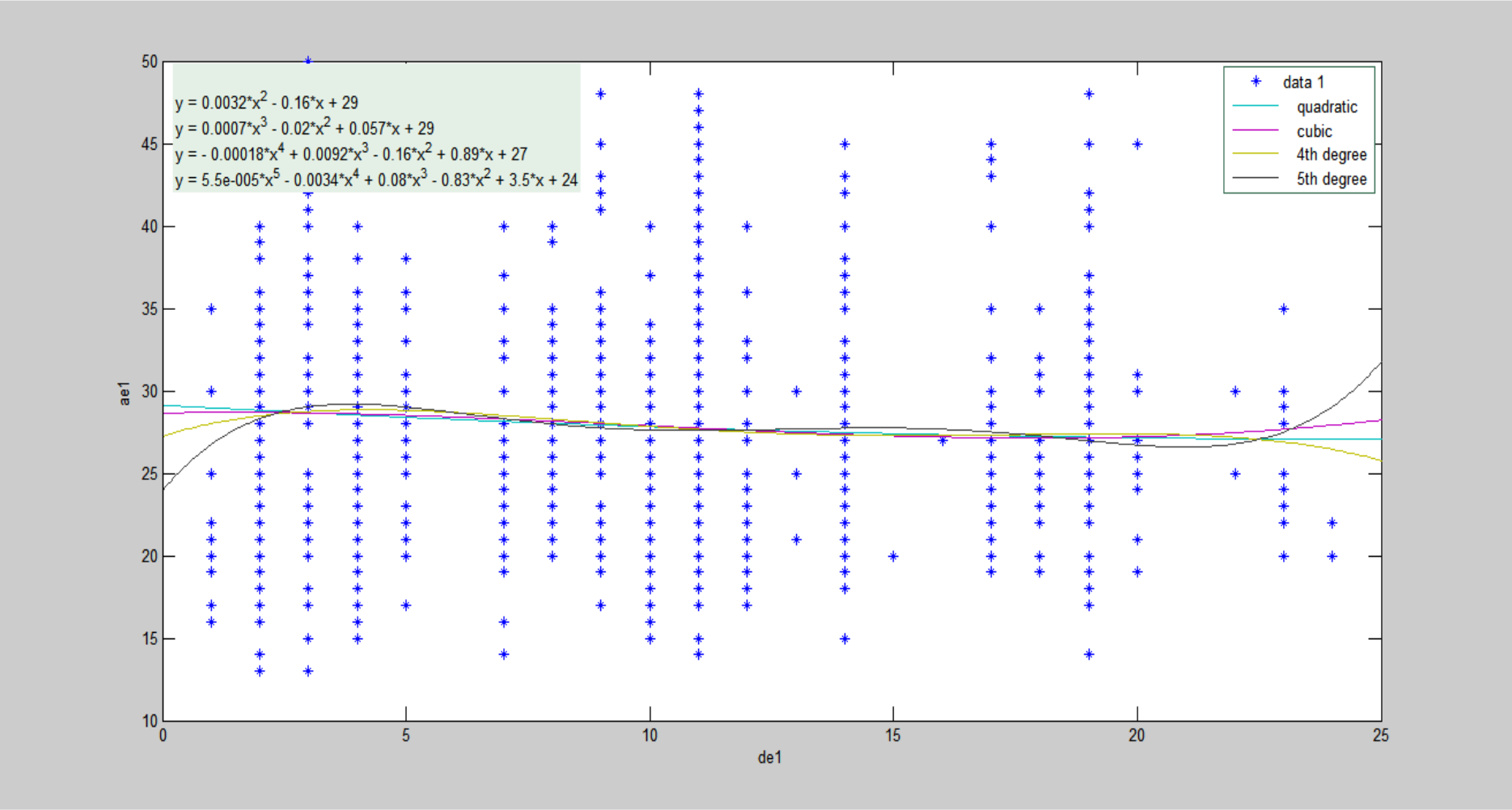

Figure (7): Curve Fitting Using Least Square Method to Cluster 1 of EM when X-axis Represents the Disease and the Y-axis Represents Age Attribute

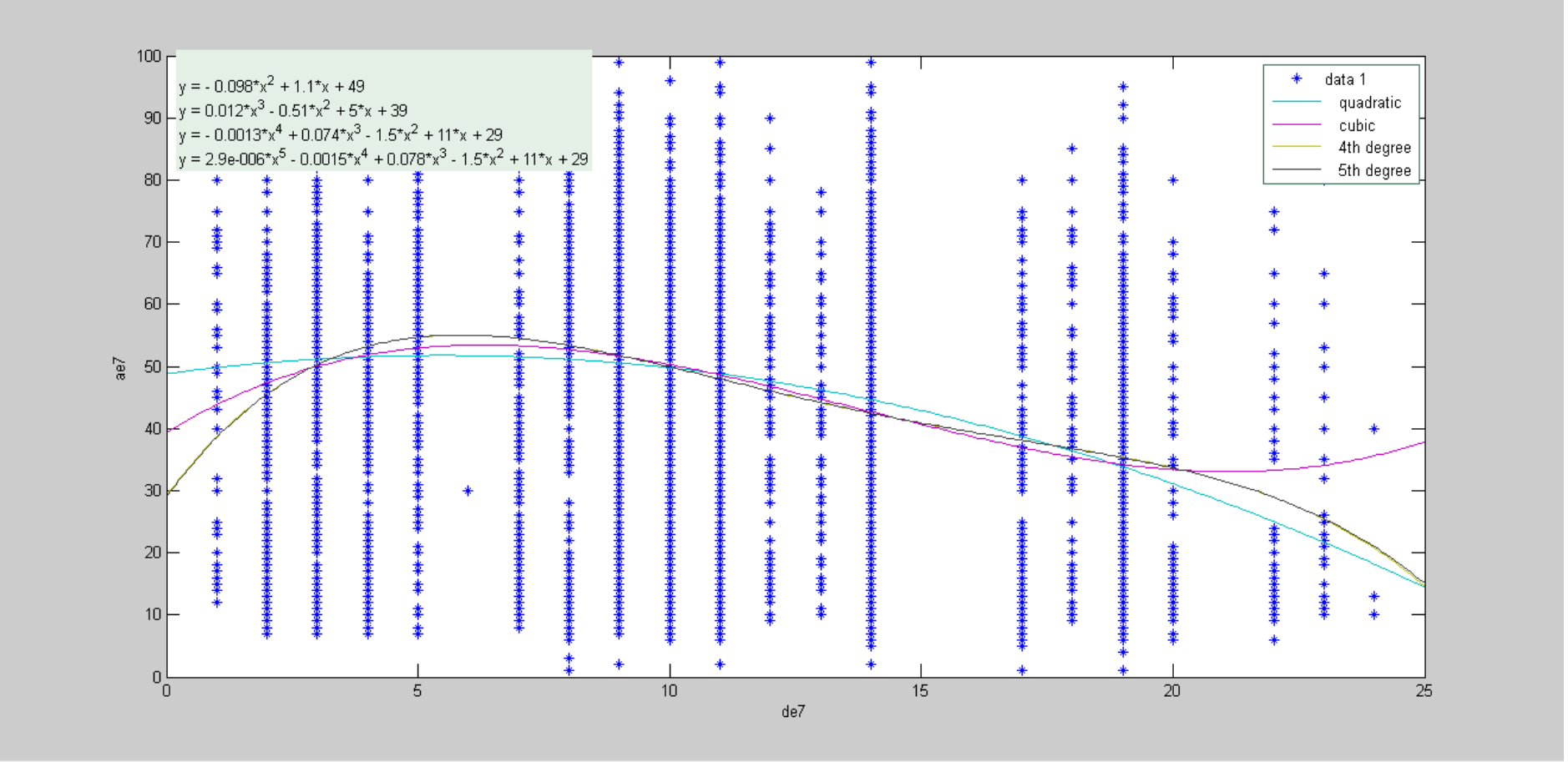

Figure (8): Curve Fitting Using Least Square Method to Cluster 7 of EM when X-axis Represents the Disease and the Y-axis Represents Age Attribute

6. Conclusions

According to the results shown in the tables in chapter three the conclusions of this thesis are:

1. Using the overall average or the most common value (mode) in replacing the missing values noted in the results extreme than reality, when we use the most common value in the whole data set, the suggested missing value will be (60) and the overall average will be (39) but the range of missing age (1-99) as it is shown in the Table (2).
2. Classification technique is much better than the average and most common attribute values because classification techniques put similar objects in one class have much similarities and the results in the Tables(3),(4),(5) and (6)
3. Clustering technique takes into account almost all attributes in grouping the objects, which means that this technique is closer to the reality and the real data.
4. To get better age estimation, it is recommended to add a constant to the age attribute as shown in Tables (9) and (10) .
5. We made a curved connecting data cluster for the function that tells us to compensate for missing values.

   Curve fitting is the way of generalizing a relationship or behavior of different attributes, and instead of replacing the missing values with some fixed and pre-estimated values, the curve fit will give a better estimation along the wide range of attribute values for example the following table shows the result of age attribute using curve fitting technique

Missing age	Curve Fitting	Cluster k-means	Cluster EM
30	34.6487	46	18

For more details see Table(11).

References:

[1] Jiawei Han and Micheline Kamber," Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann, 2006.

[2] Jiawei Han, Micheline Kamber and Jian Pei, " Data Mining: Concepts and Techniques" 3rd Edition, Morgan Kaufmann, 2012.

[3] Wei, W. --- Tang, Y., " A generic neural network approach for filling missing data in data mining", Systems, Man and

Cybernetics, 2003, IEEE International Conference on ISBN: 0780379527.

[4] Luai Al Shalabi, Mohannad Najjar and Ahmad Al Kayed, "A framework to Deal with Missing Data in Data Sets",Journal of Computer Science 2 (9): 740-745, 2006, ISSN.

[5] Kadhim B. Swadi Aljanabi, " An Improved Algorithm for Data Preprocessing in Mining Crime Data Set", Journal of Kufa for Mathematics and Computer, V o l . 1, N o . 4, N o v ., 2 011, pp.81- 87.

[6] David Hand, Heikki Mannila and Padhraic Smyth," Principles of Data Mining", MIT Press, 2001.

[7] Dr. Bushra M. Hussan, " Data Mining based Prediction of Medical data Using K-means algorithm", Basrah Journal of Science (A), Vol.30(1), 46-56 2012.

[8] Joe D. Hoffman, "Numerical Methods for Engineers and Scientists", Second Edition,, NEW YORK. BASEL, 1992, ISBN: 0-8247-0443-6.