Tribhuvan University

Institute of Science and Technology

Comparative Study of K-means, Expectation-Maximization and Density Based Clustering Algorithm

A Dissertation
Submitted to

Central Department of Computer Science and Information Technology

Tribhuvan University, Kirtipur , Kathmandu, Nepal

In partial fulfillment of the requirements

For the Master's Degree in Computer Science and Information Technology

Submitted by

Deepa Upadhaya

July, 2018

Tribhuvan University
Institute of Science and Technology

Comparative Study of K-means, Expectation-Maximization and Density Based Clustering Algorithm

Dissertation
Submitted to

Central Department of Computer Science and Information Technology
Kirtipur, Kathmandu, Nepal

In partial fulfillment of the requirements
for the Master's Degree in Computer Science and Information Technology

by
Deepa Upadhaya
Date :July,2018

Supervisor
Mr. Bikash Balami

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

Student's Declaration

I hereby declare that I am the only author of this work and that no sources other than the listed here have been used in this work.

.....

Deepa Upadhaya

Date: 24 July, 2018

Tribhuvan University
Institute of Science and Technology
Central Department of Computer Science and Information Technology

Supervisor's Recommendation

I hereby recommend that this dissertation prepared under my supervision by **Ms. Deepa Upadhaya** entitled **"Comparative Study of K-means, Expectation-Maximization and Density Based Clustering Algorithm"** in partial fulfillment of the requirements for the degree of M. Sc. in Computer Science and Information Technology be processed for the evaluation.

……...

Mr. Bikash Balami
Central Department of Computer Science and Information Technology
Tribhuvan University,

Kirtipur, Kathmandu, Nepal

Date: 24 July, 2018

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

LETTER OF APPROVAL

We certify that we have read this dissertation and in our opinion it is satisfactory in the scope and quality as a dissertation in the partial fulfillment for the requirements of Masters Degree in Computer Science and Information Technology.

Evaluation Committee

……	……
Mr.Lochan Lal Amatya Former Chief Technology Officer Nepal Telecom	**Mr. Ram Krishna Dahal** Center Department of Computer Kritipur , Nepal
……	……..
(External Examiner)	(**Internal Examiner**)

Date: 24 July, 2018

Acknowledgement

Foremost, I would like to express my sincere gratitude to my supervisor **Mr. Bikash Balami**.I have been amazingly fortunate to have a supervisor who gave me the freedom to explore on my own, and at the same time the guidance to recover when my steps faltered. His patience and support helped me overcome many crisis situations and finish this dissertation.

I would like to thank the research committee for their encouragement, insightful comments, and hard questions. I am indebted to all the people who supported and encouraged me involving directly or indirectly to complete this work. I am also obliged to **Head of Department, Asst. Prof. Nawaraj Paudel** and all respected teachers and staffs of Central Department of Computer Science and Information Technology, Tribhuvan University for their cooperation to bring this work in a tangible form.

Last but not the least , I would like to thank my family for their love and supporting me spiritually throughout my life.

Abstract

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. This dissertation entitled **—"Comparative Study of K-means, Expectation-maximization and density Based Clustering Algorithm"** is one of the implementation of Data Mining in which the datasets of "Heart Disease and Thyroid Disease Data Set" are used. There is a wide range of algorithms available for clustering. This research presents a comparative study of clustering algorithms. In experiments, the accuracy and time taken by algorithms is evaluated by comparing the results on heart disease and thyroid disease datasets , which is obtained from the UCI and KEEL repository using WEKA tool.

All total 597 data of heart disease datasets and 3772 data of Thyroid disease datasets are use for implementing the algorithm. Heart disease use 14 attributes and thyroid disease use 30 attributes.

Expectation-maximization clustering and Density based clustering takes more time to form clusters for both datasets (heart disease and thyroid disease datasets).Simple K-means clustering algorithms forms clusters with less time and more accuracy than other algorithms for heart disease and thyroid disease datasets . In terms of time and accuracy K-means produces better results as compared to other algorithms.

Keywords: Clustering, K-means algorithm, Expectation and maximization algorithm and Density based algorithm and WEKA tool.

Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviations	Full Form
MAP	Maximum a Posteriori
EM	Expectation-maximization
WEKA	Waikato Environment for Knowledge Analysis
DBSCAN	Density Based Spatial Clustering of applications with noise
OPTICS	Ordering Points To Identify Clustering Structure
LDBSCAN	Local Density Based Spatial Clustering Algorithm with Noise
GDBSCAN	Generalizing Density Based Spatial Clustering Algorithm with Noise
DBCLASD	Distribution Based Clustering of Large Spatial Databases
ARFF	Attribute-Relation File Format
CSV	Comma Separated Values
GUI	Graphical User Interface
TSH	Thyroid-stimulating hormone
TT4	Total Thyroxine
FTI	Free thyroxine index
TBG	Thyroxine-binding globulin

CHAPTER 1

1.1 INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases[5,13].

Clustering is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters**.** Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.Clustering is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) [3].

A clustering algorithm partitions a data set into several groups based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. Clustering techniques have numerous applications in various fields, including artificial intelligence, pattern recognition, bioinformatics, segmentation and machine learning [12].

This research performs a comparative study of three clustering algorithm names K-means algorithm, Expectation-maximization algorithm and Density based algorithm.

1.2 Thesis Organization

Introduction Part of this dissertation work focuses on the Data Mining and clustering algorithm along with the main processes of this work.
The rest of the material in this study is organized into five subsequent chapters.
Chapter 2 provides the background study required for this work.
Chapter 3 contains the previous work related to this dissertation in detail under literature review and research question is formulated.
Chapter 4 provides research methodology of clustering algorithm with collected datasets in heart disease and thyroid disease using weka tool.
Chapter 5 provides the performance measure of the system with different values of initial clusters seeds with table as well as graph.

At last, the concluding remarks are outlined in chapter 6.

CHAPTER 2

BACKGROUND STUDY AND PROBLEM FORMULATION

2.1 Data Mining

It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner[6,7]. This helps businesses be closer to their objective and make better decisions. Data mining involves effective data collection and warehousing as well as computer processing

2.1.1 Clustering

The process of grouping a set of objects into classes, subsets or clusters of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. The idea of clustering is to partition a free set of objects into clusters. Among many mechanisms of text mining, clustering is one of the most important techniques. There are number of clustering algorithms which are used in different areas or fields[10,11].

2.1.1.1 Algorithms

2.1.1.1.1 K-Means Clustering Algorithms

The k-means clustering algorithm is known to be efficient in clustering large data sets. This clustering algorithm was developed by MacQueen , and is one of the simplest and the best known unsupervised learning algorithms that solve the well-known clustering problem. The K-Means algorithm aims to partition a set of objects, based on their attributes features, into *k* clusters, where *k* is a predefined or user-defined constant. The main idea is to define *k* centroids, one for each cluster. The centroid of a cluster is formed in such a way that it is closely related to all objects in that cluster [16].

Algorithm:

1. Choose k number of clusters to be determined
2. Choose k objects randomly as the initial cluster center
3. Repeat
   3.1. Assign each object to their closest cluster
   3.2. Compute new clusters, i.e. Calculate mean points.
4. Until
   4.1. No changes on cluster centers (i.e. Centroids do not change location any more) OR
   4.2. No object changes its cluster (We may define stopping criteria as well)

2.1.1.1.2 Expectation–Maximization Algorithm

This is an iterative method for finding maximum likelihood or Maximum a Posteriori (MAP) estimates of parameters in statistical systems, where the system depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the expectation step. These parameters-estimates are then used to determine the distribution of the latent variables in the next expectation steps [14,15].EM Algorithm focus on the joint log-likelihood function of the observed variables X and the latent variables Z = { . . . },

$l_\theta (X, Z) = \ln p_\theta (X, Z)$.

Algorithm:

Step1. Initialize: Set i = 1 and choose an initial θ_1.

Step2. Repeat :(a) Expectation (E): Compute

$Q(\theta, \theta_i) = E_{\theta i} [\ln p_\theta (Z, X \mid X)]$
$= \int \ln p_\theta (Z, X)\, p_{\theta i} (Z \mid X) dZ.$

(b) Maximization (M): Compute

$\theta_i + 1 = \arg \max Q(\theta, \theta_i)$.

(c) i ← i + 1

2.1.1.1.3 Density Based Algorithm

A cluster is a dense region of points that is separated by low density regions from the tightly dense regions. This clustering algorithm can be used when the clusters are irregular. It finds core objects, i.e. objects that have dense neighborhoods. It connects core objects and their neighborhoods to form dense regions as clusters. Clusters are formed as maximum sets of density, connected points and can detect noise and used when outliers are encountered[4,8].

Algorithm :

Step1: Select an arbitrary point r.

Step2: Retrieve the neighborhood of r using 'ε'.

Step3: If the density of the neighborhood reaches to the threshold, clustering process start.

Else point is marked as noise.

Step4: Repeat the process until all of the points have been processed.

2.2 Problem Statement

Clustering techniques do not deal with all the requirements adequately and concurrently. Dealing with a large number of sizes and large number of data items can be problematic because of time complexity. Hence, choosing an algorithm for a particular type of data set is a difficult problem. This study presents the choice of an appropriate clustering algorithm by comparative study of clustering algorithms. If the intial clusters are not properly chosen, then after a few iterations it is found that clusters may even be left empty.There are a number of problems with clustering.The result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

2.3. Objective

The objective of this research is to perform a comparative study of clustering algorithms, namely K-means algorithm, Expectation-maximization algorithm and Density based algorithm. The output of cluster analysis is the number of clusters that form the structure of partitions of the data set. In short clustering is the technique to process the data into meaningful group for statistical analysis.The algorithms are compared in terms of time taken and accuracy in “Heart Disease” and “Thyroid Disease”data sets, using WEKA tool.

CHAPTER 3

LITERATURE REVIEW

3.1 Literature Review

In this paper work, clustering algorithm and their accuracy have been compared. In this section we have presented a review on various techniques. K-mean is the most popular partitioning method of clustering. James Mac Queen in 1967, firstly proposed this technique, though the idea goes back to Hugo Steinhaus in 1957[1,2].

It results that the K-means algorithm gives overall best results. Osama Abu Abbas has discussed k-means of clustering of objects belongs among a k-group. It is used on small and huge datasets both. The EM algorithm was explained and given its name in a classic 1977 paper by Arthur Dempster, Nan Laird, and Donald Rubin.They pointed out that the method had been "proposed many times in special circumstances" by earlier authors. A very detailed treatment of the EM method for exponential families was published by Rolf [5]. Jan Carlo Barca et al, presented a modified K-Means algorithm used to remove noise from motion capture images with the Illuminated Line Segment based Markers and classical ping-pong ball style markers. The modifications to the classical K-Means algorithm were in the form of constraints on cluster size and cluster compactness. The value for the cluster size constraint was set just above the number of data points usually found in a noise cluster for the type of data at hand. The value for the cluster compactness constraint was set just below the minimum compactness of valid data clusters. The modified K-means algorithm manage to completely remove Gaussian blur noise with varying radii, with a gradual increase of false positives as the blur radius increases. This was a better result than that produced by traditional median and mean filters[16].

The Dempster–Laird–Rubin paper in 1977 generalized the method and sketched a convergence analysis for a wider class of problems. Regardless of earlier inventions, the innovative Dempster–Laird–Rubin paper in the Journal of the Royal Statistical Society received an enthusiastic discussion at the Royal Statistical Society meeting with Sundberg calling the paper "brilliant". The Dempster–Laird–Rubin paper established the EM method as an important tool for statistical analysis.The convergence analysis of the Dempster–Laird–Rubin paper was defective and correct convergence analysis was

published by C. F. Jeff Wu in 1983. Wu's proof established the EM method's convergence outside of the exponential family, as claimed by Dempster–Laird–Rubin[15].

Density based clustering approaches are first presented nearly a one and half decades ago[4]. The density connectivity notion established by M. Ester. is served as a pioneering one and led further to many density based clustering algorithms such as DBSCAN , OPTICS (Ordering Points To Identify Clustering Structure) , LDBSCAN (Local Density Based Spatial Clustering Algorithm with Noise) , GDBSCAN (Generalizing DBSCAN) , DBCLASD (Distribution Based Clustering of Large Spatial Databases) etc. Many improvements are suggested in future research of these approaches, aiming to generalize the clustering approach such as - different types of databases as in multimedia databases , addressing the problem of clustering real time stream data as in , handling the local density problems, clustering on the time series gene expression data as in introduction of a speciality in a particular task as in moving clusters[4,9] .

CHAPTER 4

RESEARCH METHODOLOGY

4.1 Dataset

For performing the comparison analysis “Heart disease” and “Thyroid disease” datasets has been used,which is obtained from the site http: //kdd.icu.uci.edu or from the another site http://www.kdnuggets.com/datasets. All total 597 data of heart disease datasets and 3772 data of Thyroid disease datasets are uses for implementing the algorithm.Heart disease use 14 attributes and thyroid disease use 30 attributes.These are the good datasest to test time series clustering to achieve perfect accuracy. Few of the datasets are as follows:

No.	Age	Sex	Cp	Trestbps	chol	Fbs	Restecg	Thalach	Exang	oldpeak	slope	ca	thal	Num
1	63	male	typ_angina	145	233	T	left_vent_hyper	150	No	2.3	down	0	fixed_defect	<50
2	67	male	Asympt	160	286	F	left_vent_hyper	108	Yes	1.5	flat	3	normal	>50_1
3	67	male	Asympt	120	229	F	left_vent_hyper	129	Yes	2.6	flat	2	reversable_defect	>50_1
4	37	male	non_anginal	130	250	F	Normal	187	No	3.5	down	0	normal	<50
5	41	female	atyp_angina	130	204	F	left_vent_hyper	172	No	1.4	up	0	normal	<50
6	56	male	atyp_angina	120	236	F	Normal	178	No	0.8	up	0	normal	<50
7	62	female	asympt	140	268	F	left_vent_hyper	160	No	3.6	down	2	normal	>50_1
8	57	female	asympt	120	354	F	Normal	163	Yes	0.6	up	0	normal	<50
9	63	male	Asympt	130	254	F	left_vent_hyper	147	Ti	1.4	flat	1	reversable_defect	>50_1
10	53	male	Asympt	140	203	T	left_vent_hyper	155	"{	3.1	down	0	reversable_defect	>50_1
11	57	male	Asympt	140	192	F	Normal	148	No	0.4	flat	0	fixed_defect	<50
12	56	female	atyp_angina	140	294	F	left_vent_hyper	153	No	1.3	flat	0	normal	<50
13	56	male	non_anginal	130	256	T	left_vent_hyper	142	Yes	0.6	flat	1	fixed_defect	>50_1
14	44	male	atyp_angina	120	263	F	Normal	173	No	0	up	0	reversable_defect	<50
15	52	male	non_anginal	172	199	T	Normal	162	No	0.5	up	0	reversable_defect	<50

Table 4.1.1 List Of Heart Disease Datasets

No.	sex	'on thyroxine'	'query on thyroxine'	'on antithyroid medication'	sick	pregnant	'thyroid surgery'	'I131 treatment'	'query hypothyroid'
1	F	f	f	f	f	f	f	f	F
2	F	f	f	f	f	f	f	f	F
3	M	f	f	f	f	f	f	f	F
4	F	t	f	f	f	f	f	f	F
5	F	f	f	f	f	f	f	f	F
6	F	t	F	F	f	f	f	f	F
7	F	f	F	F	f	f	f	f	F
8	F	f	F	F	f	f	f	f	F
9	F	f	F	F	f	f	f	f	F

Table 4.1.2 List Of Thyroid Disease Datasets

No.	'query hyperthyroid'	lithium	goitre	tumor	hypopituitary	psych	'TSH measured'	TSH	'T3 measured'	T3
1	f	f	f	F	f	f	t	1.3	t	2.5
2	f	f	f	F	f	f	t	4.1	t	2
3	f	f	f	F	f	f	t	0.98	f	?
4	f	f	f	F	f	f	t	0.16	t	1.9
5	f	f	f	F	f	f	t	0.72	t	1.2
6	f	f	f	F	f	f	t	0.03	f	?
7	f	f	f	F	f	f	f	?	f	?
8	f	f	f	F	f	f	t	2.2	t	0.6
9	f	f	f	T	f	f	t	0.6	t	2.2

Table 4.1.3 List Of Thyroid Disease Datasets

No.	'TT4 measured'	TT4	'T4U measured'	T4U	'FTI measured'	FTI	'TBG measured'	TBG	'referral source'	Class
1	t	125	t	1.14	t	109	f	?	SVHC	negative
2	t	102	f	?	f	?	f	?	other	negative
3	t	109	t	0.91	t	120	f	?	other	negative
4	t	175	f	?	f	?	f	?	other	negative
5	t	61	t	0.87	t	70	f	?	SVI	negative
6	t	183	t	1.3	t	141	f	?	other	negative
7	t	72	t	0.92	t	78	f	?	other	negative
8	t	80	t	0.7	t	115	f	?	SVI	negative
9	t	123	t	0.93	t	132	f	?	SVI	negative

Table 4.1.4 List Of Thyroid Disease Datasets

4.1.1 Description of Attributes:

Here are the attributes that are used in data set for heart disease.

S.no.	Parameters	Parameters Description	Values
1	Age	Age in years	Continuous
2	Sex	Male or female	1= male
			0= female
3	Thestbps	Resting blood pressure	Continuous value in mmHg ,value-94-200
4	Cp	Chest pain type	1= typical type 1
			2= typical type angina
			3= non-angina pain
			4= asymptomatic
5	Chol	Serum cholesterol	Continuous value
			in mm/dL
6	Fbs	Fasting blood sugar	1≥120 mg/dL
			0≤120 mg/dL
7	Restecg	Resting electrographic results	0= normal
			1= having ST-T wave abnormal
			2= left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	Continuous value-82-188
9	old peak	ST depression induced by exercise relative to rest	Continuous value
10	Exang	Exercise induced angina	0= no
			1= yes
11	Ca	Number of major vessels colored by fluoroscopy	0–3 value
12	Slope	Slope of the peak	1= unsloping
		exercise ST segment	2= flat
			3= downsloping
13	Thal	Defect type	3= normal
			6= fixed
			7= reversible defect
14	Num	Diagnosis of heart disease	0%≤50%
			1%.50%

Table 4.1.1.1 Description of Attributes used in Heart Disease Datasets

S.N0.	Attributes	Value
1	Age	Continuous
2	Sex	Male,Female
3	'on thyroxine'	False,True
4	'query on thyroxine'	False,True
5	'on antithyroid medication'	False,True
6	Sick	False,True
7	Pregnant	False,True
8	'thyroid surgery'	False,True
9	'I131 treatment'	False,True
10	'query hypothyroid'	False,True
11	'query hyperthyroid'	False,True
12	Lithium	False,True
13	Goiter	False,True
14	Tumor	False,True
15	hypopituitary	False,True
16	Psych	False,True
17	'TSH measured'	False,True
18	TSH	Continuous
19	'T3 measured'	False,True
20	T3	Continuous
21	'TT4 measured'	False,True
22	TT4	Continuous
23	'T4U measured'	False,True
24	T4U	Continuous
25	'FTI measured'	False,True
26	FTI	Continuous
27	'TBG measured'	False,True
28	TBG	Continuous
29	'referral source'	SVHM,Other,SVI,STMW,SVHD
30	Class	negative,compensated_hypothyroid Primary_hypothyroid Secondary_hypothyroid

Table 4.1.1.2 Description of Attributes used in Thyroid Disease Datasets

4.2 Tool

WEKA is a software tool that was developed at the University of Waikato in New Zealand and written on Java . WEKA is platform-independent, open source and user friendly with a graphical interface that allows for quick set up and operation, WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to the dataset or called from your own Java code. WEKA contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization.

WEKA tool contains Attribute-relationship file format (.arff and .csv)file of the data set. Data set consists of attribute names, types, values and the data. In WEKA, the data objects are called as instances and features of data are considered as attributes.

CHAPTER 5

ANALYSIS AND RESULT

5.1 Analysis

Here are some snapshot of output of k-means,expectation-maximization and density based clustering algorithms using different value of clustering and seed.

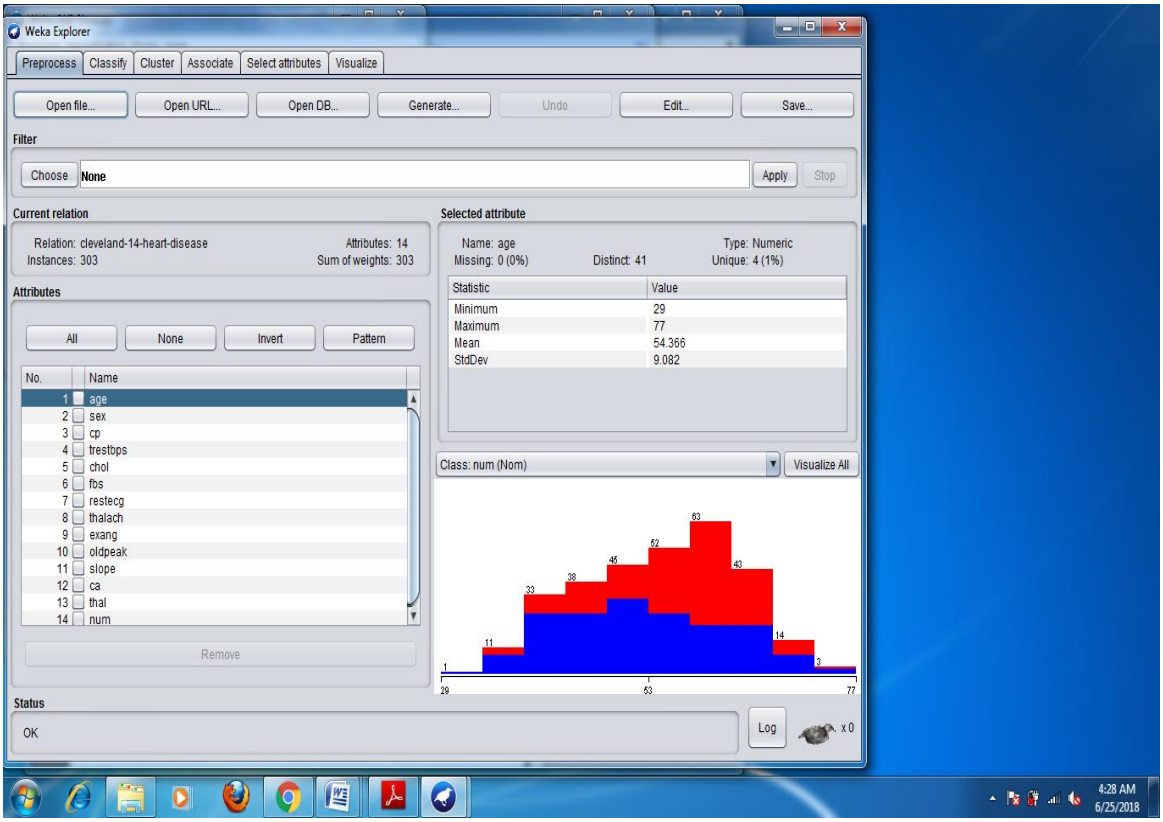

Figure 5.1.1 Screenshot of output after loading dataset.

5.1.1 Sample Output

Run Information of K-means Clustering Algorithm By Default

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Instances: 597

Attributes:

14,age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,num

Number of iterations: 9

Within cluster sum of squared errors: 1259.8346964261154

Initial starting points (random):

Cluster 0:

51,female,asympt,130,305,f,normal,142,yes,1.2,flat,0,reversable_defect,>50_1

Cluster 1:

54,male,non_anginal,120,237,f,normal,150,yes,1.5,flat,0.667774,reversable_defect,>50_1

Final Cluster Centroids:

	Clusters		
Attribute	**Full Data**	0	**1**
	597	354	243
Age	51.1457	49.3927	53.6996
Sex	Male	male	male
Cp	Asympt	atyp_angina	asympt
Trestbps	132.0956	130.1754	134.893
Chol	248.4286	241.55	258.4491
Fbs	F	f	F
Restecg	Normal	normal	Normal
Thalach	144.4765	152.7697	132.3951
Exang	No	no	Yes
Oldpeak	0.8162	0.3636	1.4757
Slope	Flat	flat	Flat
Ca	0.6678	0.494	0.9209
Thal	Normal	normal	Normal
Num	<50	<50	>50_1

Table 5.1.1.1 Final cluster centroids

Time taken to build model (full training data) **: 0.04 seconds**

Clustered Instances

0 354 (59%)

1 243 (41%)

Run Information of Expectation-maximization Clustering Algorithm By Default

Number of clusters selected by cross validation: 4

Number of iterations performed: 5

Final Cluster Centroids:

	Clusters			
Attribute	0	1	2	3
	0.21	0.18	0.41	0.19
Age				
mean	53.9956	50.7591	46.6711	57.823
std. dev.	8.9591	7.3757	7.9702	7.4142
Sex				
male	64.6451	91.9281	168.2616	99.1652
Female	64.6875	19.5586	78.1425	18.6114
[total]	129.3325	111.4867	246.404	117.7767
Cp				
typ_angina	15.9102	1.6085	11.2637	9.2177
asympt	28.3081	95.1918	58.8707	87.6294
non_anginal	52.4792	8.4285	65.4039	18.6884
atyp_angina	34.6351	8.2579	112.8658	4.2412
[total]	131.3325	113.4867	248.404	119.7767
trestbps				
mean	130.9914	136.7652	128.9762	135.4793
std. dev.	15.6546	19.0761	16.5371	18.4989
chol				
mean	249.0191	271.3623	236.786	250.6688
std. dev.	58.8656	78.7028	48.9627	46.8008
fbs				
t	18.236	11.576	14.6517	24.5362
f	111.0965	99.9107	231.7523	93.2404
[total]	129.3325	111.4867	246.404	117.7767
restecg				
left_vent_hyper	49.5873	5.0523	29.8211	72.5393
normal	74.8608	83.8102	189.3407	43.9883
st_t_wave_abnormality	5.8844	23.6243	28.2423	2.2491

[total]	130.3325	112.4867	247.404	118.7767
thalach				
mean	154.3966	123.7302	151.7121	137.9112
std. dev.	18.5999	18.7823	22.5321	21.4385
exang				
no	108.6165	18.3576	234.3664	51.6595
yes	20.7161	93.1291	12.0376	66.1172
[total]	129.3325	111.4867	246.404	117.7767
oldpeak				
mean	0.8883	1.5173	0	1.7972
std. dev.	0.6962	0.9344	0.0005	1.2935
slope				
down	8.026	2.2062	1.9342	13.8336
flat	49.726	109.0055	187.3884	78.8801
up	72.5805	1.275	58.0815	26.0629
[total]	130.3325	112.4867	247.404	118.7767
ca				
mean	0.3507	0.6449	0.5354	1.3175
std. dev.	0.5482	0.2055	0.2952	1.07
thal				
fixed_defect	3.2016	6.1316	8.951	13.7158
normal	105.2944	91.5552	223.1346	25.0157
reversable_defect	21.8366	14.7999	15.3184	80.0451
[total]	130.3325	112.4867	247.404	118.777
num				
<50	117.957	14.9793	212.6921	11.3717
>50_1	11.3756	96.5075	33.712	106.405
[total]	129.3325	111.4867	246.404	117.777

Table 5.1.1.2 Final cluster centroids

Time taken to build model (full training data) : **6.57 seconds**

Clustered Instances

0 111 (19%) 1. 90 (15%)

2 277 (46%) 3. 119 (20%)

Log likelihood: -21.6806

Run Information of Density-Based Clustering Algorithm By Default

Number of iterations: 9

Within cluster sum of squared errors: 1259.8346964261154

Initial starting points (random):

Cluster 0:

51,female,asympt,130,305,f,normal,142,yes,1.2,flat,0,reversable_defect,>50_1

Cluster 1:

54,male,non_anginal,120,237,f,normal,150,yes,1.5,flat,0.667774,reversable_defect,>50_1

Final Cluster centroids:

	Clusters		
Attribute	**Full Data**	0	**1**
	597	**354**	**243**
age	51.1457	49.3927	53.6996
sex	male	male	Male
cp	asympt	atyp_angina	Asympt
trestbps	132.0956	130.1754	134.893
chol	248.4286	241.55	258.4491
fbs	f	F	F
restecg	normal	normal	Normal
thalach	144.4765	152.7697	132.3951
exang	no	no	Yes
oldpeak	0.8162	0.3636	1.4757
slope	flat	flat	Flat
ca	0.6678	0.494	0.9209
thal	normal	normal	Normal
num	<50	<50	>50_1

Table 5.1.1.3 Final cluster centroids

Time taken to build model (full training data) : **0.13 seconds**

Clustered Instances

0 360 (60%) 1 237 (40%)

Log likelihood: -25.09342

5.2 Experimental Result

Having introduced the clustering algorithms, now turn to the discussion of these algorithms on the basis of a practical study. This section presents the experimental result of each of the three clustering algorithms using heart disease datasets. The experimental results are presented in the table below:

Column1	Time taken by K-means	Time taken by EM	Time taken by DBSCAN
Default	0.04	6.57	0.13
N=4,Seed=20	0.06	0.26	0.05
N=4,Seed=50	0.05	0.25	0.05
N=6,Seed=30	0.06	0.35	0.08
N=6,Seed=50	0.15	0.31	0.06

Table 5.2.1. Time taken by K-Means ,EM and DBSCAN Clustering Algorithm

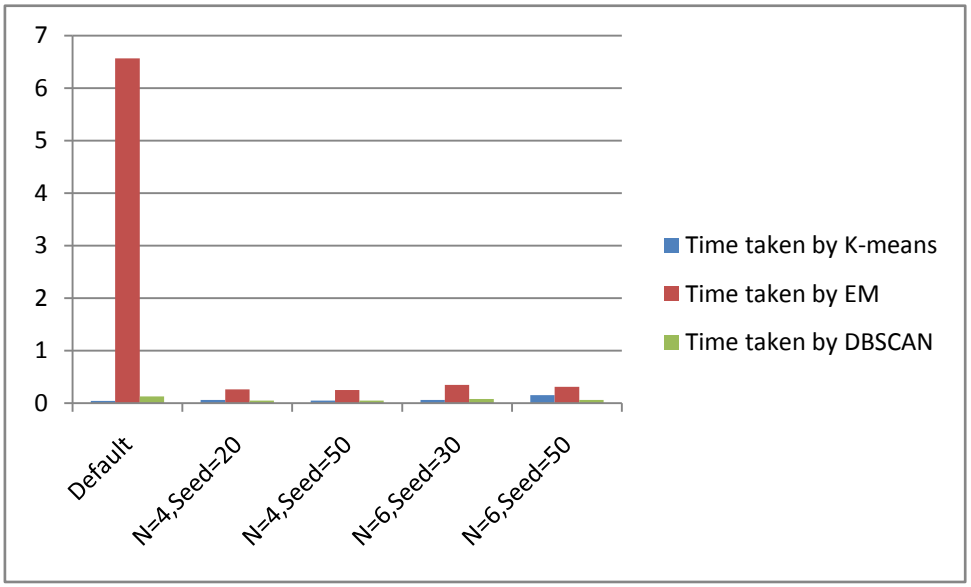

Figure 5.2.1. Graph of Time taken with different values of N and Seed for K-means ,EM and DBSCAN Clustering Algorithm

Column1	Log likelihood by DBSCAN	Log likelihood by EM
Default	-25.09342	-21.6806
N=4,Seed=20	-24.50013	-23.14797
N=6,Seed=30	-24.73684	-22.79447
N=4,Seed=50	-23.57714	-23.07362
N=6,Seed=50	-24.37416	-20.54738

Table 5.2.2.Log likelihood of DBSCAN ans Expectation-Maximization Clustering Algorithm

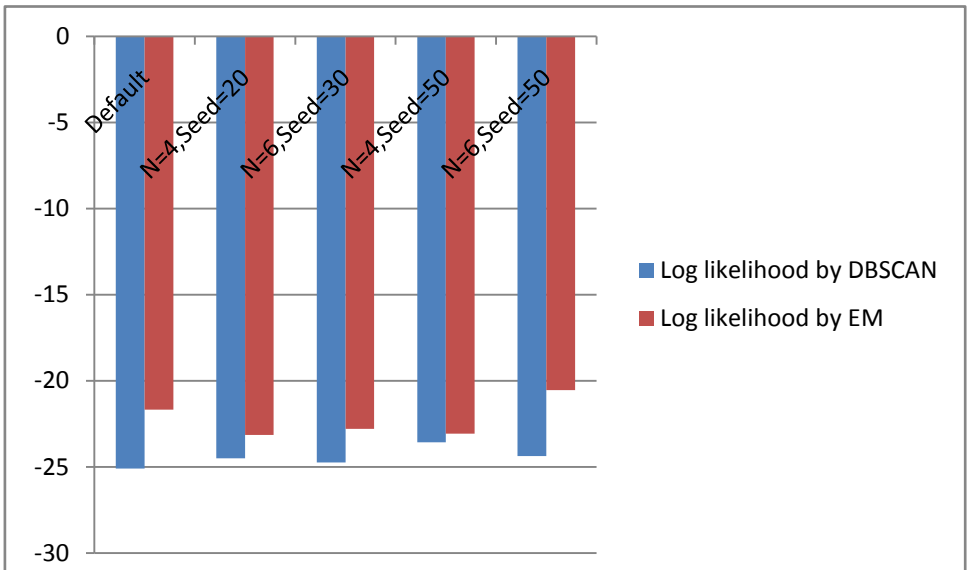

Figure 5.2.2. Graph of Log likelihood with different values of N and Seed of DBSCAN and Expectation-maximizations Clustering Algorithm

Column1	No of iteration by K-means	No of iteration by EM	No of iteration by DBSCAN
Default	9	5	9
N=4,Seed=20	6	4	6
N=4,Seed=50	7	3	7
N=6,Seed=30	13	5	13
N=6,Seed=50	11	5	11

Table 5.2.3. No.of iteration by K-means, EM and DBSCAN Clustering Algorithm

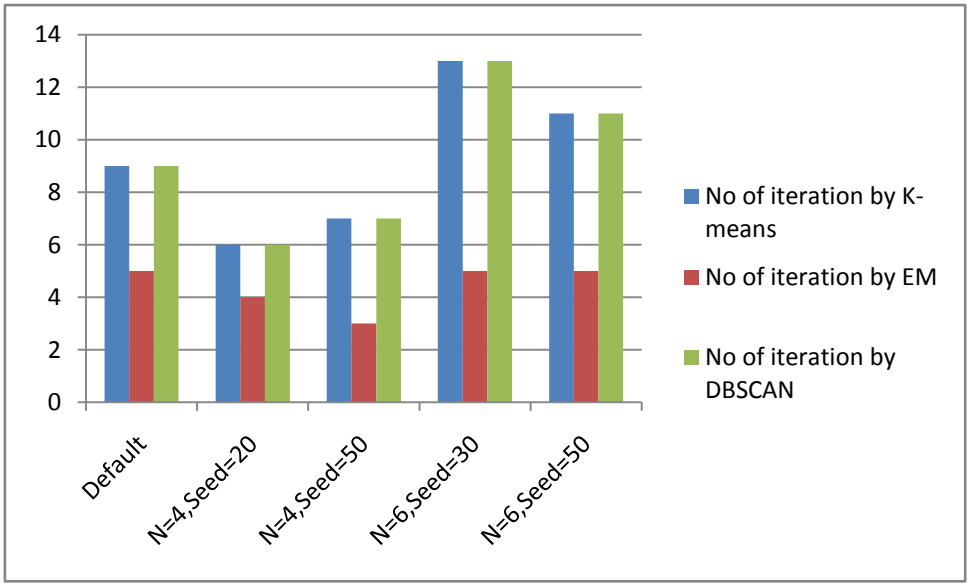

Figure 5.2.3. Graph of No.of iteration with different values of N and Seed of K-means ,EM and Density based Clustering Algorithm

Column1	Squarred error by K-means	Squrred error by DBSCAN
Default	1259.834696	1259.834696
N=4,Seed=20	983.8181909	983.8181909
N=6,Seed=30	1046.119324	1046.119324
N=4,Seed=50	917.5860719	917.5860719
N=6,Seed=50	888.0858326	888.0858326

Table 5.2.4. Squarred error by K-means and DBSCAN Clustering Algorithm

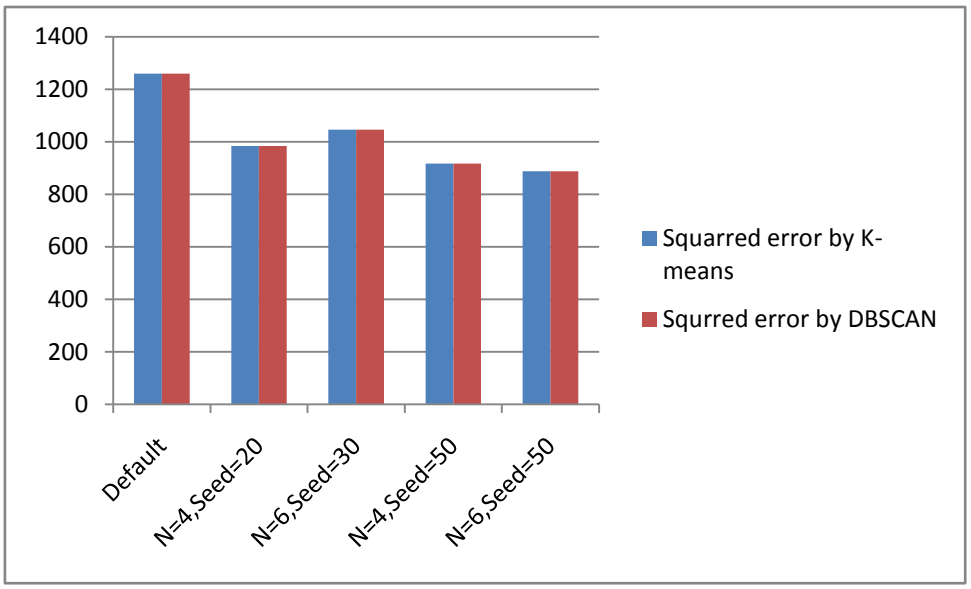

Figure 5.2.4. Graph of Squarred error with different values of N and Seed using K-means and DBSCAN

5.2.1 Evaluation Matrix

The total number of correctly instances divided by total number of instances gives the accuracy. In weka, % of correctly classified instances give the accuracy of the model. The efficiency was measured in terms of running time accuracy.

Accurracy=correctly classified instance/total number of instances

The sample results of "Heart Disease" and "Thyroid Disease" experiment have been shown as follows:

	No.Of Clusters	No.Of Iteration	Accurracy
K-means	3	9	61.9765
E and M	3	5	60.3015
Density Based	3	9	60.9715

Table 5.2.1.1 Accurracy for Heart Disease Dataset

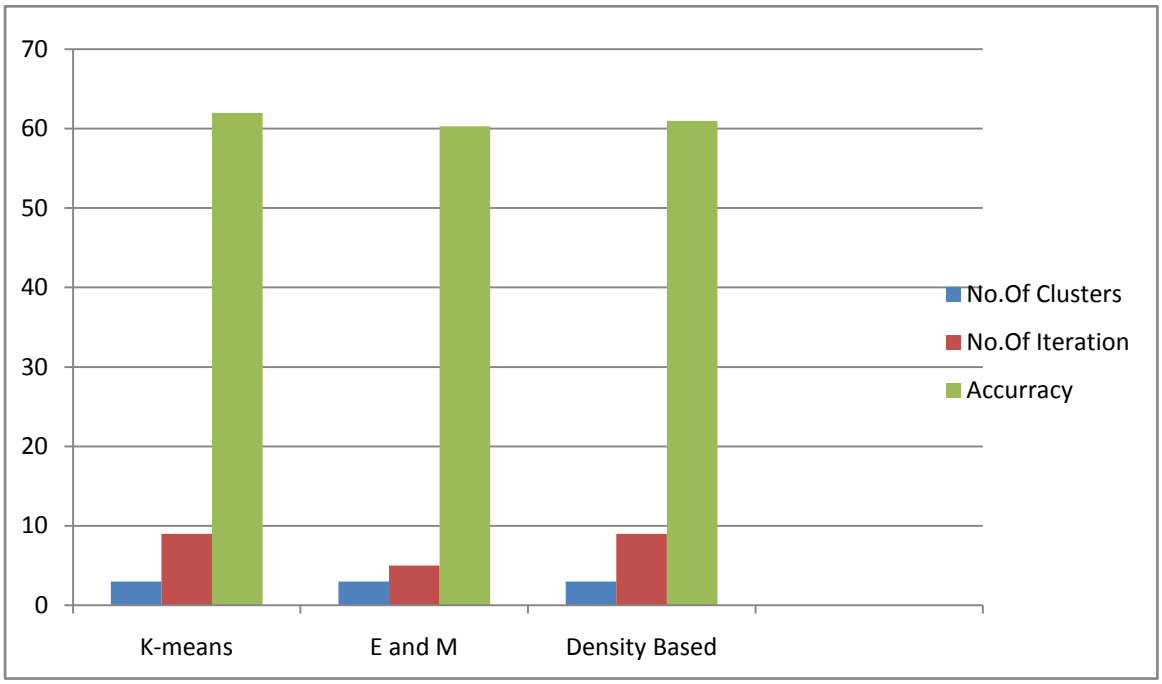

Figure 5.2.1.1 Graph for Parameters No.of iteration ,No of Cluster and Accuracy of Heart Disease Datasets.

Column1	Time Taken	No.Of Clusters
K-means	0.03	3
E and M	0.19	3
DBSCAN	0.11	3

Table 5.2.1.2 Time Taken for Heart Disease Dataset

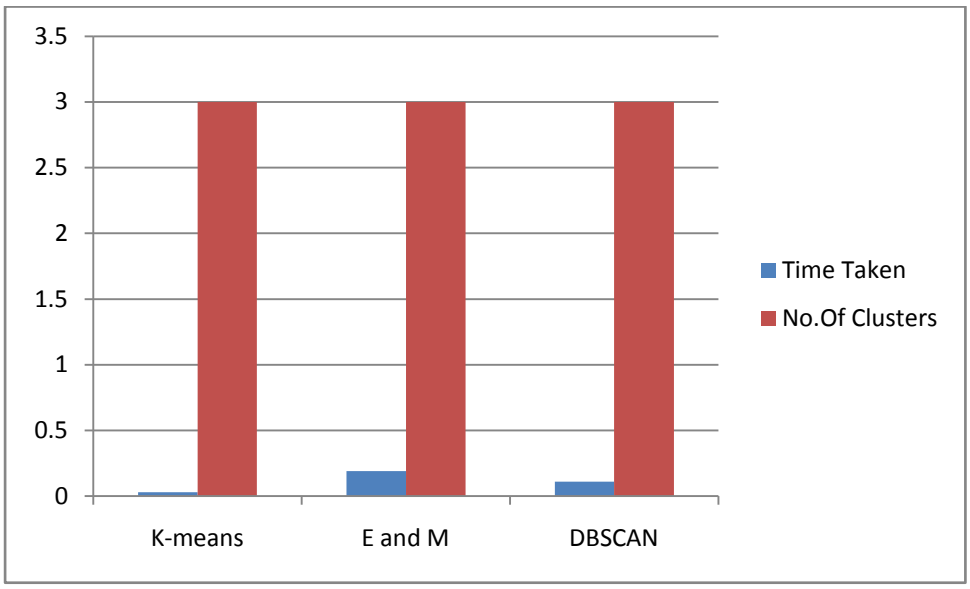

Figure 5.2.1.2 Graph for Parameters No of Cluster and Time Taken of Heart Disease Datasets .

	No.Of Clusters	No.Of Iteration	Accurracy
K-means	3	6	53.7381
E and M	3	2	47.1633
Density Based	3	6	49.4698

Table 5.2.1.3 Accurracy for Thyroid Disease Dataset

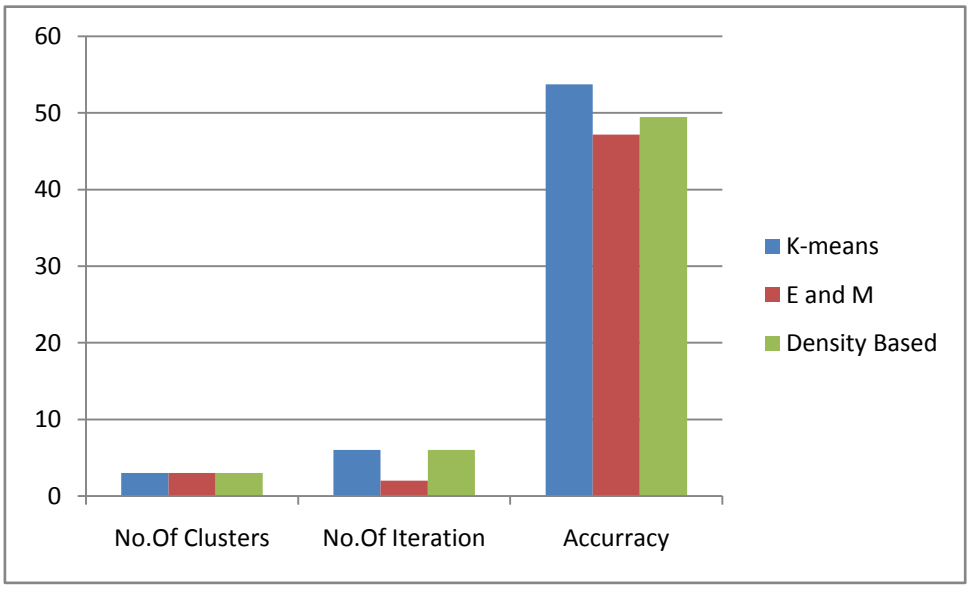

Figure 5.2.1.3 Graph for Parameters No.of iteration ,No of Cluster and Accuracy of Thyroid Disease Datasets.

Column1	Time Taken	No.Of Clusters
K-means	0.06	3
E and M	5.22	3
DBSCAN	0.07	3

Table 5.2.1.4 Time Taken for Thyroid Disease Dataset

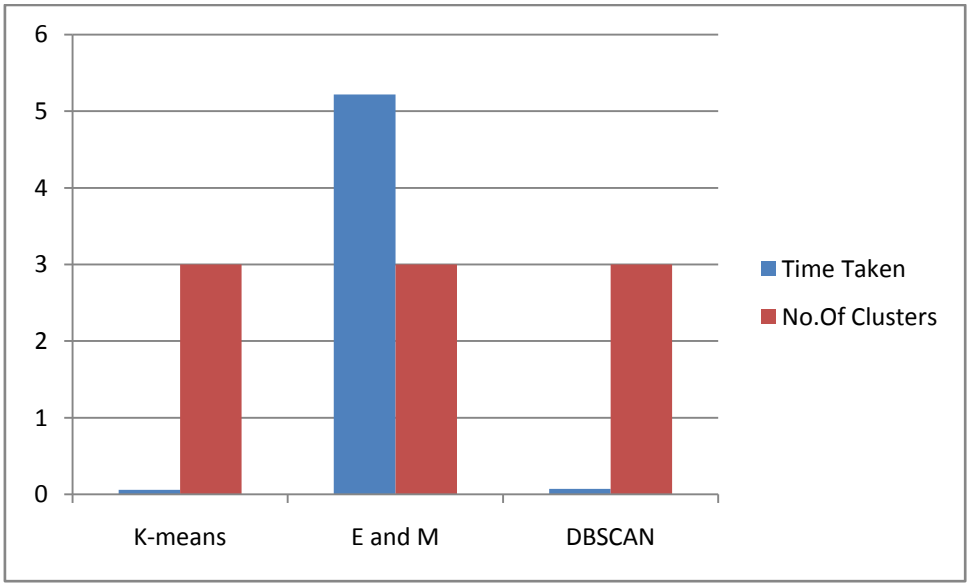

Figure 5.2.1.4 Graph for Parameters No of Cluster and Time Taken of Thyroid Disease Datasets.

CHAPTER 6

CONCLUSION

In this research,comparative study has been performed on the K-means,Expectation-Maximization and Density based clustering algorithms. Comparison is performed on “Heart disease” and “Thyroid Disease” datasets using WEKA tool and the comparative results are presented in the form of table and graph. The comparative study is performed on the basis of parameters accuracy and time taken. All total 597 data of heart disease datasets and 3772 data of Thyroid disease datasets are use for implementing the algorithm.Heart disease use 14 attributes and thyroid disease use 30 attributes.EM clustering and Density based clustering takes more time to form clusters for both datasets (heart disease and thyroid disease datasets).Simple K-means clustering algorithms forms clusters with less time and more accuracy than other algorithms for heart disease and thyroid disease datasets . In terms of time and accuracy K-means produces better results as compared to other algorithms.

References

1. AmandeepKaurMann ,NavneetKaur ,"Survey Paper on Clustering Techniques "Volume 2, Issue 4, April 2013 ISSN: 2278 – 7798.

2. B. Ramesh1, *Clustering Algorithms – A Literature Review*,Dept. of Computer Science, Chikkanna Govt. Arts college (Bharathiyar University), Tirupur, India.

3. Chauhan R, Kaur H, Alam M A, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications , (0975 – 8887) Vol.10– No.6, November 2010.

4. Dr.N.RajalingamK.Ranjini, "Density Based Clustering Algorithm - A Comparative Study" Volume 19– No.3, April 2011, ISSN: 0975 – 8887.

5. Jain A.K., Murty M.N., and Flynn P.J., "Data Clustering: A Review", ACM Computing Surveys, 31 (3). pp. 264-323, 1999.

6. Jiawei Han, MichelineKamber," Data Mining: Concepts and Techniques" Second Edition.

7. P. Nithya, R. Umamaheswari, Dr. N. Shanthi, A *Data Mining Objective Function with Feature Selection Algorithm using Document Clustering*, Gnanamani College of Technology, Namakkal, Tamilnadu, India, April 2015

8. Raj bala ,*A Comparative Analysis of Clustering Algorithms* ,Research Scholar (M.Tech) Amity University Haryana, India

9. Rui Xu, *Survey of Clustering Algorithms*, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE.

10. Sharma Priyanka, *Comparative Analysis of Various Clustering Algorithms Using WEKA* ,Chaudhary Devi Lal University, Sirsa, Haryana, India

11. Steinbach Michael, Karypis George, Kumar Vipin, A *Comparision of Document Clustering Techniques*, Department of Computer Science and Engineering, University of Minnesota.No. 11, pp. 27-34,1996.

12. Sharmila, R.C Mishra “Performance Evaluation of Clustering Algorithms” International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue7- July 2013, ISSN: 2231-5381.

13. S.Revathi, Dr.T.NalinI, “Performance Comparison of Various Clustering Algorithm” Volume 3, Issue 2, February 2013, ISSN: 2277 128X.

14. Thomas Schön, “Machine Learning, Lecture 6 Expectation Maximization (EM) and clustering”,Available at:

15. Tung ThanhKhuat, Hung Duc Nguyen, HanhThi My Le, *A Comparision of Algorithms used to measure the similarity between two documents*, IJARCET, April 2015.

16. Yadav Anshul, Dhingra Sakshi .*A REVIEW ON K-MEANS CLUSTERING TECHNIQUE* ,Department of Computer Science & Application, Sirsa Chaudhary Devi Lal University, Sirsa.

Bibliography

- https://sites.google.com/site/dataclusteringalgorithms/density-based-clustering-algorithmen
- http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/preprocess.html.
- http://transact.dl.sourceforge.net/sourcefor ge/weka/WekaManual-3.6.0.pdf
- https://en.wikipedia.org/wiki/Data_mining
- https://en.wikipedia.org/wiki/Clustering
- https://archive.ics.uci.edu/ml/index.php
- https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

APPENDIX

Implementation Output

K-means clustering Algorithm When Number of Cluster is 2

Run information

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: heart-ch.arff

Instances: 597

Attributes: 14: :age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal

Ignored: num

Test mode: Classes to clusters evaluation on training data

kMeans

Number of iterations: 7

Within cluster sum of squared errors: 1226.7303907482483

Initial starting points (random):

Cluster 0: 51,female,asympt,130,305,f,normal,142,yes,1.2,flat,0,reversable_defect

Cluster 1:

54,male,non_anginal,120,237,f,normal,150,yes,1.5,flat,0.667774,reversable_defect

Final cluster centroids:

Attribute Full	Data	0	1
	597	218	379
age	51.1457	49.4358	52.1293
sex	Male	female	male
cp	Asympt	atyp_angina	asympt
trestbps	132.0956	129.5509	133.5594
chol	248.4286	247.8453	248.764
fbs	F	f	f
restecg	Normal	normal	normal
thalach	144.4765	154.2224	138.8707
exang	No	no	no
oldpeak	0.8162	0.2959	1.1156
slope	Flat	flat	flat
ca	0.6678	0.4532	0.7912
thal	Normal	normal	normal

Time taken to build model (full training data) : 0.03 seconds

Clustered Instances

0 218 (37%)

1 379 (63%)

Class attribute: num

Classes to Clusters:

0 1 <-- assigned to cluster

196 157 | <50

22 222 | >50_1

Cluster 0 <-- <50

Cluster 1 <-- >50_1

Incorrectly clustered instances : 179.0 29.9832 %

Density Based Clustering AlgorithmWhen N= 2

Run Information:

Scheme: weka.clusterers.MakeDensityBasedClusterer -M 1.0E-6 -W weka.clusterers.MakeDensityBasedClusterer -- -M 1.0E-6 -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Instances: 597

Attributes: 14:: age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal

Ignored: num

Test mode: Classes to clusters evaluation on training data

Clustering model (full training set)

MakeDensityBasedClusterer:

Number of iterations: 7

Within cluster sum of squared errors: 1226.7303907482483

Initial starting points (random):

Cluster 0: 51,female,asympt,130,305,f,normal,142,yes,1.2,flat,0,reversable_defect

Cluster 1: 54,male,non_anginal,120,237,f,normal,150,yes,1.5,flat,0.667774,reversable_defect

Final cluster centroids:

Attribute	Full Data	0	1
	597	218	379
age	51.1457	49.4358 5	52.1293
sex	male	female	Male
cp	asympt	atyp_angina	Asympt
trestbps	132.0956	129.5509	133.5594
chol	248.4286	247.8453	248.764
fbs	f	f	F
restecg	normal	normal	Normal
thalach	144.4765	154.2224	138.8707
exang	no	no	No
oldpeak	0.8162	0.2959	1.1156
slope	flat	flat	Flat
ca	0.6678	0.4532	0.7912
thal	normal	normal	Normal

Time taken to build model (full training data) : 0.1 seconds

Model and evaluation on training set

Clustered Instances

0 291 (49%)

1 306 (51%)

Log likelihood: -24.67402

Class attribute: num

Classes to Clusters:

0 1 <-- assigned to cluster

260 93 | <50

31 213 | >50_1

Cluster 0 <-- <50

Cluster 1 <-- >50_1

Incorrectly clustered instances : 124.0 20.7705 %

Expectation-Maximization Clustering Algorithm When N= 2

Run information

Scheme: weka.clusterers.EM -I 100 -N 2 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 10

Relation: heart-ch.arff

Instances: 597

Attributes: 14:: age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal

Ignored: num

Test mode: Classes to clusters evaluation on training data

EM

Number of clusters: 2

Number of iterations performed: 13

Final Cluster Centroids:

Attribute	0	1
	0.54	0.46
Age		
Mean	54.1937	47.5615
std. dev.	8.5085	8.3857
Sex		
Male	234.9702	187.0298
Female	89.6669	89.3331
[total]	324.637	276.363
Cp		
typ_angina	24.0582	11.9418
Asympt	192.935	75.065
non_anginal	68.7749	74.2251
atyp_angina	40.869	117.131
[total]	326.637	278.363
Trestbps		
Mean	134.3481	129.4469
std. dev.	17.9475	16.6724
Chol		
Mean	255.6135	239.9795
std. dev.	62.6771	52.1637
Fbs		
T	47.2617	19.7383
F	277.3753	256.6247
[total]	324.637	276.363
Restecg		

left_vent_hyper	111.5456	43.4544
Normal	185.9411	204.0589
st_t_wave_abnormality	28.1504	29.8496
[total]	325.637	277.363
Thalach		
Mean	138.9026	151.0312
std. dev.	23.0079	22.9738
Exang		
No	160.7252	250.2748
Yes	163.9118	26.0882
[total]	324.637	276.363
Oldpeak		
Mean	1.5104	0
std. dev.	1.0288	0
Slope		
Down	22.0023	1.9977
Flat	224.0135	198.9865
Up	79.6213	76.3787
[total]	325.637	277.363
Ca		
Mean	0.7882	0.5261
std. dev.	0.8347	0.3184
Thal		
fixed_defect	21.013	8.987
Normal	195.1266	247.8734
reversable_defect	109.4975	20.5025
[total]	325.637	277.363

Time taken to build model (full training data) : 0.17 seconds
Model and evaluation on training set
Clustered Instances
0 314 (53%)
1 283 (47%)
Log likelihood: -20.18037
Class attribute: num
Classes to Clusters:
0 1 <-- assigned to cluster
123 230 | <50
191 53 | >50_1
Cluster 0 <-- >50_1
Cluster 1 <-- <50
Incorrectly clustered instances : 176.0 29.4807 %

K-means Clustering Algorith When number of cluster is 3

Run information

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: heart-ch.arff

Instances: 597

Attributes: 14::age ,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal

Test mode: Classes to clusters evaluation on training data

kMeans

Number of iterations: 9

Within cluster sum of squared errors: 1037.9794358525755

Initial starting points (random):

Cluster 0: 51,female,asympt,130,305,f,normal,142,yes,1.2,flat,0,reversable_defect

Cluster 1:

54,male,non_anginal,120,237,f,normal,150,yes,1.5,flat,0.667774,reversable_defect

Cluster 2:

58,male,non_anginal,132,224,f,left_vent_hyper,173,no,3.2,up,2,reversable_defect

Final cluster centroids:

Attribute	Full	0	1	2
	597	238	225	134
Age	51.1457	46.9874	53.6267	54.3657
Sex	Male	Male	male	Male
Cp	Asympt	atyp_angina	asympt	non_anginal
trestbps	132.0956	128.5882	135.8	132.1052
Chol	248.4286	238.6206	260.1295	246.2015
Fbs	F	f	f	F
Restecg	Normal	normal	normal	left_vent_hyper
Thalach	144.4765	150.6261	130.6044	156.8468
Exang	No	no	yes	No
oldpeak	0.8162	0.2639	1.5071	0.6373
Slope	Flat	flat	flat	Up
Ca	0.6678	0.561	0.9131	0.4454
Thal	Normal	normal	normal	normal
Num	<50	<50	>50_1	<50

Time taken to build model (full training data) : 0.03 seconds

Model and evaluation on training set

Clustered Instances

0 206 (35%)

1 248 (42%)

2 143 (24%)

Class attribute: num

Classes to Clusters:

0 1 2 <-- assigned to cluster

184 62 107 | <50

22 186 36 | >50_1

Cluster 0 <-- <50

Cluster 1 <-- >50_1

Cluster 2 <-- No class

Incorrectly clustered instances : 227.0 38.0235 %

Density Based Clustering Algorithm When N=3

Run information

Scheme: weka.clusterers.MakeDensityBasedClusterer -M 1.0E-6 -W weka.clusterers.MakeDensityBasedClusterer -- -M 1.0E-6 -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: heart-ch.arff

Instances: 597

Attributes: 14:: age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal

Ignored: num

Test mode: Classes to clusters evaluation on training data

MakeDensityBasedClusterer:

Number of iterations: 9

Within cluster sum of squared errors: 1037.9794358525755

Initial starting points (random):

Cluster 0: 51,female,asympt,130,305,f,normal,142,yes,1.2,flat,0,reversable_defect

Cluster 1:

54,male,non_anginal,120,237,f,normal,150,yes,1.5,flat,0.667774,reversable_defect

Cluster 2:

58,male,non_anginal,132,224,f,left_vent_hyper,173,no,3.2,up,2,reversable_defect

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data	0	1	2
	597	238	225	134
Age	51.1457	46.9874	53.6267	54.3657
Sex	Male	male	male	Male
Cp	Asympt	atyp_angina	asympt	non_anginal
Trestbps	132.0956	128.5882	135.8	132.1052
Chol	248.4286	238.6206	260.1295	246.2015
Fbs	F	f	f	f
Restecg	Normal	normal	normal	left_vent_hyper
Thalach	144.4765	150.6261	130.6044	156.8468
Exang	No	no	yes	no
Oldpeak	0.8162	0.2639	1.5071	0.6373
Slope	Flat	flat	flat	up
Ca	0.6678	0.561	0.9131	0.4454
Thal	Normal	normal	normal	normal
Num	<50	<50	>50_1	<50

Time taken to build model (full training data) : 0.11 seconds

Model and evaluation on training set

Clustered Instances

0 224 (38%)

1 224 (38%)

2 149 (25%)

Log likelihood: -24.42464

Class attribute: num

Classes to Clusters:

0 1 2 <-- assigned to cluster

201 41 111 | <50

23 183 38 | >50_1

Cluster 0 <-- <50

Cluster 1 <-- >50_1

Cluster 2 <-- No class

Incorrectly clustered instances : 213.0 35.6784 %

Expectation-Maximization Clustering Algorithm When N=3

Run information

Scheme: weka.clusterers.EM -I 100 -N 3 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 10

Relation: heart-ch.arff

Instances: 597

Attributes: 14::age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal

Ignored: num

Test mode: Classes to clusters evaluation on training data

EM

Number of clusters: 3

Number of iterations performed: 5

Final Cluster Centroids:

	Clusters		
Attribute	0	1	2
	0.32	0.36	0.33
Age			
Mean	46.2263	55.2721	51.3742
std. dev.	7.8126	8.2635	8.7367
Sex			
Male	133.2282	169.0238	120.748
Female	56.8901	46.4261	76.6838
[total]	190.1183	215.4499	197.4318
Cp			
typ_angina	8.9918	10.7452	17.263
Asympt	48.811	172.2234	47.9656
non_anginal	46.2994	24.3946	73.3061
atyp_angina	88.0162	10.0867	60.8971
[total]	192.1183	217.4499	199.4318
Trestbps			
Mean	130.2313	136.5244	129.0531
std. dev.	16.4279	19.079	15.7753
Chol			
Mean	242.7352	259.4822	241.8361
std. dev.	61.362	64.3102	46.2738

Fbs			
T	11.1809	33.2223	23.5968
F	178.9375	182.2275	173.835
[total]	190.1183	215.4499	197.4318
Restecg			
left_vent_hyper	6.2004	74.8386	74.961
Normal	155.1162	116.4802	119.4036
st_t_wave_abnormality	29.8018	25.131	4.0672
[total]	191.1183	216.4499	198.4318
Thalach			
Mean	145.5237	129.2279	160.123
std. dev.	22.9509	20.2641	16.4663
Exang			
No	177.2858	62.1827	172.5316
Yes	12.8326	153.2672	24.9002
[total]	190.1183	215.4499	197.4318
Oldpeak			
Mean	0.0003	1.7786	0.5506
std. dev.	0.0097	1.0881	0.6797
Slope			
Down	1	16.5229	7.4771
Flat	186.6308	179.5561	57.8131
Up	3.4875	20.3709	133.1416
[total]	191.1183	216.4499	198.4318
Ca			
Mean	0.6678	0.9384	0.3722
std. dev.	0.664	0.8711	0.5904
Thal			
fixed_defect	6.9948	19.5184	4.4868
Normal	176.9814	115.1554	151.8631
reversable_defect	7.1421	81.776	42.0819
[total]	191.1183	216.4499	198.4318

Time taken to build model (full training data) : 0.19 seconds

Clustered Instances

0 229 (38%)

1 206 (35%)

2 162 (27%)

Log likelihood: -23.30603

Class attribute: num

Classes to Clusters:

0 1 2 <-- assigned to cluster

192 38 123 | <50

37 168 39 | >50_1

Cluster 0 <-- <50

Cluster 1 <-- >50_1

Cluster 2 <-- No class

Incorrectly clustered instances : 237.0 39.6985 %

Implementation Output for Thyroid Disease Datasets

K-means clustering Algorithm When Number of Cluster is 3

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: hypothyroid

Instances: 3772

Attributes: 30

age	thyroid surgery	psych	T4U
sex	I131 treatment	TSH measured	FTI measured
on thyroxine	query hypothyroid	TSH	FTI
query on thyroxine	query hyperthyroid	T3 measured	TBG measured
on antithyroid medication	lithium	T3	TBG
sick	goitre	TT4 measured	referral source
pregnant	tumor	TT4	T4U measured
hypopituitary			

Ignored:

Class

Test mode: Classes to clusters evaluation on training data

kMeans

Number of iterations: 6

Within cluster sum of squared errors: 5161.876331676682

Initial starting points (random):

Cluster 0: 69,F,t,f,f,f,f,f,f,f,f,f,f,f,f,t,1.5,t,1.8,t,136,t,0.92,t,149,f,0,other

Cluster 1: 27,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.15,t,1.6,t,101,t,1.09,t,92,f,0,SVHC

Cluster 2: 82,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.03,t,1.4,t,74,t,0.52,t,143,f,0,other

Missing values globally replaced with mean/mode
Final cluster centroids:

S.No.	Attribute	Full Data	Cluster 0	1	2
		3772	447	1186	2139
1	age	51.7359	52.194	59.7656	47.1879
2	sex	F	F	F	F
3	on thyroxine	F	T	f	f
4	query on thyroxine	F	F	f	f
5	on antithyroid medication	F	F	f	f
6	sick	F	F	f	f
7	pregnant	F	F	f	f
8	thyroid surgery	F	F	f	f
9	I131 treatment	F	F	f	f
10	query hypothyroid	F	F	f	f
11	query hyperthyroid	F	F	f	f
12	lithium	F	F	f	f
13	goitre	F	F	f	f
14	tumor	F	F	f	f
15	hypopituitary	F	F	f	f
16	psych	F	F	f	f
17	TSH measured	T	T	t	t
18	TSH	5.0868	6.1298	3.9196	5.516
19	T3 measured	T	T	t	t
20	T3	2.0135	2.0559	1.6738	2.193
21	TT4 measured	T	T	t	t
22	TT4	108.3193	129.5116	98.6431	109.2558
23	T4U measured	T	T	t	t
24	T4U	0.995	1.0214	0.9277	1.0268
25	FTI measured	T	T	t	t
26	FTI	110.4696	127.2989	107.9607	108.3438
27	TBG measured	F	F	f	f
28	TBG	0	0	0	0
29	referral source	Other	Other	SVI	other

Time taken to build model (full training data) : 0.6 seconds

Clustered Instances

0 447 (12%)

1 1186 (31%)

2 2139 (57%)

Class attribute: Class

Classes to Clusters:

0 1 2 <-- assigned to cluster

440 1084 1957 | negative

0 63 131 | compensated_hypothyroid

7 39 49 | primary_hypothyroid

0 0 2 | secondary_hypothyroid

Cluster 0 <-- primary_hypothyroid

Cluster 1 <-- compensated_hypothyroid

Cluster 2 <-- negative

Incorrectly clustered instances : 1745.0 46.2619 %

EM

Scheme: weka.clusterers.EM -I 100 -N 3 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Ignored:

Class

Test mode: Classes to clusters evaluation on training data

EM

Number of clusters: 3

Number of iterations performed: 2

Final Cluster centroids:

	Cluster		
Attribute	0	1	2
	-0.07	-0.29	-0.64
age			
mean	47.0557	52.097	52.0737
std. dev.	19.3987	17.823	21.0495
Sex			
F	217.1364	60.3837	2355.4799
M	43.9454	1051.9634	49.0913
[total]	261.0817	1112.347	2404.5712
on thyroxine			
f	237.2369	1038.7405	2035.0226
t	23.8449	73.6065	369.5486
[total]	261.0817	1112.347	2404.5712
query on thyroxine			
f	260.0817	1086.5699	2378.3483
t	1	25.7771	26.2229
[total]	261.0817	1112.347	2404.5712
on antithyroid medication			
f	254.355	1104.2893	2373.3557
t	6.7268	8.0577	31.2155
[total]	261.0817	1112.347	2404.5712
Sick			
f	250.0808	1063.4228	2314.4964
t	11.0009	48.9243	90.0748
[total]	261.0817	1112.347	2404.5712
Pregnant			
f	259.0817	1111.3465	2351.5718
t	2	1.0006	52.9994
[total]	261.0817	1112.347	2404.5712
thyroid surgery			
f	259.081	1102.9136	2360.0054
t	2.0008	9.4334	44.5658
[total]	261.0817	1112.347	2404.5712
I131 treatment			
f	259.2481	1099.2222	2357.5298
t	1.8337	13.1249	47.0415
[total]	261.0817	1112.347	2404.5712
query hypothyroid			
f	247.9593	1067.825	2225.2157
t	13.1224	44.522	179.3556
[total]	261.0817	1112.347	2404.5712

query hyperthyroid			
f	236.0398	1072.6039	2229.3562
t	25.0419	39.7431	175.215
[total]	261.0817	1112.347	2404.5712
Lithium			
f	260.0817	1106.6436	2390.2747
t	1	5.7034	14.2966
[total]	261.0817	1112.347	2404.5712
Goiter			
f	257.0817	1098.814	2385.1043
t	4	13.533	19.467
[total]	261.0817	1112.347	2404.5712
Tumor			
f	248.0818	1103.1583	2327.7599
t	12.9999	9.1888	76.8113
[total]	261.0817	1112.347	2404.5712
Hypopituitary			
f	260.0817	1110.3478	2403.5705
t	1	1.9992	1.0008
[total]	261.0817	1112.347	2404.5712
Psych			
f	260.0817	1008.8228	2322.0954
t	1	103.5242	82.4758
[total]	261.0817	1112.347	2404.5712
TSH measured			
t	55.0545	1049.0121	2301.9334
f	206.0272	63.3349	102.6379
[total]	261.0817	1112.347	2404.5712
TSH			
mean	4.5548	2.0794	6.534
std. dev.	1.4276	2.4303	29.0203
T3 measured			
t	63.0151	960.3389	1982.6461
f	198.0667	152.0082	421.9252
[total]	261.0817	1112.347	2404.5712
T3			
mean	2.0234	1.9235	2.054
std. dev.	0.2146	0.5701	0.8336
TT4 measured			
t	36.0978	1110.2892	2397.6131
f	224.984	2.0579	6.9582
[total]	261.0817	1112.347	2404.5712

TT4			
mean	108.0631	99.4207	112.4595
std. dev.	2.0541	23.3372	39.5209
T4U measured			
t	1.1237	1065.9183	2320.958
f	259.9581	46.4287	83.6132
[total]	261.0817	1112.347	2404.5712
T4U			
mean	0.995	0.9284	1.0258
std. dev.	0.0002	0.1369	0.2053
FTI measured			
t	1.1237	1066.9472	2321.9291
f	259.9581	45.3998	82.6422
[total]	261.0817	1112.347	2404.5712
FTI			
mean	110.4698	108.4858	111.3865
std. dev.	0.0316	22.0075	36.2852
TBG measured			
f	260.0817	1111.347	2403.5712
[total]	260.0817	1111.347	2403.5712
TBG			
mean	0	0	0
std. dev.	0	0	0
referral source			
SVHC	2	209.6133	177.3867
other	252.2494	478.1643	1473.5864
SVI	5.8324	415.0767	616.0909
STMW	2	1.0166	111.9834
SVHD	2	11.4762	28.5238
[total]	264.0817	1115.347	2407.5712

Time taken to build model (full training data) : 5.22 seconds

Clustered Instances

0 274 (7%)

1 1627 (43%)

2 1871 (50%)

Log likelihood: -9.04454

Class attribute: Class

Classes to Clusters:

0 1 2 <-- assigned to cluster

270 1606 1605 | negative

3 19 172 | compensated_hypothyroid

1 1 93 | primary_hypothyroid

0 1 1 | secondary_hypothyroid

Cluster 0 <-- primary_hypothyroid

Cluster 1 <-- negative

Cluster 2 <-- compensated_hypothyroid

Incorrectly clustered instances : 1993.0 52.8367 %

DBSCAN

Scheme: weka.clusterers.MakeDensityBasedClusterer -M 1.0E-6 -W weka.clusterers.MakeDensityBasedClusterer -- -M 1.0E-6 -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Ignored:

Class

Test mode: Classes to clusters evaluation on training data

MakeDensityBasedClusterer:

Wrapped clusterer: MakeDensityBasedClusterer:

Wrapped clusterer:

kMeans

Number of iterations: 6

Within cluster sum of squared errors: 5161.876331676682

Initial starting points (random):

Cluster 0: 69,F,t,f,f,f,f,f,f,f,f,f,f,f,f,t,1.5,t,1.8,t,136,t,0.92,t,149,f,0,other

Cluster 1: 27,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.15,t,1.6,t,101,t,1.09,t,92,f,0,SVHC

Cluster 2: 82,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.03,t,1.4,t,74,t,0.52,t,143,f,0,other

Missing values globally replaced with mean/mode

Final cluster centroids:

Cluster				
Attribute	Full Data	0	1	2
	-3772	-447	-1186	-2139
age	51.7359	52.194	59.7656	47.1879
sex	F	F	F	F
on thyroxine	f	t	f	F
query on thyroxine	f	f	f	F
on antithyroid medication	f	f	f	F
sick	f	f	f	F
pregnant	f	f	f	F
thyroid surgery	f	f	f	F
I131 treatment	f	f	f	F
query hypothyroid	f	f	f	f
query hyperthyroid	f	f	f	f
lithium	f	f	f	f
goitre	f	f	f	f
tumor	f	f	f	f
hypopituitary	f	f	f	f
psych	f	f	f	f
TSH measured	t	t	t	t
TSH	5.0868	6.1298	3.9196	5.516
T3 measured	t	t	t	t
T3	2.0135	2.0559	1.6738	2.193
TT4 measured	t	t	t	t
TT4	108.3193	129.5116	98.6431	109.2558
T4U measured	t	t	t	t
T4U	0.995	1.0214	0.9277	1.0268
FTI measured	t	t	t	t
FTI	110.4696	127.2989	107.9607	108.3438
TBG measured	f	f	f	f
TBG	0	0	0	0
referral source	other	other	SVI	other

Time taken to build model (full training data) : 0.7 seconds

Clustered Instances

0 471 (12%)

1 1366 (36%)

2 1935 (51%)

Log likelihood: -10.49713

Class attribute: Class

Classes to Clusters:

0 1 2 <-- assigned to cluster

427 1294 1760 | negative

2 64 128 | compensated_hypothyroid

42 8 45 | primary_hypothyroid

0 0 2 | secondary_hypothyroid

Cluster 0 <-- primary_hypothyroid

Cluster 1 <-- compensated_hypothyroid

Cluster 2 <-- negative

Incorrectly clustered instances : 1906.0 50.5302 %