CHAPTER 1
INTRODUCTION

Introduction

Clustering is most powerful technique in data mining which is used to grouping data with their similarity (i.e. similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster). It plays a significant role for data analysis in various fields such as Data mining [1], image segmentation [2], computational biology, mobile communication, medicine, economics and many more. Clustering is also known as unsupervised learning because the class label information is not present. So we can say clustering is form of learning by observation, rather than learning by example. Clustering is emerging topic for researcher so many researchers have developed number of algorithms [3, 4] and face some problems for real situation. For example, K-means is very simple and robust but it requires users to provide the number of cluster, which usually unknown in advance.

Minimum Spanning Tree (MST) is sub-graph that spans over all the vertices of given graph without any cycle and minimum sum of weight over all the connected edge. MST is an useful data structure that has been used to design many clustering algorithm. The first cluster was designed by Zhan using MST [5]. In his paper MST was constructed from given dataset and then removes inconsistent edges to create connected component. Now days several researchers are extensively studying MST based clustering algorithm for biological data analysis [6], image processing, pattern recognition, outlier detection, etc. Three approaches were proposed by Xu et al [6] i.e., clustering by removing longest MST edges, iterative clustering and globally optimal clustering. Algorithm is effective but users do not know how to select inconsistency edges for their removal without any prior knowledge of the structure of the data patterns.

In this dissertation the threshold values are applied randomly on MST based clustering algorithm, motivated with studying many MST based clustering algorithm. This algorithm has no specific requirement of prior knowledge of some parameters and dimensionality of data set. This algorithm can be run using different threshold values on MST and get many group of object. Then those groups of object are analyzed by validity index. Minimum validity index selects the best threshold value of minimum spanning tree for clustering.

1.2 Problem Definition

The problem of choosing best threshold for clustering is the main problem faced in this dissertation. There is no certain rule and regulation for solving such kind of problem. There are so many MST based clustering algorithms for data mining had been developed in past some years and also they have successful implementation. However, choosing the best threshold value for different datasets is always a challenging task as different algorithm can have different performance and accuracies with different type of datasets

1.3 Objective

The objectives of this research are:

- To represent n-dimensional data point in the form of dissimilarity matrix and generated graph from it and then MST from graph.
- To implement and analysis MST based algorithm with various threshold values
- To select best threshold value for best clustering according to their minimum validity index.

1.4 Motivation

Clustering is powerful tool that play vital role for data analysis in various field such as data mining, image segmentation, computational biology, mobile communication and many more. Different graph based and non-graph based algorithms have different problem such as k-means algorithm requires user's to provide the number of clusters, MST based algorithm such as dynamic clustering technique requires iteration method for many threshold values and analysis them by validity index for choose best threshold value. Motivated with these algorithms, some threshold values are taken from mean, SD and mean + SD of MST and analysis them by validity index for best threshold value.

1.5 Report Organization

The background part of this dissertation work focuses on MST, Threshold, and validity index. This entire dissertation works are organized as follows.

Chapter 1:- This chapter is focused on overview of Clustering Analysis, Minimum Spanning Tree (MST), Problem definition, Motivation and Objective of this dissertation.

Chapter 2:- This chapter consists of literature review where previous research works are explained.

Chapter 3:- This chapter contains details information about research methodology.

Chapter 4:- This chapter explains about prim's algorithm and clustering algorithm.

Chapter 5:- This chapter describes Implementation section where different tools, programming language and data-structure are explained.

Chapter 6:- This chapter contains Result Analysis and Comparison section and describes how data are collected and analysis them.

Chapter 7:- Finally this chapter shows Conclusion of this whole dissertation work and guidelines for future research.

CHAPTER 2
LITERATURE REVIEW

2.1 Literature Review

Clustering analysis is simply a process of identifying similar object in one group and dissimilar object in other group from large collection of data. Many researchers are applying this term using different methods for filtering unknown class label data. Clustering term is always in top of the researcher's list because we are living in data age and everything we can predict to analyze the data.

T.Soni Madhulatha [7] explained various types of clustering methods and their meaningful interpretation. He has generally overview of clustering methods on his work. He classified clustering methods are as follows: - Partitioned, Hierarchical, Density based, Grid based and Model based. Partitioned clustering methods are distance based and concerned with K-clusters randomly and assign each data to the nearest cluster center. It may use mean or mediod to represent cluster centers. Hierarchical clustering method creates hierarchical decomposition of given data set. It may incorporate other technique like micro clustering or consider object linkages. Its main drawback is that once the merge/split is done can never be undone. Density based methods are used to find arbitrary shaped clusters and filter out outliers. Density based clustering algorithms allocates object depending on its density. Grid based clustering algorithms are used for multi-resolution grid data structure. This algorithm allocates object depending on its location grid. Model based algorithms are concerned with its weight vector and attempt to optimize fit between the given data and mathematical model.

Tan et al [8] also explained various types of cluster and clustering algorithms with their application areas. They explained quality and types of cluster for clustering algorithm such as Well- separated, Prototyped based, Shared property, graph based and density based cluster. The concept used was eigenvector matrix. Some of the clustering algorithms are K-means and its variants, hierarchical and density based. They explained various application of cluster analysis such as in Biology, Information retrieve, Climate, Psychology and medicine, Business.

A.K Jain et al. [9] have been broadly explained about data clustering. The explanation includes lots of other concept like history of clustering, application area of clustering including artificial intelligence. They have explained various types clustering algorithms are as follows: Hierarchical, Partitional,

Mixture Resolving, Nearest-Neighbor and Fuzzy. Similarly, the works included partitional concept, Graph theoretic, Squared Error, Mixture Resolving and mode seeking.

Graph based and non-graph based algorithms have been compared by Nabin Ghimire [10]. He explained graph based algorithm such as Prim's algorithm and non-graph based algorithm such as K-mediods and Prim's algorithm. These algorithms have been analyzed by widely popular statistical measure-chi square test. For the evaluation of the algorithms, numerical co-ordinate data is used and tested their convergence.

MST based clustering technique has been explained by V.M.K et al. [11]. Prim's algorithm and Clustering algorithm have been explained in their paper. Where MST based clustering algorithm has been compared with K- means algorithm. The evaluation of this algorithm has been evaluated by ratio between intra-cluster distance and inter-cluster distance. In this paper, initially they set the threshold value (except 0) and then remove those edges from MST whose length are greater than the threshold value. Next they calculated validity index and recorded validity index as well as the threshold value. The above procedure repeats continuously until threshold value is maximum than MST edges.

MST has been constructed by Bhaskar Adepu et al. [12] using new method which reduces the time complexity compared with traditional construction method. For generating cluster from MST they used threshold value (i.e., sum of mean and SD). In their method the user does not require to provide number of cluster like K-means algorithm. They have been compared their proposed method with k-means algorithm, which performs better than k- means algorithm.

Oleksandr Grygorash et al. [13] proposed EMST based clustering algorithm. In this paper, they classified EMST based clustering algorithm into two minimum spanning tree based clustering algorithms. The first algorithm produces k-partition of a set of points for any given k, i.e. HEMST clustering algorithm. The algorithm constructs a minimum spanning tree of the point set and removes edges that satisfy a predefined criterion. The process is repeated until k clusters are produced. The second algorithm partitions a point set into a group of clusters by maximizing the overall standard deviation reduction, without a given k value i.e. MSDR clustering algorithm. They presented their experimental results comparing their proposed algorithms to k-means and EM.

They also applied their algorithms to image color clustering and compared their algorithms to the standard minimum spanning tree clustering algorithm.

Xiaochun Wang et.al.[14], purposed a new approach called divided and conquer method to facilitate efficient MST-based clustering by using the idea of the revers-delete algorithm. The divide and conquer algorithm mainly divided into two phases. The first phase includes the sequential initialization and DHCA spanning tree updating and second phase uses the MDHCA to locate the longest edges and partition the obtained approximate MST to form sensible cluster. The proposed algorithm works better than the classical clustering algorithm in both the execution time and classification results.

Prasanta K. Jana et.al.[15], Purposed an Efficient Minimum Spanning Tree based clustering algorithm. The proposed algorithm consists of two phase, the first phase they generated MST using Kruskal algorithm and second phase they generated different clusters using threshold value on MST. In this paper Validity Index (ratio of intra cluster distance and inter cluster distance) uses for evaluation of best threshold value on MST. They showed their proposed algorithm performs better than the k-means algorithm by some experimental results.

S.Ray et al. [16] purposed the validity measure based on intra cluster-distance and inter cluster-distance which allows the number of cluster to be determined automatically. This method is based on k-means algorithm and it overcomes the limitation of having to indicate the number of clusters by incorporating a validity measure. In this paper they evaluated the synthetic images and natural images of clusters. The performance of their proposed cluster validity measure of synthetic images and natural images are compared with Bouldin and Dunn's indexes. In synthetic images and natural images validity index measure performs well than Bouldin and Dunn's indexes.

C.Zong et al. [17], proposed A graph-theoretical clustering method based on two rounds of minimum spanning trees. In this paper they explained two clustering algorithm which deals with separated clusters and touching clusters in two phases. In the first phase, two round Minimum Spanning trees are employed to construct a graph and detect separated clusters which cover distance separated and density separated clusters. In the second phase, touching clusters which are subgroups produced in first phase can be partitioned by comparing cuts, respectively on the two round MST. In this paper a two round

MST based graph is robust to different cluster shapes, sizes and densities, and it does not request user-defined cluster number.

The randomly generated data sets as well as real world data set such as Wine is taken from UCI machine leaning repository [18]. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. The University of California - Irvine holds a repository of datasets which are used by practitioners and researchers in the fields of Artificial Intelligence, Pattern Recognition, Machine Learning, Neural Networks, Data Mining, Bio-informatics and others these are referred to as the UCI datasets.

CHAPTER 3

RESEARCH METHODOLOGY

This dissertation work is fully experimental. It is a traced driven approach in which random numbers of input size (n) and UCI repository data are collected and traced MST using Prim's algorithm. With different threshold values on MST different clusters are traced using MST based clustering algorithm. To analyze those clusters validity index uses with python programming language. Finally, conclusion is drawn with help of analyzed clusters.

3.1 Data collection

In this dissertation three different data sets are collected, one data set is collected from UCI machine learning repository and two data sets are randomly generated. The datasets have been chosen such that they differ in size, mainly in terms of number of instances and number of attributes.

3.2 Data analysis

MST based Clustering algorithm used to analyze the different data sets on python programming language. With different threshold values on MST different clusters are traced using MST based clustering algorithm. Intra and inter distances are calculated from different clusters to find out the validity index. From overall validity index, Minimum validity index is selected best clustering for best threshold value.

3.3 Performance Metrics

To evaluate best threshold value for best number of cluster from MST, validity index is used in this dissertation. There are many metrics for evaluating clustering algorithm like Dunn's index, Davies-Bouldin index, Silhouette index, Maulik-Bandyopadhyay index and Score Function [19]. Validity Index is explained in this dissertation which purposed by Ray and Turi.

3.3.1 Validity Index

Validity index is used to evaluate the quality of clusters produced by clustering algorithm. The validity index is based on compactness and isolation of cluster. Compactness of cluster is measured by the intra-cluster distance and isolation between the cluster is measured my inter-cluster distance. They are defines as follows.

Intra-cluster distance: This is average distance of all the points within the cluster from the cluster center and is calculated by

$$\textbf{Intra-distance} = \frac{1}{N}\sum_{i=1}^{k}\sum_{x\epsilon z_i}\left|\left|x - z_i\right|\right|^2$$

Where z_i is the center of cluster, N is number data point, K is total number of cluster and X is the weight of the data point.

Inter-cluster distance: This is the minimum of the pair wise distance between any two cluster centers and is calculated by

$$\textbf{Inter-distance} = \min \| z_i - z_j\|^2$$

Where i=1,2,…k-1; j=i+1, i+2, i+3,…..k

For the evaluation of clustering algorithm, the validity index proposed by Ray and Turi [16] has been used here as follows.

$$\textbf{Validity index} = \frac{\text{Intra\_distance}}{\text{inter\_distance}}$$

3.3.2 Threshold values

This denotes the limit when two points get disconnected if the distance between them is greater than this limit. This types of limit is taking by following way

i. **Mean**: - Mean is average value of MST.
ii. **Standard Deviation**: -It is the square root of variance. Variance is the average of squared difference from the mean.

$$\textbf{SD } (\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x-\mu)^2}$$

iii. **Mean + Standard Deviation**: - This is combination of both mean and SD.

Using the different threshold value on MST get number of groups of clusters then those groups of clusters are analyzed by validity index and choose the best threshold values for best number of clusters.

CHAPTER 4

ALGORITHMS

4.1 Prim's Algorithm

It is a greedy algorithm which uses to generate the Minimum Spanning Tree. Using Prim's algorithm, K-Minimum Spanning tree (K-MST) algorithm is adopted. The spanning trees are cut k-1 edges to form K-clusters.

1. Set S=Ø
2. Randomly choose any vertex in graph.
3. Add the minimum edges incident to that vertex to S
4. Continue to add the edges into S until all the vertices of G included with n-1 edges without form of cycle.

Here has been given a graph from which we can construct minimum spanning tree using prim's algorithm.

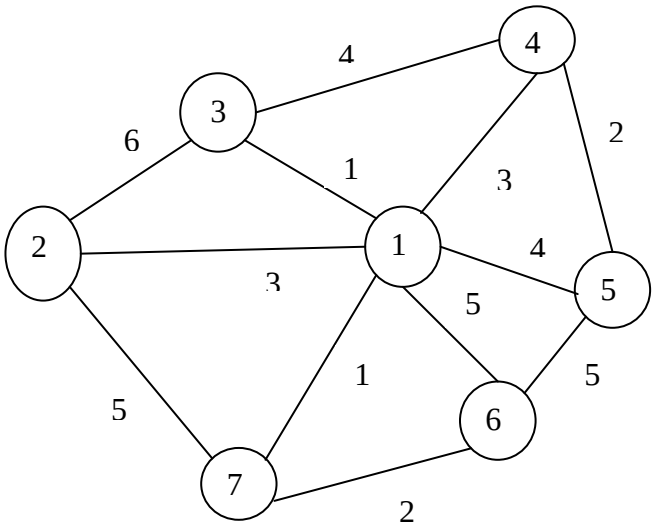

Figure1:- Graph

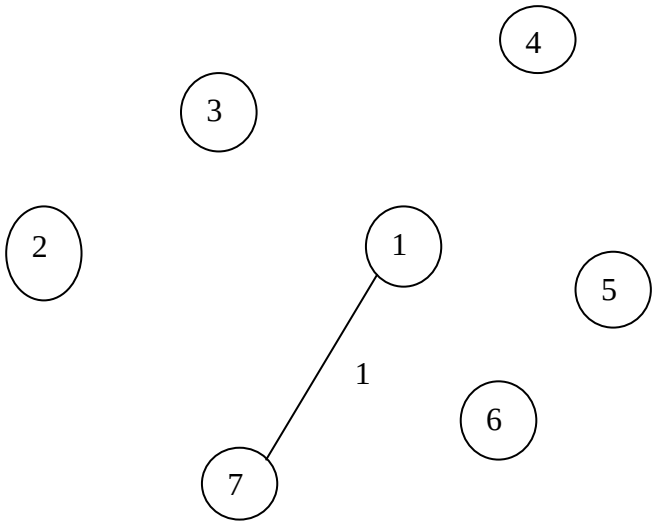

Figure1.1:-Constructing MST using prim's algorithm, U = {1}

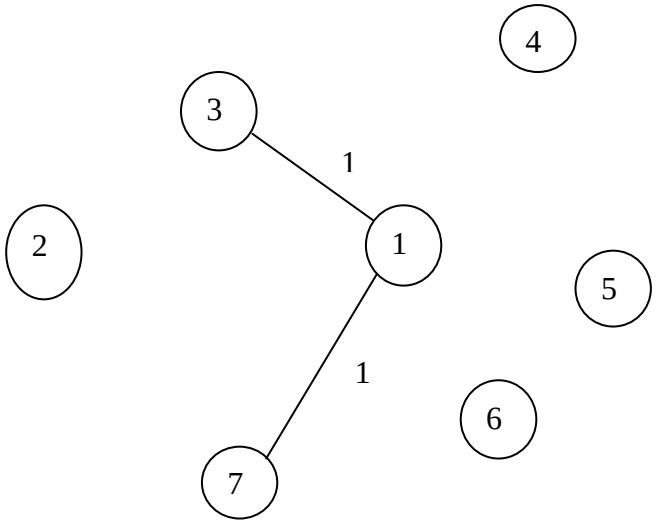

Figure1.2:-Constructing MST using prim's algorithm, U = {1, 1}

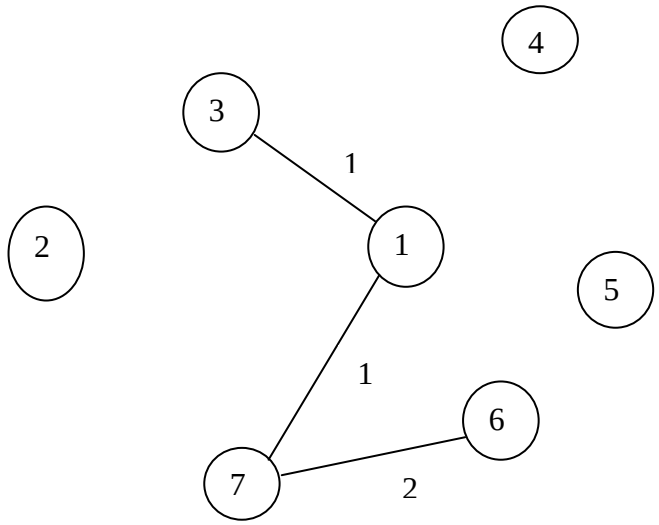

Figure1.3:-Constructing MST using prim's algorithm, U = {1, 1, 2}

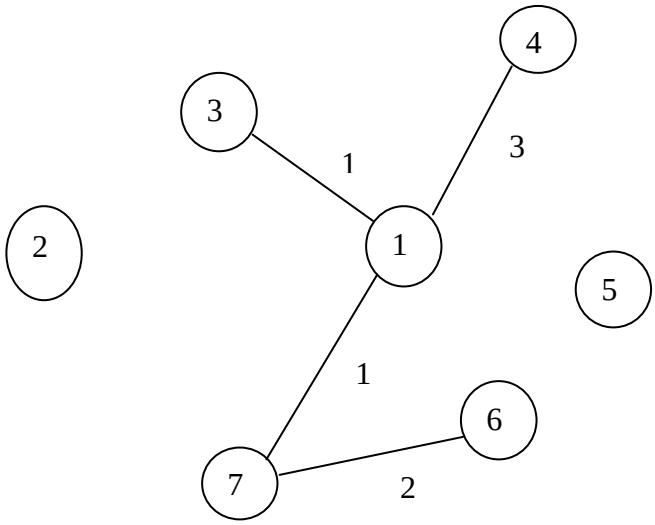

Figure1.4:-Constructing MST using prim's algorithm U = {1, 1, 2, 3}

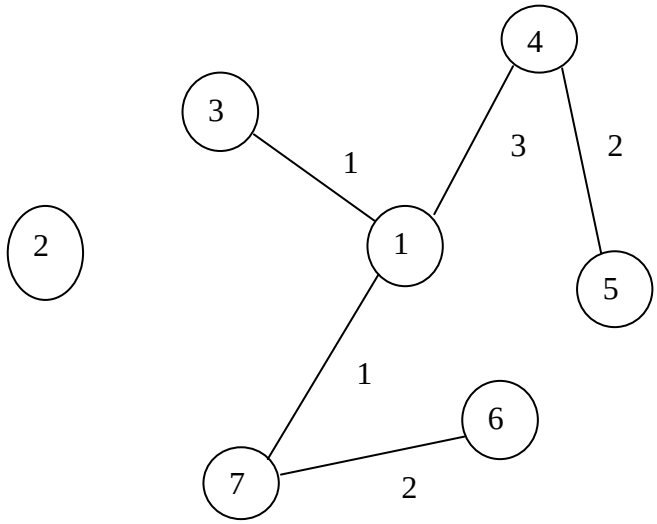

Figure1.5:-Constructing MST using prim's algorithm U = {1, 1, 2, 3, 2}

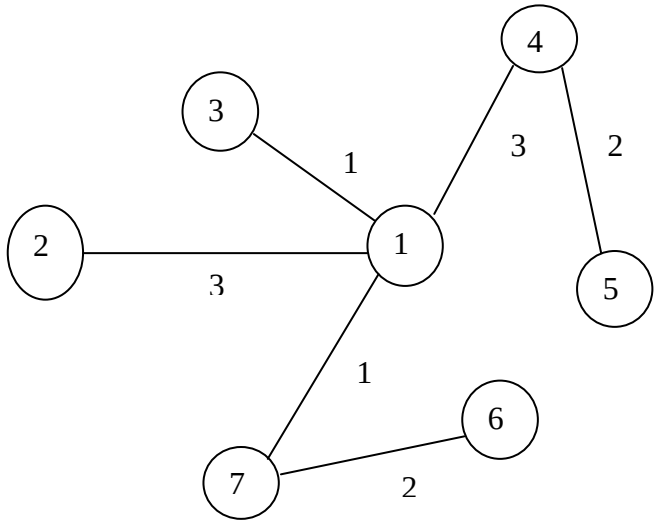

Figure1.6:-MST obtained by using prim's algorithm, U = {1, 1, 2, 3, 2, 3}

The graphs are cut into largest k-1 edges to form a k cluster, which is termed as K-MST. The graphs are made considering every point or pixel as vertex in this work. Further different clusters are generates using threshold values on this Minimum Spanning Tree.

4.2 MST based clustering algorithm

In MST based clustering, the weight for each edge is considered as the Euclidean distance between end points forming that edge. Applying a threshold value on MST inconsistence edges are removed from the MST. Generally those edges should be greater than threshold value. The connected components of the MST obtained by removing these edges are treated as the clusters. In this dissertation MST based clustering algorithm work as follows.

1. Storage of graph using list
2. Choose the threshold and remove the edge whose weight is greater than threshold
3. Make clusters

   //a is an array of graph edges

   //cl is a 2-d array to store clusters

a. c=0, Put the first element of a to cl[0] and remove that element from graph storage list then intersect that element with second element of a.

b. If both elements have intersection number then union of both elements and construct cluster_1 and adding number of element in cluster_1 by repeating above same procedure.

   Else

   Construct cluster_2 and adding number of element by repeating above same procedure.

   Construct cluster_n by above same procedure until removed all elements of a in graph storage list.

   Here has been given Minimum Spanning Tree, where we apply threshold value and obtain number of group of cluster using MST based clustering algorithm.

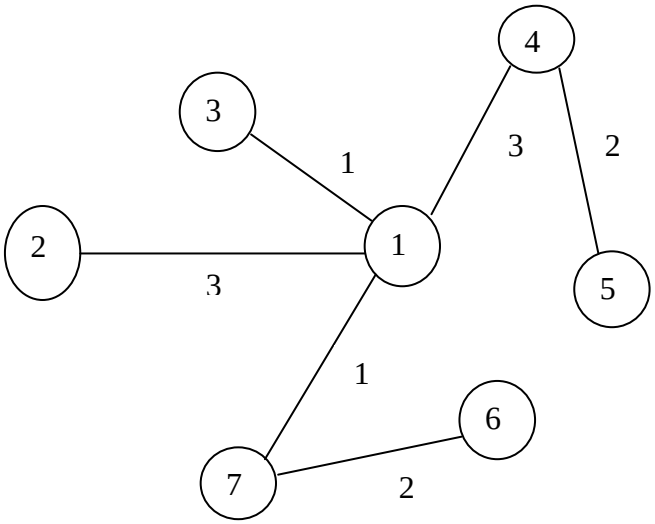

Figure 2:- Minimum Spanning Tree

Storage a = [[node1, node2, edge1], [….], …]

a = [[2,1,3], [3,1,1], [1,4,3], [4,5,2], [1,7,1], [7, 6, 2]]

Suppose threshold value is mean of MST i.e. 2

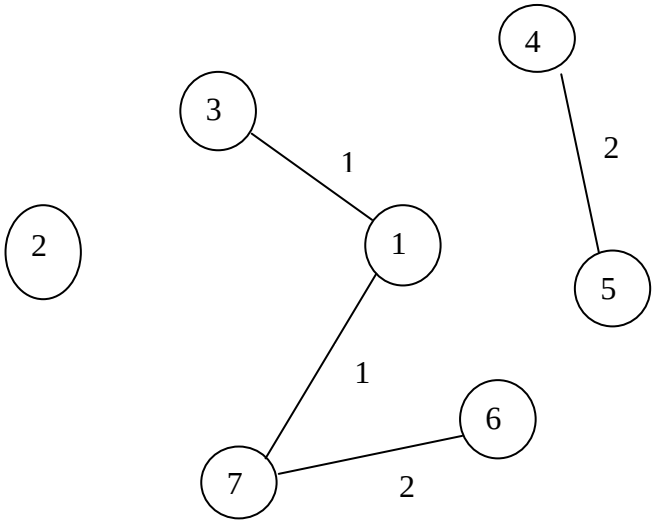

Figure 2.1:- Number of cluster are obtained by MST based clustering algorithm

a = [[2, 2, 0], [1, 1, 0], [3, 1, 1], [1, 7, 1], [1, 1, 0], [4, 4, 0], [7, 6, 2], [4, 5, 2]]

x = [[2, 2], [1, 1], [3, 1], [1, 7], [1, 1], [7, 6], [4, 4], [4, 5]]

Cluster = [[ ]]

Cluster1= [[2, 2]]

Union = {2, 2}

Intersection = (2, 2) ⋂ (1, 1) =Ø

Cluster2 = [[1, 1]]

Union = {1, 1}

Intersection = (1, 1) ⋂ (3, 1) ≠ ϕ

Union = {1, 3}

Cluster2 = [[1, 3]]

Intersection = (1, 3) ⋂ (1, 7) ≠ ϕ

Union = {1, 3, 7}

Cluster2 = [[1, 3, 7]]

Intersection = (1, 3, 7) ⋂ (1, 1) ≠ ϕ

Union = {1, 3, 7}

Cluster2 = [[1, 3, 7]]

Intersection = (1, 3, 7) ⋂ (7, 6) ≠ ϕ

Union = {1, 3, 6, 7}

Cluster2 = [[1, 3, 6, 7]]

Intersection = (1, 3, 6, 7) ⋂(4, 4)= ϕ

Cluster3 = [[4,4]]

Intersection $= (4,4) \bigcap (4,5) \neq \phi$

Union $= \{4, 5\}$

Cluster3 = [[4, 5]]

CHAPTER 5
IMPLEMENTATION

5.1 Tools Used

In this dissertation all algorithms are implemented in Python language using python IDLE 2.7.10 with some modules (i.e. scipy and numpy) and data structure (i.e. python list).

5.2 Programming language

All algorithms are implemented by python programming language in windows 7 operating system, 64 bit machine having 2GB RAM with Intel CORE^{tm} i3 processor. Python is general purpose, concurrent, class based, object oriented, platform independent and high level programming language. It is used in many areas such as web and internet development, Scientific and numeric, software development etc. Python is powerful modern computer programming language. Python allow us to use variables without declaring them (i.e. it determines types simplicitly), and relies on indentation as a control structure. We are not forced to define class in python (unlike java) but we are free to do so when convenient. Python is Just a Scripting Language. Scripts written in Python (.PY files) can be parsed and run immediately. They can also be saved as a compiled programs (.PYC files), which are often used as programming modules that can be referenced by other Python programs.

Python is much simpler to use and more compact programming language. When reusing an old variable in python programming language, it becomes easier to learn and more forgiving. We need to write fewer lines code in python than in java and python code is bit easier to read and understand than java due to the removal of the braces. It is open source, which means it is free to use, even for commercial applications. Python can run on Mac, Windows, and UNIX systems and has also been ported to Java and .NET virtual machines.

5.2.1 Python IDLE

IDLE (Integrated Development Environment OR Integrated Development and Learning Environment) is an integrated development environment for python. An IDE combines a program editor and a language environment as a convenience to the programmer. Using IDLE is not a requirement for using Python. There are many other IDEs that can be used to write Python programs, not to mention a variety

of text-based programmer's editors that many programmers prefer to IDEs. We are covering IDLE because it comes with Python, and because it is not too complex for beginning programmers to use effectively. Some features of IDLE are:

- Multi-window text editor with multiple undo, Python colorizing and many other features, e.g. smart indent and call tips
- Integrated debugger with stepping, persistent breakpoints, and call stack visibility.
- Python shell with syntax highlighting.

5.2.2 Scipy

It is an open source Python library used by scientists, analysts, and engineers doing scientific computing and technical computing or SciPy is a library of algorithms and mathematical tools built to work with NumPy arrays. It adds significant power to the interactive Python session by providing the user with high-level commands and classes for manipulating and visualizing data. The additional benefit of basing SciPy on Python is that this also makes a powerful programming language available for use in developing sophisticated programs and specialized applications. Scientific applications using SciPy benefit from the development of additional modules in numerous niche's of the software landscape by developers cross the world. Everything from parallel programming to web and data-base subroutines and classes has been made available to the Python programmer. All of this power is available in addition to the mathematical libraries in SciPy. The SciPy (Scientific Python) package extends the functionality of NumPy with a substantial collection of useful algorithms, like minimization, Fourier transformation, regression, and other applied mathematical techniques.

5.2.3 NumPy

NumPy is the fundamental package needed for scientific computing with Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and basic linear algebra functions, basic Fourier transforms, sophisticated random number capabilities, tools for integrating Fortran code, tools for integrating C/C++ code, etc. Some NumPy functions are abs( ), add( ), binomial( ), floor( ), histogram( ), min( ), max( ), multipy( ) etc. NumPy arrays have a fixed size at creation, unlike Python lists (which can grow dynamically). Changing the size of an *nd array* will create a new array and delete the original. The elements in a NumPy array are all required to be of the same data type, and thus will be the same size in memory. The exception: one

can have arrays of (Python, including NumPy) objects, thereby allowing for arrays of different sized elements. Some Advantages of using Numpy with Python are :- array oriented computing, efficiently implemented multi-dimensional arrays, designed for scientific computation.

5.2.4 Python list

A list in Python is an ordered group of items (or element). It is a very general structure, and list elements don't have to be of the same type: you can put numbers, letters, strings and nested lists all on the same list. Creating a list is as simple as putting different comma-separated values between square brackets. Ex list1 = ['English', 'Computer', 1989, 2016]; list2 = [4, 5, 6, 7, 8]; list3 = ["x", "y", "z"]; It is similar to string indices, so list indices start from zero (0). While we execute above example we will get following result list1[2]; computer, list2[3]; 6. Lists are "mutable" - we *can* change an element of a list using the index operator.

CHAPTER 6
RESULT AND ANALYSIS

In this dissertation three different datasets are analyzed and three different threshold values are comparing on them with minimum validity index. From these data sets the following results analyzed.

6.1 Data set 1:

The first data set is a small one dimensional data set that generated randomly. This data set has been analyzed here as

73, 79, 70, 55, 85, 91, 17, 53, 82, 75, 89, 76, 45, 12, 76, 48, 89, 90, 87, 23

Result:

THRESHOLD = MEAN = 8.22
Intra-dist. = 32.4
Inter-dist. = 625
Validity-index = 0.05184
Cluster = 3
[[[0, 2], [0, 9], [1, 9], [1, 8], [4, 8], [4, 5]], [[3, 7]], [[6]]]

THRESHOLD = MEAN + SD = 18.76
Intra-dist. = 132.2
Inter-dist. = 3136
Validity-Index = 0.0421556122449
Cluster = 2
[[[0, 2], [0, 9], [1, 9], [1, 8], [2, 3], [3, 7], [4, 8], [4, 5]], [[6]]]

THRESHOLD = SD= 10.53
Intra-dist. = 32.4
Inter-dist. = 625
Validity-index = 0.05184

Cluster=3

[[[0, 2], [0, 9], [1, 9], [1, 8], [4, 8], [4, 5]], [[3, 7]], [[6]]]

Data set-1 has been analyzed with different threshold values using MST based clustering algorithm then got different validity index and clusters given below in table and chart

Data Set-1 (1D)		
Threshold	V.I	Clusters
Mean	0.052	3
Mean + SD	0.043	2
SD	0.052	3

Table1:- V.I and Clusters with different threshold values for data set1

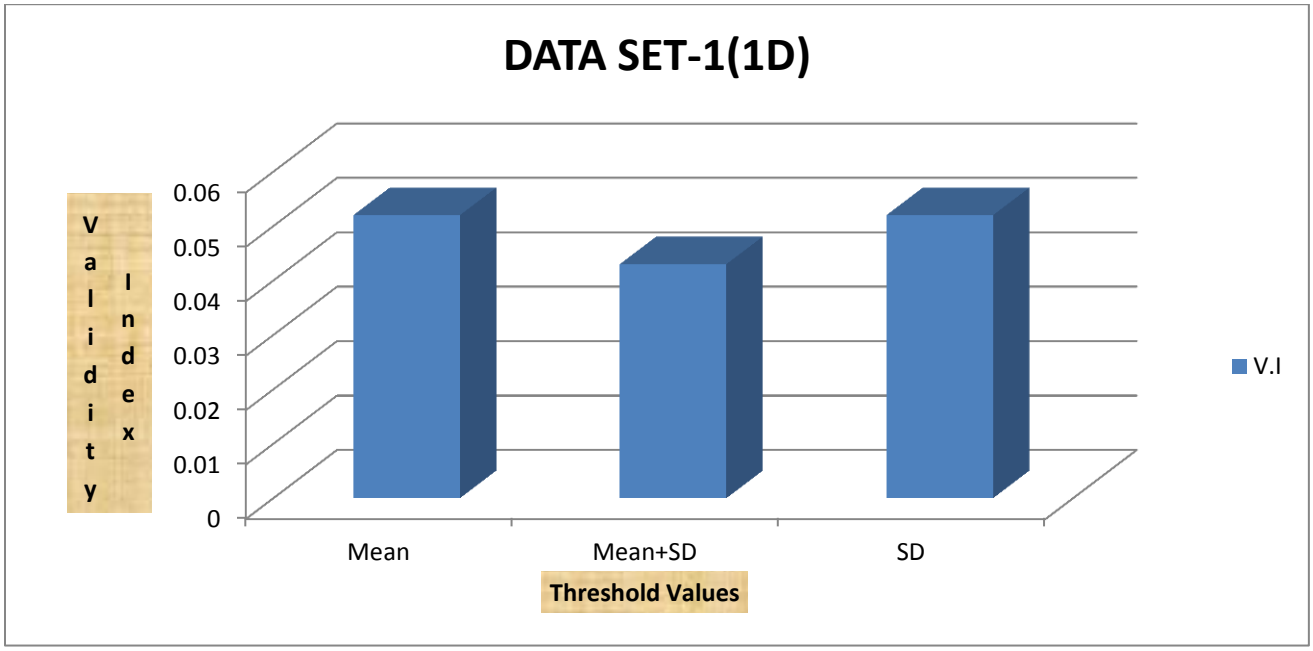

Figure3:- Validity Index vs. Threshold values for data set 1

In this data set, the above chart shows that Mean + SD threshold value has minimum validity index and SD and Mean has little bit more validity index. So minimum validity index determined Mean + SD is best threshold value for best clusters.

6.2 Data set 2:

The second data set is a small five dimensional data set that generated randomly. This data set has been analyzed here as

45.2571542991, 77.4854606536, 9.65371374137, 18.2802131148, 3.13685240185
65.0224429342, 84.7389865208, 81.6857637986, 92.3864062492, 23.2150722587
80.1341370612, 33.4238362643, 38.292428333, 19.3795246702, 30.0947291929
58.5047391778, 37.9573540228, 29.4247823751, 15.5483348045, 74.1001301864
1.73581883567, 60.7975196477, 77.1649146561, 87.521924627, 55.917463708
21.1090134421, 72.7241684788, 3.9424565675, 30.6654195217, 90.935923934
32.0407369474, 11.4068553545, 63.4629376047, 36.7072268219, 8.00891283714
30.5579786207, 19.7190017753, 2.99669522221, 36.8858922173, 94.9991817563
13.5089260328, 56.106005863, 83.439766687, 27.8299738515, 49.2179675908
41.4147374186, 78.1938121888, 7.21270984568, 15.1542483817, 57.013275149
5.1108538842, 39.8444720162, 34.5303732631, 2.49346649999, 67.2674293214
67.7617560559, 55.8905262323, 27.3136038489, 49.5804563057, 55.9444463227
26.925360641, 94.7237840648, 72.9556761117, 92.3821029339, 57.9389650353
55.7007196683, 22.3911503332, 99.7693779043, 57.258955522, 12.5430780541
12.8480270788, 51.4445576384, 99.6585807299, 6.15648820579, 99.8561590916
40.0039540264, 38.0382852169, 45.7636800291, 86.2523012237, 42.2017697873
38.9786620646, 63.8577863191, 13.9223780627, 24.9177642668, 0.868626399469
97.1404807328, 57.8421015679, 55.3186177205, 26.6886265221, 30.2694046462
5.96315320517, 59.6518441084, 10.306024929, 44.1064999668, 25.6651636797
76.4098701163, 16.9298019512, 53.2820757921, 57.4930800668, 26.8914710366
70.7266146882, 49.6145084837, 14.3528402067, 75.2198167165, 76.8321187676
39.6663269583, 68.9265114871, 87.9494346782, 76.0152261049, 29.224338398
81.0069747134, 96.0111510384, 38.6195061004, 49.6160498712, 38.3483477357
90.1337476285, 75.0508052054, 90.9066945015, 52.7580308106, 22.2241817706
95.3014786282, 57.1612075092, 13.5150321818, 21.4591918172, 34.5389634966

45.4229963227, 77.0921575139, 46.0138745102, 57.2437996565, 14.6628007439
81.0830374821, 22.0497255575, 54.1605280242, 69.0894558296, 7.02583573088
61.6039709996, 23.8815604467, 85.7959732954, 64.6724790426, 37.4084158202
9.82397437673, 33.1375719392, 13.3935423693, 57.5213889397, 3.94180413288
72.8568546885, 29.7273666838, 67.816794276, 31.5573212694, 22.3149626623
73.4656067753, 76.556790676, 15.7717497168, 92.0960507308, 97.0399662992
11.4348871512, 86.5427702823, 89.8033494631, 72.0492276432, 75.3513913427
32.9482485083, 93.838564136, 33.9395342989, 50.5661560732, 76.9628584469
43.8237526308, 57.2961626484, 11.5098587781, 40.9397069916, 98.5860013793
0.0824530760281, 15.5825984111,1 0.5058498744, 59.4405320005, 17.0273846257
74.0128902269, 80.0653694723, 30.9564128615, 70.8971810577, 91.4213708476
48.903069984, 48.4279016534, 26.7433141733, 77.4539135205, 20.0536179934
13.9258405156, 50.2826575812, 30.7773985483, 90.1606815479, 11.0074060531
44.723214047, 95.1438273244, 13.822610025, 8.01739364503, 20.249281278
7.30805432762, 77.7423576197, 85.6523729356, 80.0186409148, 59.5467944027
25.2565883964, 32.7484790912, 42.9421758148, 53.1342224184, 70.834076334
14.3909981456, 71.4579196267, 45.2609190775, 88.0306252086, 53.2534351416
27.3046641522, 86.3064097037, 90.4580175006, 58.8707010559, 80.7970859373
55.7715801824, 17.1270344464, 28.7946045393, 93.8265339913, 55.1214036517
99.3271911386, 26.1552821844, 96.7180230297, 82.7027667642, 57.4227555153
87.0514112655, 42.1545974513, 78.6563723286, 46.7131577203, 97.2724818187
34.4730720259, 0.19754539166, 15.3350574053, 23.9630029574, 54.7298290903
1.12704828392, 44.7961417141, 50.1062089757, 11.1103853457, 19.4936830279
8.34692877534, 82.0383640848, 86.5545981263, 34.0673448713, 37.5698534796
70.3719888869, 87.4768910441, 22.463632842, 63.2811929943, 96.6251957206

Result

THRESHOLD=STANDARD-DEVIATION=2.03347072179

Intra Distance= 2.18

Inter Distance= 9

Validity Index= 0.242222222222

CLUSTERS=18

[[[0, 4]], [[1, 22], [1, 48], [5, 22], [5, 31], [19, 48], [19, 42], [38, 42], [27, 38]], [[2, 13], [13, 24], [15, 24], [11, 15], [11, 33], [33, 37], [12, 37]], [[7, 36], [36, 43], [43, 46]], [[8, 47]], [[9, 45], [10, 45], [10, 28]], [[16, 30], [30, 49], [17, 49]], [[18, 44], [34, 44]], [[20, 23]], [[21, 25], [25, 35], [32, 35]], [[41], [41]], [[3], [3]], [[39], [39]], [[6], [6]], [[26], [26]], [[14], [14]], [[40], [40]], [[29], [29]]]

THRESHOLD: MEAN+SD=4.06467660069

Intra Distance= 31.68

Inter Distance= 144

Validity Index= 0.22

CLUSTERS: =6

[[[0, 4], [0, 18], [18, 44], [34, 44]], [[1, 22], [1, 48], [5, 22], [5, 31], [19, 48], [19, 42], [38, 42], [27, 38], [6, 27], [6, 20], [20, 23], [23, 40], [16, 40], [16, 30], [30, 49], [17, 49]], [[2, 13], [2, 41], [13, 24], [15, 24], [11, 15], [11, 33], [29, 41], [33, 37], [12, 37], [8, 12], [8, 47]], [[7, 26], [7, 36], [14, 26], [36, 43], [43, 46], [9, 46], [9, 45], [10, 45], [10, 28], [28, 39]], [[21, 25], [25, 35], [32, 35]], [[3], [3]]]

THRESHOLD= MEAN=2.0312058789

Intra Distance= 2.14

Inter Distance= 4

Validity Index= 0.535

CLUSTERS:=19

[[[0, 4]], [[1, 22], [1, 48], [5, 22], [5, 31], [19, 48], [19, 42], [38, 42], [27, 38]], [[2, 13], [13, 24], [15, 24], [11, 15], [11, 33], [33, 37], [12, 37]], [[7, 36], [36, 43], [43, 46]], [[8, 47]], [[9, 45], [10, 45], [10, 28]], [[16, 30], [30, 49], [17, 49]], [[18, 44], [34, 44]], [[21, 25], [25, 35], [32, 35]], [[41], [41]], [[3], [3]], [[39], [39]], [[6], [6]], [[20], [20]], [[26], [26]], [[14], [14]], [[40], [40]], [[23], [23]], [[29], [29]]]

Data Set-2 has been analyzed with different threshold values using MST based clustering algorithm and got different validity index and clusters given below in table and chart.

Data Set-2(5D)		
Threshold	V.I	Clusters
Mean	0.535	19
Mean+SD	0.22	6
SD	0.243	18

Table 2:-V.I and Clusters with different threshold values for data set 2

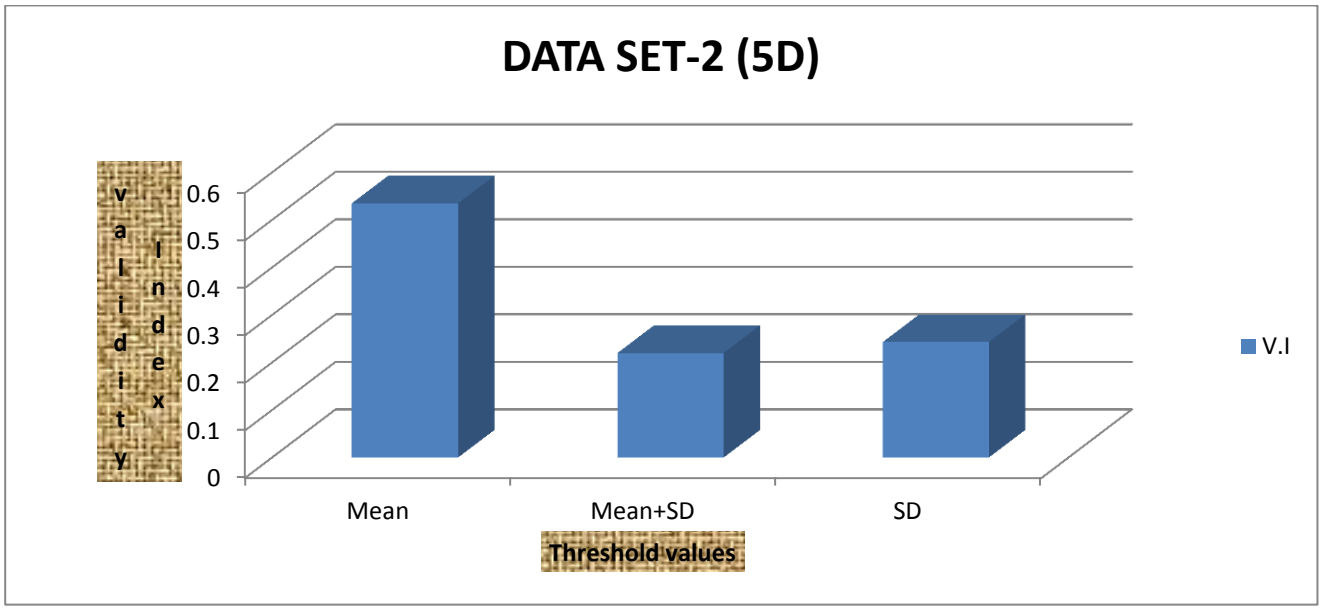

Figure 4:- Validity Index vs. Threshold values for data set 2

In this data set, the above chart shows that mean + SD threshold value has minimum validity index and SD has little bit more validity index and MEAN has more than others. So minimum validity index determined Mean + SD is best threshold value for best clusters.

6.3 Data set 3:

The third data set is multi-dimensional medium sized data set that is WINE data set from UCI machine learning repository having 178 numbers of instances and 13 numbers of attributes.

14.23,1.71,2.43,15.6,127,2.8,3.06,.28,2.29,5.64,1.04,3.92,1065

13.2,1.78,2.14,11.2,100,2.65,2.76,.26,1.28,4.38,1.05,3.4,1050

13.16,2.36,2.67,18.6,101,2.8,3.24,.3,2.81,5.68,1.03,3.17,1185
14.37,1.95,2.5,16.8,113,3.85,3.49,.24,2.18,7.8,.86,3.45,1480
13.24,2.59,2.87,21,118,2.8,2.69,.39,1.82,4.32,1.04,2.93,735
14.2,1.76,2.45,15.2,112,3.27,3.39,.34,1.97,6.75,1.05,2.85,1450
14.39,1.87,2.45,14.6,96,2.5,2.52,.3,1.98,5.25,1.02,3.58,1290
14.06,2.15,2.61,17.6,121,2.6,2.51,.31,1.25,5.05,1.06,3.58,1295
14.83,1.64,2.17,14,97,2.8,2.98,.29,1.98,5.2,1.08,2.85,1045
13.86,1.35,2.27,16,98,2.98,3.15,.22,1.85,7.22,1.01,3.55,1045
14.1,2.16,2.3,18,105,2.95,3.32,.22,2.38,5.75,1.25,3.17,1510
14.12,1.48,2.32,16.8,95,2.2,2.43,.26,1.57,5,1.17,2.82,1280
13.75,1.73,2.41,16,89,2.6,2.76,.29,1.81,5.6,1.15,2.9,1320
14.75,1.73,2.39,11.4,91,3.1,3.69,.43,2.81,5.4,1.25,2.73,1150
14.38,1.87,2.38,12,102,3.3,3.64,.29,2.96,7.5,1.2,3,1547
13.63,1.81,2.7,17.2,112,2.85,2.91,.3,1.46,7.3,1.28,2.88,1310
14.3,1.92,2.72,20,120,2.8,3.14,.33,1.97,6.2,1.07,2.65,1280
13.83,1.57,2.62,20,115,2.95,3.4,.4,1.72,6.6,1.13,2.57,1130
14.19,1.59,2.48,16.5,108,3.3,3.93,.32,1.86,8.7,1.23,2.82,1680
13.64,3.1,2.56,15.2,116,2.7,3.03,.17,1.66,5.1,.96,3.36,845
14.06,1.63,2.28,16,126,3,3.17,.24,2.1,5.65,1.09,3.71,780
12.93,3.8,2.65,18.6,102,2.41,2.41,.25,1.98,4.5,1.03,3.52,770
13.71,1.86,2.36,16.6,101,2.61,2.88,.27,1.69,3.8,1.11,4,1035
12.85,1.6,2.52,17.8,95,2.48,2.37,.26,1.46,3.93,1.09,3.63,1015
13.5,1.81,2.61,20,96,2.53,2.61,.28,1.66,3.52,1.12,3.82,845
13.05,2.05,3.22,25,124,2.63,2.68,.47,1.92,3.58,1.13,3.2,830
13.39,1.77,2.62,16.1,93,2.85,2.94,.34,1.45,4.8,.92,3.22,1195
13.3,1.72,2.14,17,94,2.4,2.19,.27,1.35,3.95,1.02,2.77,1285
13.87,1.9,2.8,19.4,107,2.95,2.97,.37,1.76,4.5,1.25,3.4,915
14.02,1.68,2.21,16,96,2.65,2.33,.26,1.98,4.7,1.04,3.59,1035
13.73,1.5,2.7,22.5,101,3,3.25,.29,2.38,5.7,1.19,2.71,1285
13.58,1.66,2.36,19.1,106,2.86,3.19,.22,1.95,6.9,1.09,2.88,1515
13.68,1.83,2.36,17.2,104,2.42,2.69,.42,1.97,3.84,1.23,2.87,990
13.76,1.53,2.7,19.5,132,2.95,2.74,.5,1.35,5.4,1.25,3,1235

13.51,1.8,2.65,19,110,2.35,2.53,.29,1.54,4.2,1.1,2.87,1095
13.48,1.81,2.41,20.5,100,2.7,2.98,.26,1.86,5.1,1.04,3.47,920
13.28,1.64,2.84,15.5,110,2.6,2.68,.34,1.36,4.6,1.09,2.78,880
13.05,1.65,2.55,18,98,2.45,2.43,.29,1.44,4.25,1.12,2.51,1105
13.07,1.5,2.1,15.5,98,2.4,2.64,.28,1.37,3.7,1.18,2.69,1020
14.22,3.99,2.51,13.2,128,3,3.04,.2,2.08,5.1,.89,3.53,760
13.56,1.71,2.31,16.2,117,3.15,3.29,.34,2.34,6.13,.95,3.38,795
13.41,3.84,2.12,18.8,90,2.45,2.68,.27,1.48,4.28,.91,3,1035
13.88,1.89,2.59,15,101,3.25,3.56,.17,1.7,5.43,.88,3.56,1095
13.24,3.98,2.29,17.5,103,2.64,2.63,.32,1.66,4.36,.82,3,680
13.05,1.77,2.1,17,107,3,3,.28,2.03,5.04,.88,3.35,885
14.21,4.04,2.44,18.9,111,2.85,2.65,.3,1.25,5.24,.87,3.33,1080
14.38,3.59,2.28,16,102,3.25,3.17,.27,2.19,4.9,1.04,3.44,1065
13.9,1.68,2.12,16,101,3.1,3.39,.21,2.14,6.1,.91,3.33,985
14.1,2.02,2.4,18.8,103,2.75,2.92,.32,2.38,6.2,1.07,2.75,1060
13.94,1.73,2.27,17.4,108,2.88,3.54,.32,2.08,8.90,1.12,3.1,1260
13.05,1.73,2.04,12.4,92,2.72,3.27,.17,2.91,7.2,1.12,2.91,1150
13.83,1.65,2.6,17.2,94,2.45,2.99,.22,2.29,5.6,1.24,3.37,1265
13.82,1.75,2.42,14,111,3.88,3.74,.32,1.87,7.05,1.01,3.26,1190
13.77,1.9,2.68,17.1,115,3,2.79,.39,1.68,6.3,1.13,2.93,1375
13.74,1.67,2.25,16.4,118,2.6,2.9,.21,1.62,5.85,.92,3.2,1060
13.56,1.73,2.46,20.5,116,2.96,2.78,.2,2.45,6.25,.98,3.03,1120
14.22,1.7,2.3,16.3,118,3.2,3,.26,2.03,6.38,.94,3.31,970
13.29,1.97,2.68,16.8,102,3,3.23,.31,1.66,6,1.07,2.84,1270
13.72,1.43,2.5,16.7,108,3.4,3.67,.19,2.04,6.8,.89,2.87,1285
12.37,.94,1.36,10.6,88,1.98,.57,.28,.42,1.95,1.05,1.82,520
12.33,1.1,2.28,16,101,2.05,1.09,.63,.41,3.27,1.25,1.67,680
12.64,1.36,2.02,16.8,100,2.02,1.41,.53,.62,5.75,.98,1.59,450
13.67,1.25,1.92,18,94,2.1,1.79,.32,.73,3.8,1.23,2.46,630
12.37,1.13,2.16,19,87,3.5,3.1,.19,1.87,4.45,1.22,2.87,420
12.17,1.45,2.53,19,104,1.89,1.75,.45,1.03,2.95,1.45,2.23,355
12.37,1.21,2.56,18.1,98,2.42,2.65,.37,2.08,4.6,1.19,2.3,678

13.11,1.01,1.7,15,78,2.98,3.18,.26,2.28,5.3,1.12,3.18,502
12.37,1.17,1.92,19.6,78,2.11,2,.27,1.04,4.68,1.12,3.48,510
13.34,.94,2.36,17,110,2.53,1.3,.55,.42,3.17,1.02,1.93,750
12.21,1.19,1.75,16.8,151,1.85,1.28,.14,2.5,2.85,1.28,3.07,718
12.29,1.61,2.21,20.4,103,1.1,1.02,.37,1.46,3.05,.906,1.82,870
13.86,1.51,2.67,25,86,2.95,2.86,.21,1.87,3.38,1.36,3.16,410
13.49,1.66,2.24,24,87,1.88,1.84,.27,1.03,3.74,.98,2.78,472
12.99,1.67,2.6,30,139,3.3,2.89,.21,1.96,3.35,1.31,3.5,985
11.96,1.09,2.3,21,101,3.38,2.14,.13,1.65,3.21,.99,3.13,886
11.66,1.88,1.92,16,97,1.61,1.57,.34,1.15,3.8,1.23,2.14,428
13.03,.9,1.71,16,86,1.95,2.03,.24,1.46,4.6,1.19,2.48,392
11.84,2.89,2.23,18,112,1.72,1.32,.43,.95,2.65,.96,2.52,500
12.33,.99,1.95,14.8,136,1.9,1.85,.35,2.76,3.4,1.06,2.31,750
12.7,3.87,2.4,23,101,2.83,2.55,.43,1.95,2.57,1.19,3.13,463
12,.92,2,19,86,2.42,2.26,.3,1.43,2.5,1.38,3.12,278
12.72,1.81,2.2,18.8,86,2.2,2.53,.26,1.77,3.9,1.16,3.14,714
12.08,1.13,2.51,24,78,2,1.58,.4,1.4,2.2,1.31,2.72,630
13.05,3.86,2.32,22.5,85,1.65,1.59,.61,1.62,4.8,.84,2.01,515
11.84,.89,2.58,18,94,2.2,2.21,.22,2.35,3.05,.79,3.08,520
12.67,.98,2.24,18,99,2.2,1.94,.3,1.46,2.62,1.23,3.16,450
12.16,1.61,2.31,22.8,90,1.78,1.69,.43,1.56,2.45,1.33,2.26,495
11.65,1.67,2.62,26,88,1.92,1.61,.4,1.34,2.6,1.36,3.21,562
11.64,2.06,2.46,21.6,84,1.95,1.69,.48,1.35,2.8,1,2.75,680
12.08,1.33,2.3,23.6,70,2.2,1.59,.42,1.38,1.74,1.07,3.21,625
12.08,1.83,2.32,18.5,81,1.6,1.5,.52,1.64,2.4,1.08,2.27,480
12,1.51,2.42,22,86,1.45,1.25,.5,1.63,3.6,1.05,2.65,450
12.69,1.53,2.26,20.7,80,1.38,1.46,.58,1.62,3.05,.96,2.06,495
12.29,2.83,2.22,18,88,2.45,2.25,.25,1.99,2.15,1.15,3.3,290
11.62,1.99,2.28,18,98,3.02,2.26,.17,1.35,3.25,1.16,2.96,345
12.47,1.52,2.2,19,162,2.5,2.27,.32,3.28,2.6,1.16,2.63,937
11.81,2.12,2.74,21.5,134,1.6,.99,.14,1.56,2.5,.95,2.26,625
12.29,1.41,1.98,16,85,2.55,2.5,.29,1.77,2.9,1.23,2.74,428

12.37,1.07,2.1,18.5,88,3.52,3.75,.24,1.95,4.5,1.04,2.77,660
12.29,3.17,2.21,18,88,2.85,2.99,.45,2.81,2.3,1.42,2.83,406
12.08,2.08,1.7,17.5,97,2.23,2.17,.26,1.4,3.3,1.27,2.96,710
12.6,1.34,1.9,18.5,88,1.45,1.36,.29,1.35,2.45,1.04,2.77,562
12.34,2.45,2.46,21,98,2.56,2.11,.34,1.31,2.8,.8,3.38,438
11.82,1.72,1.88,19.5,86,2.5,1.64,.37,1.42,2.06,.94,2.44,415
12.51,1.73,1.98,20.5,85,2.2,1.92,.32,1.48,2.94,1.04,3.57,672
12.42,2.55,2.27,22,90,1.68,1.84,.66,1.42,2.7,.86,3.3,315
12.25,1.73,2.12,19,80,1.65,2.03,.37,1.63,3.4,1,3.17,510
12.72,1.75,2.28,22.5,84,1.38,1.76,.48,1.63,3.3,.88,2.42,488
12.22,1.29,1.94,19,92,2.36,2.04,.39,2.08,2.7,.86,3.02,312
11.61,1.35,2.7,20,94,2.74,2.92,.29,2.49,2.65,.96,3.26,680
11.46,3.74,1.82,19.5,107,3.18,2.58,.24,3.58,2.9,.75,2.81,562
12.52,2.43,2.17,21,88,2.55,2.27,.26,1.22,2,.9,2.78,325
11.76,2.68,2.92,20,103,1.75,2.03,.6,1.05,3.8,1.23,2.5,607
11.41,.74,2.5,21,88,2.48,2.01,.42,1.44,3.08,1.1,2.31,434
12.08,1.39,2.5,22.5,84,2.56,2.29,.43,1.04,2.9,.93,3.19,385
11.03,1.51,2.2,21.5,85,2.46,2.17,.52,2.01,1.9,1.71,2.87,407
11.82,1.47,1.99,20.8,86,1.98,1.6,.3,1.53,1.95,.95,3.33,495
12.42,1.61,2.19,22.5,108,2,2.09,.34,1.61,2.06,1.06,2.96,345
12.77,3.43,1.98,16,80,1.63,1.25,.43,.83,3.4,.7,2.12,372
12,3.43,2,19,87,2,1.64,.37,1.87,1.28,.93,3.05,564
11.45,2.4,2.42,20,96,2.9,2.79,.32,1.83,3.25,.8,3.39,625
11.56,2.05,3.23,28.5,119,3.18,5.08,.47,1.87,6,.93,3.69,465
12.42,4.43,2.73,26.5,102,2.2,2.13,.43,1.71,2.08,.92,3.12,365
13.05,5.8,2.13,21.5,86,2.62,2.65,.3,2.01,2.6,.73,3.1,380
11.87,4.31,2.39,21,82,2.86,3.03,.21,2.91,2.8,.75,3.64,380
12.07,2.16,2.17,21,85,2.6,2.65,.37,1.35,2.76,.86,3.28,378
12.43,1.53,2.29,21.5,86,2.74,3.15,.39,1.77,3.94,.69,2.84,352
11.79,2.13,2.78,28.5,92,2.13,2.24,.58,1.76,3,.97,2.44,466
12.37,1.63,2.3,24.5,88,2.22,2.45,.4,1.9,2.12,.89,2.78,342
12.04,4.3,2.38,22,80,2.1,1.75,.42,1.35,2.6,.79,2.57,580

12.86,1.35,2.32,18,122,1.51,1.25,.21,.94,4.1,.76,1.29,630
12.88,2.99,2.4,20,104,1.3,1.22,.24,.83,5.4,.74,1.42,530
12.81,2.31,2.4,24,98,1.15,1.09,.27,.83,5.7,.66,1.36,560
12.7,3.55,2.36,21.5,106,1.7,1.2,.17,.84,5,.78,1.29,600
12.51,1.24,2.25,17.5,85,2,.58,.6,1.25,5.45,.75,1.51,650
12.6,2.46,2.2,18.5,94,1.62,.66,.63,.94,7.1,.73,1.58,695
12.25,4.72,2.54,21,89,1.38,.47,.53,.8,3.85,.75,1.27,720
12.53,5.51,2.64,25,96,1.79,.6,.63,1.1,5,.82,1.69,515
13.49,3.59,2.19,19.5,88,1.62,.48,.58,.88,5.7,.81,1.82,580
12.84,2.96,2.61,24,101,2.32,.6,.53,.81,4.92,.89,2.15,590
12.93,2.81,2.7,21,96,1.54,.5,.53,.75,4.6,.77,2.31,600
13.36,2.56,2.35,20,89,1.4,.5,.37,.64,5.6,.7,2.47,780
13.52,3.17,2.72,23.5,97,1.55,.52,.5,.55,4.35,.89,2.06,520
13.62,4.95,2.35,20,92,2,.8,.47,1.02,4.4,.91,2.05,550
12.25,3.88,2.2,18.5,112,1.38,.78,.29,1.14,8.21,.65,2,855
13.16,3.57,2.15,21,102,1.5,.55,.43,1.3,4,.6,1.68,830
13.88,5.04,2.23,20,80,.98,.34,.4,.68,4.9,.58,1.33,415
12.87,4.61,2.48,21.5,86,1.7,.65,.47,.86,7.65,.54,1.86,625
13.32,3.24,2.38,21.5,92,1.93,.76,.45,1.25,8.42,.55,1.62,650
13.08,3.9,2.36,21.5,113,1.41,1.39,.34,1.14,9.40,.57,1.33,550
13.5,3.12,2.62,24,123,1.4,1.57,.22,1.25,8.60,.59,1.3,500
12.79,2.67,2.48,22,112,1.48,1.36,.24,1.26,10.8,.48,1.47,480
13.11,1.9,2.75,25.5,116,2.2,1.28,.26,1.56,7.1,.61,1.33,425
13.23,3.3,2.28,18.5,98,1.8,.83,.61,1.87,10.52,.56,1.51,675
12.58,1.29,2.1,20,103,1.48,.58,.53,1.4,7.6,.58,1.55,640
13.17,5.19,2.32,22,93,1.74,.63,.61,1.55,7.9,.6,1.48,725
13.84,4.12,2.38,19.5,89,1.8,.83,.48,1.56,9.01,.57,1.64,480
12.45,3.03,2.64,27,97,1.9,.58,.63,1.14,7.5,.67,1.73,880
14.34,1.68,2.7,25,98,2.8,1.31,.53,2.7,13,.57,1.96,660
13.48,1.67,2.64,22.5,89,2.6,1.1,.52,2.29,11.75,.57,1.78,620
12.36,3.83,2.38,21,88,2.3,.92,.5,1.04,7.65,.56,1.58,520
13.69,3.26,2.54,20,107,1.83,.56,.5,.8,5.88,.96,1.82,680

12.85,3.27,2.58,22,106,1.65,.6,.6,.96,5.58,.87,2.11,570

12.96,3.45,2.35,18.5,106,1.39,.7,.4,.94,5.28,.68,1.75,675

13.78,2.76,2.3,22,90,1.35,.68,.41,1.03,9.58,.7,1.68,615

13.73,4.36,2.26,22.5,88,1.28,.47,.52,1.15,6.62,.78,1.75,520

13.45,3.7,2.6,23,111,1.7,.92,.43,1.46,10.68,.85,1.56,695

12.82,3.37,2.3,19.5,88,1.48,.66,.4,.97,10.26,.72,1.75,685

13.58,2.58,2.69,24.5,105,1.55,.84,.39,1.54,8.66,.74,1.8,750

13.4,4.6,2.86,25,112,1.98,.96,.27,1.11,8.5,.67,1.92,630

12.2,3.03,2.32,19,96,1.25,.49,.4,.73,5.5,.66,1.83,510

12.77,2.39,2.28,19.5,86,1.39,.51,.48,.64,9.899999,.57,1.63,470

14.16,2.51,2.48,20,91,1.68,.7,.44,1.24,9.7,.62,1.71,660

13.71,5.65,2.45,20.5,95,1.68,.61,.52,1.06,7.7,.64,1.74,740

13.4,3.91,2.48,23,102,1.8,.75,.43,1.41,7.3,.7,1.56,750

13.27,4.28,2.26,20,120,1.59,.69,.43,1.35,10.2,.59,1.56,835

13.17,2.59,2.37,20,120,1.65,.68,.53,1.46,9.3,.6,1.62,840

14.13,4.1,2.74,24.5,96,2.05,.76,.56,1.35,9.2,.61,1.6,560

Result

THRESHOLD = MEAN=14.4545515812

Intra-Dist. = 0.47191011236

Inter-Dist. = 1

Validity Index = 0.47191011236

Cluster = 53

[[[0, 54]], [[1, 8], [1, 48], [8, 9], [8, 29], [22, 29], [29, 41], [46, 48]], [[2, 26], [2, 52]], [[6, 27], [11, 27], [11, 57], [27, 30], [30, 58], [16, 58], [49, 57], [51, 57]], [[10, 31]], [[13, 50]], [[17, 55]], [[19, 144], [19, 176], [175, 176], [25, 175]], [[23, 38]], [[28, 35]], [[32, 47]], [[34, 42], [37, 42]], [[36, 44], [36, 70], [44, 74], [74, 157]], [[39, 78]], [[43, 60], [43, 161], [60, 65], [65, 109], [65, 153], [88, 109], [88, 104], [88, 167], [98, 104], [98, 172], [135, 167], [148, 172], [134, 148], [158, 172], [161, 163]], [[59, 84], [84, 142], [84, 160], [131, 142], [137, 142], [137, 170], [160, 165], [83, 165], [83, 106], [67, 106], [66, 67], [66, 92], [92, 116], [86, 116], [107, 116], [90, 107], [90, 156], [156, 171], [72, 171], [72, 127], [79, 127], [79, 85], [61, 85], [85, 91], [85, 102], [75, 102], [102, 113], [97, 113], [63, 97], [63, 103], [71,

103], [71, 115], [99, 115], [103, 146]], [[62, 120], [62, 147], [62, 154], [82, 147], [82, 89], [147, 159], [159, 164]], [[64, 117], [64, 122], [94, 117], [94, 128], [126, 128]], [[68, 168], [168, 174], [173, 174]], [[76, 114], [114, 124], [124, 125], [118, 125], [123, 125]], [[77, 150]], [[80, 93]], [[81, 100], [81, 136], [136, 155]], [[87, 101], [87, 132], [101, 119], [110, 132], [110, 162], [132, 177], [143, 177]], [[96, 130], [130, 169]], [[105, 108], [105, 111]], [[112, 133], [133, 139], [133, 140]], [[129, 138]], [[3], [3]], [[5], [5]], [[4]], [[53], [53]], [[7], [7]], [[15], [15]], [[12], [12]], [[14], [14]], [[18]], [[20], [20]], [[40], [40]], [[21], [21]], [[141]], [[24]], [[145], [145]], [[33], [33]], [[45], [45]], [[56], [56], [56]], [[73], [73]], [[69]], [[95]], [[152]], [[121], [121]], [[149], [149]], [[151]], [[166]]]

THRESHOLD= MEAN + SD = 28.2227712938

Intra-Dist. = 0.432584269663

Inter-Dist. = 1

Validity Index= 0.432584269663

Cluster:-17

[[[0, 54], [48, 54], [1, 48], [1, 8], [8, 9], [8, 29], [22, 29], [29, 38], [23, 38], [23, 32], [29, 41], [32, 47], [46, 48], [45, 46], [34, 45], [34, 42], [37, 42], [37, 55], [17, 55], [47, 56]], [[2, 26], [2, 52]], [[4, 68], [39, 68], [20, 39], [20, 40], [39, 78], [68, 168], [168, 174], [21, 174], [21, 141], [173, 174], [155, 173], [136, 155], [81, 136], [81, 100], [100, 135], [135, 167], [88, 167], [88, 104], [88, 109], [65, 109], [60, 65], [43, 60], [43, 161], [65, 153], [98, 104], [98, 172], [148, 172], [134, 148], [148, 154], [62, 154], [62, 120], [62, 147], [82, 147], [82, 89], [147, 159], [154, 169], [130, 169], [96, 130], [158, 172], [159, 164], [112, 164], [112, 133], [133, 139], [133, 140], [138, 139], [119, 138], [101, 119], [87, 101], [87, 132], [110, 132], [110, 149], [110, 162], [129, 138], [131, 149], [131, 142], [84, 142], [59, 84], [84, 160], [132, 177], [137, 142], [137, 170], [77, 170], [77, 150], [143, 177], [160, 165], [83, 165], [83, 106], [67, 106], [66, 67], [66, 92], [92, 116], [86, 116], [107, 116], [90, 107], [90, 156], [156, 171], [72, 171], [72, 127], [79, 127], [79, 85], [61, 85], [79, 121], [85, 91], [85, 102], [75, 102], [75, 152], [102, 113], [97, 113], [63, 97], [63, 103], [71, 103], [71, 115], [99, 115], [76, 99], [76, 114], [103, 146], [114, 124], [121, 151], [124, 125], [118, 125], [118, 126], [123, 125], [126, 128], [94, 128], [94, 117], [64, 117], [64, 122], [111, 128], [105, 111], [105, 108], [93, 108], [80, 93], [161, 163], [161, 166]], [[6, 27], [11, 27], [11, 57], [27, 30], [30, 58], [16, 58], [7, 16], [7, 15], [12, 15], [49, 57], [51, 57]], [[10, 31]], [[13, 50]], [[19, 144], [19, 176], [70, 144], [36, 70], [36, 44], [44, 74], [74, 157], [175, 176], [25, 175], [145, 175], [24, 145]], [[28, 35]], [[3], [3]], [[5], [5]], [[53], [53]], [[14], [14]], [[18]], [[33], [33]], [[73], [73]], [[69]], [[95]]]

THRESHOLD = SD=13.7682197126

Intra-Dist. = 0.522471910112

Inter-Dist. = 1

Validity Index = 0.522471910112

Cluster: = 56

[[[0, 54]], [[1, 8], [1, 48], [8, 9], [8, 29], [22, 29], [29, 41], [46, 48]], [[2, 26], [2, 52]], [[6, 27], [11, 27], [11, 57], [27, 30], [30, 58], [16, 58], [49, 57], [51, 57]], [[10, 31]], [[13, 50]], [[17, 55]], [[19, 144], [19, 176], [175, 176], [25, 175]], [[23, 38]], [[28, 35]], [[32, 47]], [[34, 42], [37, 42]], [[36, 44], [36, 70], [44, 74], [74, 157]], [[39, 78]], [[43, 60], [43, 161], [60, 65], [65, 109], [65, 153], [88, 109], [88, 104], [88, 167], [98, 104], [98, 172], [135, 167], [148, 172], [134, 148], [158, 172], [161, 163]], [[59, 84], [84, 142], [84, 160], [131, 142], [137, 142], [137, 170], [160, 165], [83, 165], [83, 106], [67, 106], [66, 67], [66, 92], [92, 116], [86, 116], [107, 116], [90, 107], [90, 156], [156, 171], [72, 171], [72, 127], [79, 127]], [[61, 85], [85, 91], [85, 102], [75, 102], [102, 113], [97, 113], [63, 97], [63, 103], [71, 103], [71, 115], [99, 115], [103, 146]], [[62, 120], [62, 147], [82, 147], [82, 89], [147, 159], [159, 164]], [[64, 117], [64, 122], [94, 117], [94, 128], [126, 128]], [[68, 168], [168, 174], [173, 174]], [[76, 114], [114, 124], [124, 125], [118, 125], [123, 125]], [[80, 93]], [[81, 100], [81, 136], [136, 155]], [[87, 101], [87, 132], [101, 119], [110, 132], [110, 162], [132, 177], [143, 177]], [[96, 130], [130, 169]], [[105, 108], [105, 111]], [[112, 133], [133, 139], [133, 140]], [[129, 138]], [[3], [3]], [[5], [5]], [[4]], [[53], [53]], [[7], [7]], [[15], [15]], [[12], [12]], [[14], [14]], [[18]], [[20], [20]], [[40], [40]], [[21], [21]], [[141]], [[24]], [[145], [145]], [[33], [33]], [[45], [45]], [[56], [56], [56]], [[73], [73]], [[154], [154], [154]], [[69]], [[95]], [[152]], [[77], [77]], [[150]], [[121], [121]], [[149], [149]], [[151]], [[166]]]

This data set has been analyzed with different threshold values using MST based clustering algorithm and got different validity index and clusters given below in table and charts

Data Set-3(13D)		
Threshold	V.I	Clusters
Mean	0.472	53
Mean + SD	0.433	17
SD	0.523	56

Table 3:- V.I and clusters with different threshold values for data set 3

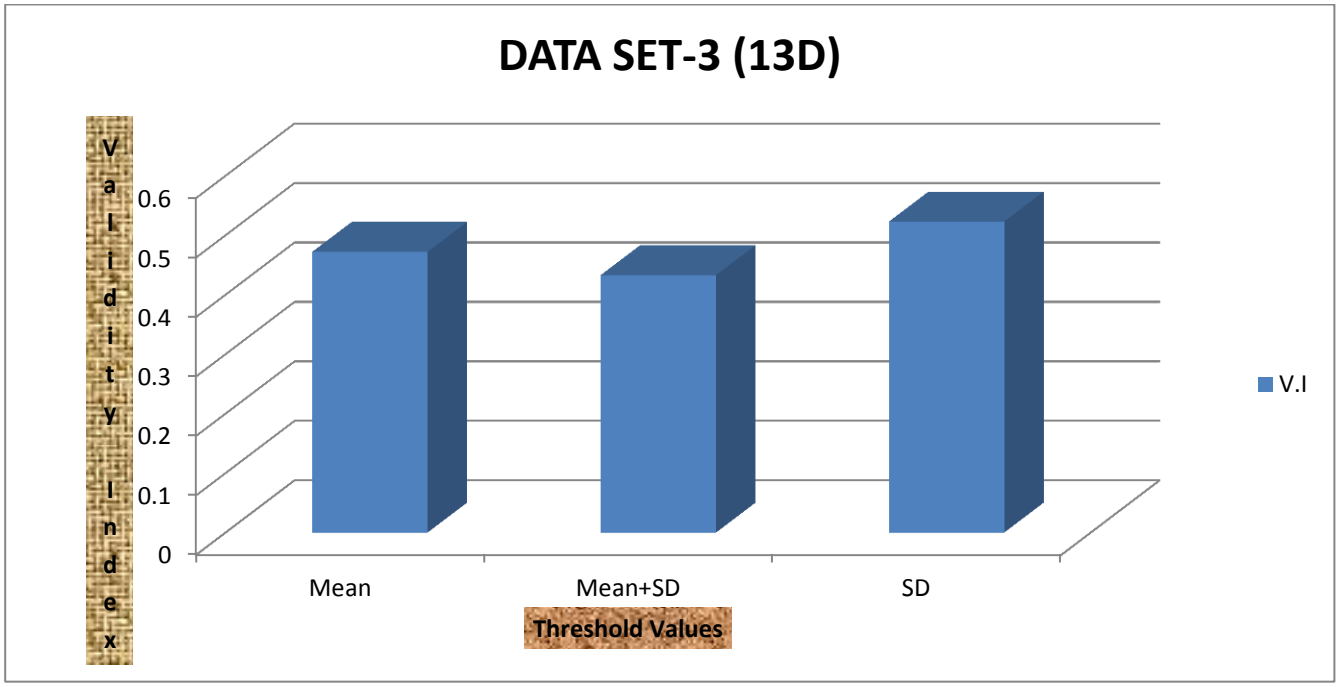

Figure 5:-Validity Index vs. Threshold values for data set 3

In this data set, the above chart shows that mean + SD threshold value has minimum validity index and Mean has little bit more validity index and SD has more than others. So minimum validity index determined Mean + SD is best threshold value for best clusters.

6.4 Result

The different data sets are evaluated by different threshold values using MST based clustering algorithm. From above table and chart we can understand that Mean + SD threshold value is best for clustering because Mean + SD threshold value has minimum validity index. If value of validity index is minimum then intra cluster distance has high similarity and inter cluster distance has low similarity. So we can say that Mean + SD threshold value is best for clustering than other threshold values.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

Prim's and MST based clustering algorithms are used in this paper. The algorithm has been tested on the randomly generated data sets and real world data sets such as Wine data set from UCI machine learning repository. The algorithm has been shown to be very effective in clustering multidimensional data sets. Threshold values plays vital role in this dissertation. Which threshold value has minimum validity index that determined best threshold value for best clusters.

In this dissertation, the different data sets are evaluated by MST based clustering algorithm using different threshold values (mean, SD, and mean + SD,) on them. The Minimum validity index is found on each data set while we applied Mean + SD threshold value on MST. Mean and SD threshold values has the little bit different validity index on each data set. The experimental results clearly show that Mean + SD threshold value is best threshold value for best clusters.

7.2 Future Work

More algorithms from the clustering can be incorporated for the future study to the studied datasets. So, we can have experiment over various values for better finding of threshold. For the limitation of Storage and Time Complexity, the same procedure can be implemented in Parallel Scenario.

References:

1. J.Han, M.Kamber, and J.Pei "Data mining: Concepts and Techniques", 3rd edition, Morgan-Kaufman publisher, 2013.
2. Xuegang Hu, Lei Li. "Improved Fuzzy C-Means Algorithm for Image Segmentation", *Journal of Electrical and Electronic Engineering*. Vol. 3, No. 1, 2015, pp. 1-5. doi: 10.11648/j.jeee.(2015) 0301.11
3. G.Kerr, H.J. Ruskin, M. Crane and P. Doolan,"Techniques for clustering gene expression data, Computers in Biology and Medicine", 38(2008) 283-293.
4. ZhihuaDu,YiweiWang, Zhen Ji, PK-means: "A new algorithm for gene clustering, Computational Biology and Chemistry", 32(2008) 243–247.
5. C. T. Zahn, "Graph-theriotical methods for detecting and describing getalt clusters", *IEEE Trans.* on Computers, 20 (1971), 68-86.
6. Y. Xu, V. Olman and D. Xu "Clustering gene expression data using a graph-theriotic approach: An application of minimum spanning trees, Bioinformatics", 18(2002) 536-545
7. T. Soni Madhulatha, "An overview on clustering methods", *IOSR Journal of engineering*, Apr . 2012, vol.2(4), pp.719-725.
8. T., Stinbach., Kumar " Data Mining Cluster Analysis: Basic concept and algorithm" Lecturer Notes for Chapter 8, 2004.
9. A.K Jain, M.N Murty, P.J. Flynn, “Data Clustering: A Review”, ACM Computing, Vol. 31, No. 3, September 1999.
10. Nabin Ghimire, "Analysis of various clustering algorithms in data mining", CDCSIT TU, Sep. 2014.
11. V.M.K. Prasad Goura, N.M.Rao, M.R.Reddy, "A Dynamic Clustering Technique using Minimum Spanning Tree", 2nd International Conference on Biotechnology and Food Science IPCBEE vol.7, 2011, Singapore.
12. B.Adepu, Kiran Kumar Bejjanki "A Novel Approach for Minimum Spanning Tree based Clustering Algorithm", Associate Professor Department of MCA, Kakatiya Institute of technology and science, Warangal, Andra Pradesh, India 506015.
13. O.Grygorash, Y.Zhou, Zach Jorgensen "Minimum Spanning Tree Based Clustering Algorithms" School of Computer and Information Sciences University of South Alabama, Mobile, AL 36688 USA

14. Xiaochum Wang, Xiali Wang and D.Mitchell Willkes, "A Divide and Conquer approach for minimum spanning tree–based clustering," *IEEE Trans*. Knowledge and Data Engg., 21(2009) 945-958.
15. Prasanta K. Jana, Azad Naik "An Efficient Minimum Spanning Tree based Clustering Algorithm," In: Proc. Of Intel. Conference on Methods and Models in Computer Science (ICM2CS09), 2009 pp 1-5
16. S.Ray, R.H.Turi, "Determination of number of cluster in K- means clustering and application in color image segmentation", Pros.4thIntel.Conf. ICAPRDT '99, Calcutta India, 1999, pp.137-143.
17. C.Zong, D.Miao, R.Wang, A graph-theoretical clustering method based on two rounds of minimum spanning trees, Pattern Recognition,43(2010), 752-766.
18. UCI Machine Leaning Repository, *https://archive.ics.uci.edu/ml/datasets.html*

Bibliography:

- K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, 1988.
- J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
- Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.
- Sholom M. Weiss and Nitin Indurkhya, "Predictive Data Mining: A Practical Guide", Morgan Kaufmann Publishers, 1998.
- J.Han, M.Kamber, and J.Pei Data mining: Concepts and Techniques, Morgan-Kaufman, 2013.
- https://archive.ics.uci.edu/ml/datasets.html.
- https://en.wikipedia.org/wiki/Minimum_spanning_tree