DR.T. KARTHIGEYAN
Research Guide and Supervisor
Associate Professor in Computer Science
PSG College of Arts and Science
Coimbatore-641 014, India.

CERTIFICATE

This thesis entitled "**MST: A CONTEMPORARY APPROACH FOR DATA MINING BASED ON CLUSTERING METHOD"** for the award of the degree of Doctor of Philosophy in Computer Science of Manonmaniam Sundaranar University is a record of bonafide research work done by him and it has not been submitted for the award of any degree, diploma, associateship or fellowship of any other University / Institution.

Place: Palayamkottai DR.T. KARTHIGEYAN

Date :

S. CHIDAMBARANATHAN
Research Scholar
Dept. of Computer Science
St. Xavier's College
Palayamkottai-627 002, Tamil Nadu
India

DECLARATION

I hereby declare that the thesis entitled "**MST: A CONTEMPORARY APPROACH FOR DATA MINING BASED ON CLUSTERING METHOD"** submitted by me for the degree of Doctor of Philosophy in Computer Science is the result of my original and independent research work carried out under the guidance of **DR.T. KARTHIGEYAN,** Associate Professor in Computer Science, PSG College of Arts and Science, Coimbatore-641 014, India, and it has not been submitted for the award of any degree, diploma, associateship or fellowship of any other University / Institution.

Place: Palayamkottai S. CHIDAMBARANATHAN

Date :

ACKNOWLEDGEMENT

First and foremost, I register my sincere thanks to my Research Guide and Supervisor **Dr.T. Karthigeyan** for his continuous support and priceless guidance in my Ph.D. Programme which is a golden opportunity that has enabled me to pursue my research. I thank him especially for his support at the early stage of this research work as well as for giving me the extra-ordinary experience of satisfaction and confidence throughout my research work.

I am very grateful to the Management of St. Xavier's College: **Rev.Dr. Gilburt S.J.,** Principal, **Rev.Fr. Jesu Michael Das S.J**., Secretary, **Rev.Dr. Danis Ponniah S.J.,** Rector, **Rev.Dr. Antonysamy S.J**., former Principal and **Rev.Dr. Selvanayagam** S.J., former Secretary. I am thankful to them for allowing me to use the Resources of the College. I have really benefitted a lot and it has helped me to complete my research work within the given time.

I would like to thank **Dr. S.P. Victor,** Director, Computer Science Research Centre, St. Xavier's College, Palayamkottai for his continuous help and encouraging words and special care during my study period.

I gratefully acknowledge **Prof.S.A. Thiaga Rajan,**(Formerly of the Dept. of English, St. Xavier's College) for his tremendous support and selfless help in every stage of preparing my thesis.

Let me express my gratitude to **Mrs.K. Sujatha,** Dept. of Comp. Science, Govt. Engg. College, Tirunelveli for her consistent help in my research. She has

helped me in solving many research oriented issues. She always encouraged me to hurry up the work and provided some research documents, that were really helpful in my research work.

I would like to show my special gratitude also to **Dr.S. John Peter**, Asst. Professor, Dept. of Comp. Science, St. Xavier's College for inspiring in me the love for science and supporting me in all my academic achievements.

I wish to thank my friends **Prof. Arockia Stephen Raj, Dr. Ashok Kumar, Prof. Reeta** and **Prof. Veniston** for their continuous help and encouragement in my research work.

I am thankful to my well wishers **Mr. Rayan, Mr. Palani, Mr. Vargeese, Mr. Antony, Mr. John, Mr. Natarajan,** and **Mr. Xavier** for their help in many ways.

Last but not least, I would like to thank the members of my family. Their constant inspiration and guidance kept me focused and motivated. My heartiest gratitude goes to my parents and I thank my wife and daughter for their love and persistent confidence in me that supported me mentally during the course of this research.

Finally, I would like to thank everybody who was important to the successful of this thesis.

S. CHIDAMBARANATHAN

CONTENTS

LIST OF ABBREVIATIONS

BF	Bacteria Foraging
CBIR	Content Based Image Retrieval
CRF	Conditional Random Field
CS	Cluster Separation
CT	Confuted Tomography
DEM	Digital Elevation Meta
DHC	Density-based Hierarchical Clustering
DHCMST	Divisive Hierarchical Clustering using Minimum Spanning Tree
DHMST	Decisive Hierarchical Minimum Spanning Tree
DIS	Difference In Strength
DM	Dissimilarity Matrix
DNA	Deoxyribonucleic Acid
DOF	Degree of Freedom
DOG	Difference-Of-Gaussian
DPSO	Darwinian Particle Swarm Optimization
EM	Expectation-Maximization
EMST	Euclidean Minimum Spanning Tree
EMST1	Euclidean Minimum Spanning Tree1
EMST2	Euclidean Minimum Spanning Tree2
FCM	Fuzzy C-Menas
FMRI	Functional Magnetic Resonance Imaging
FODPSO	Fractional-Order Darwinian Particle Swarm Optimization
GIS	Geographic Information System
HAC	Hierarchical Agglomerative Clustering
HEMST	Hierarchical Euclidean-distance based Minimum Spanning Tree
ID	Induced Dependencies
IRS	Information Retrieval System
KDD	Knowledge Discovery in Database
KFCM	Kernelized Fuzzy C-Means
LDA	Latent Direichlet Allocation
LIM	Lorenz Information Measure
MAP	Maximum Aposteriori Probability
MaRACel	Markov Random field driven region-based Active Contour model
MELC	Multi-epitode Ligand Cartography
MLE	Maximum Likelihood Examination
MRF	Markov Random Field
MRI	Magnetic Resonance Imaging
MSDR	Maximum Standard Deviation Reduction

MST	Minimum Spanning Tree
MSTODDN	Minimum Spanning Tree based Outlier Detection using Degree Number
MSTSSCIMLRO	Minimum Spanning Tree based Structural Similarity Clustering for Image Mining with Local Region Outliers
NLP	Natural Language Processing
OFP	Outlier Finding Process
PAM	Partitioning Around Medoids
PSO	Particle Swarm Optimization
QBIC	Query By Image Content
RAC	Region-based Active Contour
RF	Random Forests
SFMST	Scale Free Minimum Spanning Tree
SIFT	Scale-Invariant Feature Transform
SLDA	Spatial Latent Direichlet Allocation
STF	Semantic Texts on Forests
TIS	Toponome Imaging System
UDT	Uncertain Decision Tree
UI	Ultrasound Imaging
WIPFCM	Weighted Image Patch-based Fuzzy C-Means

CHAPTER I

1.1 INTRODUCTION

Nowadays it is clear that the world needs effective as well as efficient decisions. It has hence become vital that such decisions are made. Researchers are up to such decision making and in quest of the same, such researchers are concentrating on data mining. The data mining leads to decision making. The main goal of data mining is to discover the unknown know-how as extracted from the huge collected and stored historical data. Relying on such historical data, one can certainly initiate or take action. Such data can prompt one to take the proper initiation or actions. With such proper initiations one can accomplish decisively, for such data are the proper decisions of the great scholars and initiations based on such decisions always lead one to decisive accomplishments. Mining is strictly the knowledge mining from the stored data or the historical data and this data mining is popularly known as Knowledge Discovery in Database(KDD).

1.2 IMAGE MINING AND DATA MINING

Knowledge discovery in database is data mining and knowledge discovery in image database is image mining. In the course of image mining, the very same researchers are trying their best to extract useful and deep-set know-how or data from the already existing image database in order to take certain decisions in areas like medical diagnosis, space research, remote sensing, agriculture, industry and education.

Image mining is an expansion of data mining related to the image domain. Image mining is not only used to extract the relevant images but also used to create image patterns. The very same image mining process extracts the patterns too from the collection of stored images. The image mining process also extracts the specific features from a single image. Ji Zhang, Wynne Hsu and Mong Li Lee have urged the need for image mining in the era of the rapid growth of the amount of image data. Image mining is a modern approach in data mining. There is an increase in the number of images as well as the image databases and such an increase has rather caused a necessity for image mining techniques. Image mining process is an extended data mining process which is concerned solely with knowledge discovery out of digital images.

The image data plays a very important role in the engineering field too. Image mining is a new technology for the analysis and interpretation of the knowledge out of the images. Image mining is a new approach in data mining. Image processing can be broadly defined as the manipulation of signals which are inherently multidimensional.

Nowadays a large portion of information is all in the visual guise. This visual form is very essential and certainly pleasing to the eyes so as to make users search for the images with pleasure. Image mining is applied in many fields and sectors such as medical diagnosis, biology, remote sensing, space research, etc.

1.3 WEB MINING

Web mining as represented in Fig.1.1 is a data mining technique on WWW to extract knowledge from web data that includes web documents, hyperlinks, weblogs, etc.

There is a continuous and a constant growth of the content stored and shared on the web and also on other document repositories. The information retrieval system (IR) goes about collecting the web pages from the document corpus and answers or retrieves aptly for the Input query. This IR acts as the chief component of the search engine. The user goes about searching for the information on the web either by using the search engines or by using browsing directories. The overview of a search engine and web-searching are shown in Fig. 1.2 and Fig. 1.3. There are different IR retrieval models known as boolean model, vector space model, latent semantic indexing and probabilistic model. Information retrieval is the process of representing, storing, organizing, and producing certain access to the information items. The principal functions of the Google search engine are Crawling, Indexing, and Searching. So far as web mining is concerned, the web content mining extracts useful information from the corpus-wise stored contents. The web usage mining discovers interesting patterns from the web access logs on the servers. The data collected are of three kinds:

Web server data

Application server data

Application level data.

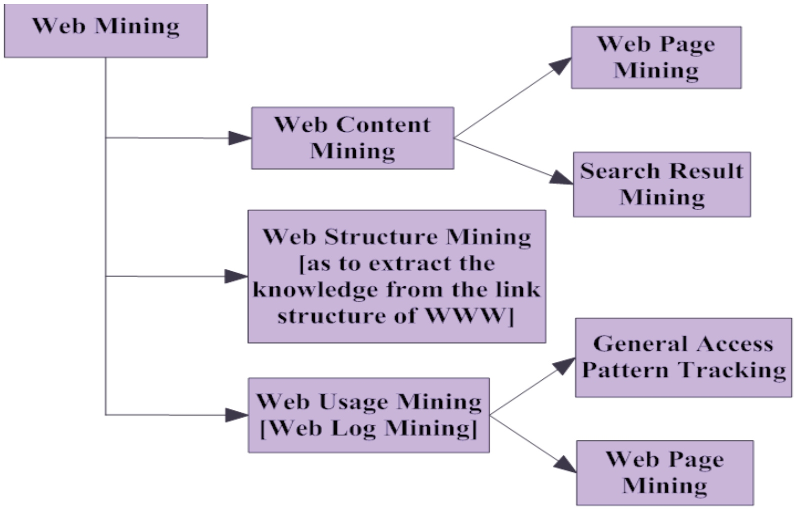

Web Mining
Web Page Mining
Web Content Mining
Search Result Mining
Web Structure Mining [as to extract the knowledge from the link structure of WWW]
General Access Pattern Tracking
Web Usage Mining [Web Log Mining]
Web Page Mining

Fig. 1.1 Web Mining Process

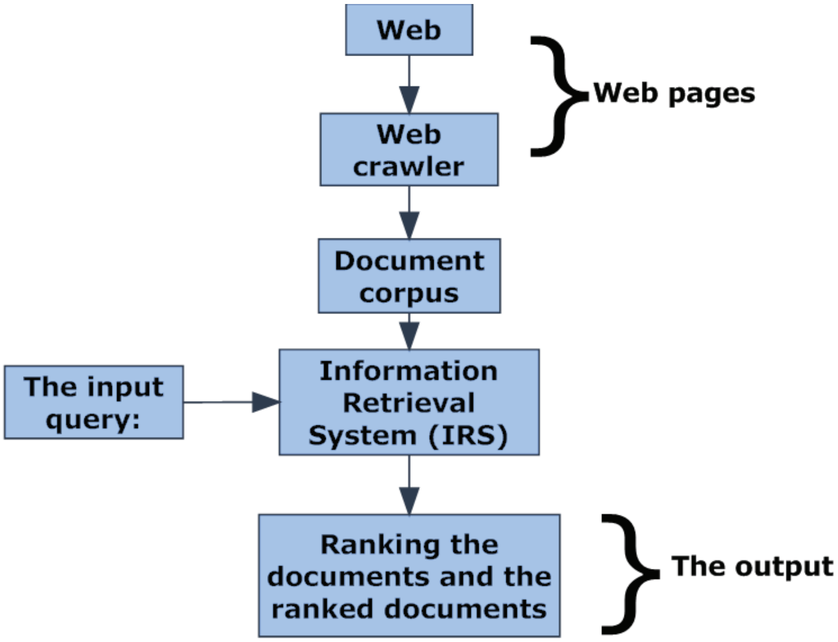

Web
Web crawler
Document corpus
The input query:
Information Retrieval System (IRS)
Ranking the documents and the ranked documents
Web pages
The output

Fig. 1.2 Overview of the Search Engine

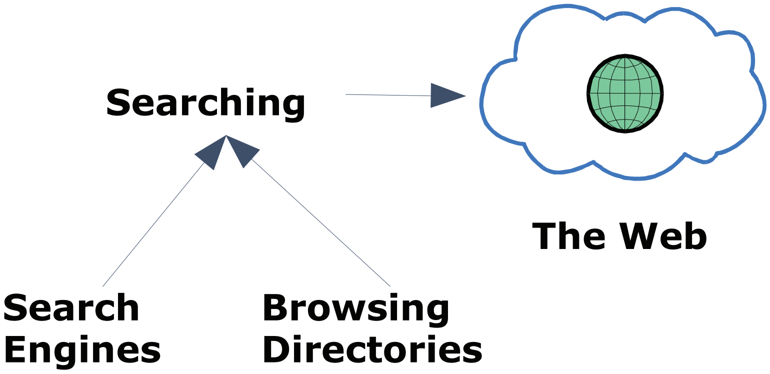


Fig. 1.3 The Web-searching Scheme

1.3.1 Information Retrieval System (IRS)

Preprocessing is a process of an operation on images at the lowest level which not only improves the image but also suppresses the unwilling distortions. It enhances some image features and focuses on image feature processing. The noise is voided or filtered by means of linear and non-linear filtering techniques. Here median filtering is used to reduce the noise. An overview of IRS is given in Fig.1.4.

1.3.2 Entropy classification

Entropy is a statistical randomness measure that can characterize the texture of the input image. Entropy is defined as the histogram counts obtained from the histogram calculation.

In the past, there were problems in target recognition, object recognition, face recognition, and face detection or face verification. Such problems are now solved by means of this image mining process. Image mining is rapidly gaining more attention from the researchers as the field of data mining or information retrieval and multimedia databases.

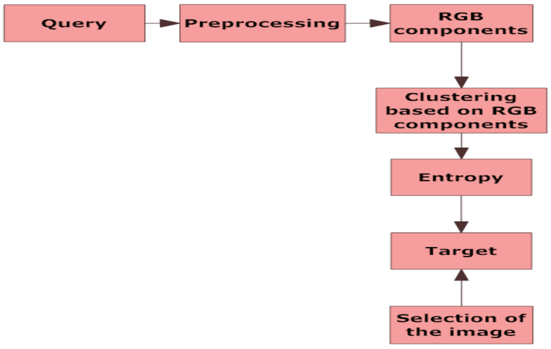


Fig. 1.4 Overview of IRS

1.4 IMAGE TRANSFORMATION PROCESS

Image mining is the process of applying the algorithms in which the data mining algorithms are applied on the images. Hence, we can define image mining as the process of applying the data mining algorithms. The overall flow of image mining is shown in Fig.1.5.

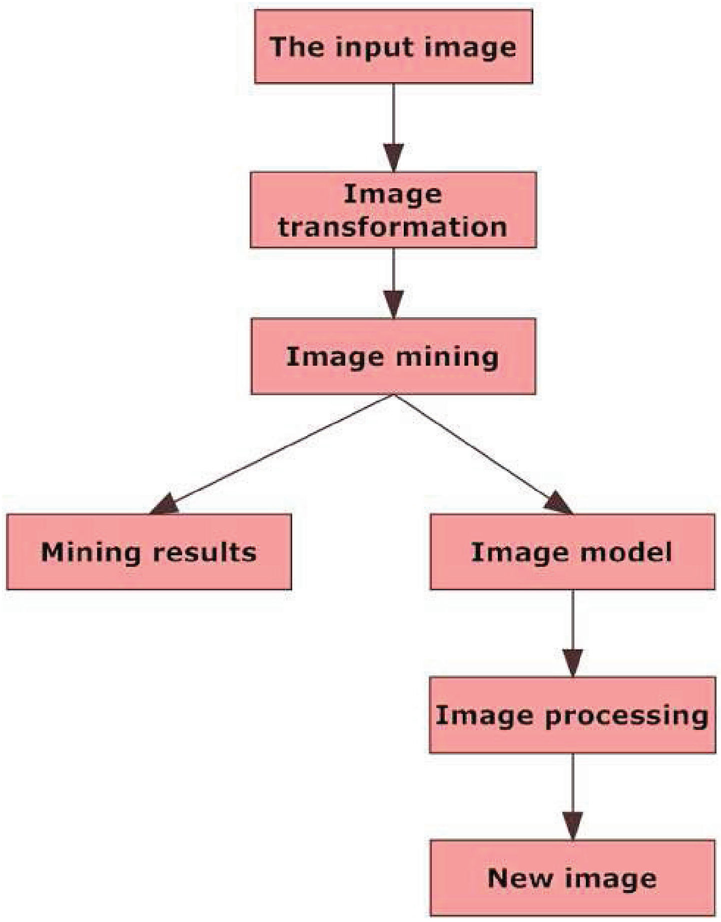


Fig. 1.5 Overview of Image Mining

1.4.1 The existing image mining techniques

Most of the available data mining techniques have been designed for mining categorical or numerical data and are not at all well suited for image mining. On the other hand, there are various image processing techniques such as image segmentation, image enhancement, image restoration, image compression all of which focus on image manipulation rather than on image data analyzing. Of the many image processing techniques already designed and developed, a few can be adopted to mine the image data. In the course of image mining, the input image consists of the raw and the label image pairs. This input undergoes a transformation called image transformation. In the image transformation process, the pixels from a neighboring area will generate similar feature vectors that cause remarkable redundant information in the research table. After having such database like table in accordance with the desired input image dataset, mining algorithms can be used on the resultant table and the decision tree also can be used for mining algorithm.

1.5 LIM BASED IMAGE MATCHING TECHNIQUE:

This image mining process is strictly undertaken solely to determine the exact image while mining an image (multimedia) database. LIM (Lorenz Information Measure) based image matching technique is a novel approach to mine images with neural networks. The resultant performance is deserving, noteworthy and comparable. This LIM technique tries to be an independently functioning process and it is independent enough without a set parameter to create a robust solution.

Image matching is an important requirement in the field of image mining as the application process. A number of the matching techniques have already been developed till today but no optimized matching technique has been developed so far. So, serious research is going on for such an optimized matching technique. For the required content there should be an apt image. Without an apt image it is very hard to proceed and the apt images alone drive the nail home: that is, explain well enough the theme and reach the target almost absolutely. Hence the image matching techniques have to be further and further developed. The image matching model is described in Fig.1.6.

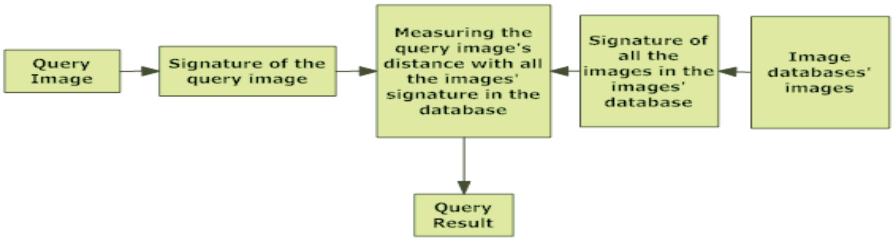

Query Image
Signature of the query image
Measuring the query image's distance with all the images' signature in the database
Signature of all the images in the images' database
Image databases' images
Query Result

Fig. 1.6: Image Matching Model

1.5.1 Kinds of matching techniques

The nearest neighborhood technique is the most commonly used matching technique which is still an important technique utilized in such applications where objects to be matched are represented as n-dimensional vectors. Other matching techniques are least square mode, coefficient of correlation technique, approximate nearest neighbor technique, relational graph isomorphism technique etc.

1.6 CBIR

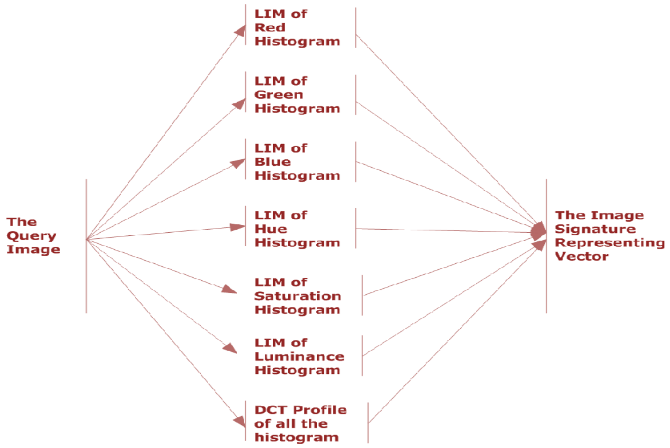
The growth of the internet system not only brings about an explosive volume of digital multimedia data but also provides more ways to retrieve images. Using the manual text annotations, indexing and retrieving image data are done. The annotations can be used to detect the images indirectly. But there are numerous problems with this particular approach. The first problem is that it is very difficult to find out or depict the content of an image or a video scene using only a few key words. The second problem is that the manual annotation process is very slanted, confusing, and deficient. These problems have created a great need for automatic and effective techniques for Content Based Image Retrieval (CBIR). The image's color information is got through the color histograms. Color histogram is a type of bar graph, where each bar represents a particular color of the color space that is being used.

One can extract the content based images and renowned images using the retrieval system is called Content Based Image Retrieval(CBIR) system. Business images do indeed touch every one of the business aspects in the present scenario. Such images play a vital role in the present scenario. CBIR aims at searching image data base for the given query and thus for the specific images that are very similar to the queried image. It also aims at developing new techniques of searching and browsing through the large digital image libraries which are based upon automatically derived imagery features. CBIR focuses only on image features and image data bases to retrieve the queried image. These features can be classified as high level and low level features and users can make their queries based on these features.

1.7 MEDICAL IMAGE SEGMENTATION OR MINING

There are various methodologies for medical image segmentation or mining but there are struggles due to the noise prevailing in the medical images. Hence, to reduce the noise, the medical image is preprocessed and then a multi- level histogram is generated in the foremost phase of this particular process. Image segmentation is prevalently used in many clinical and research applications to the data sets pertaining to the medical image. There are various image segmentation algorithms which have their basics in signal processing theory and modes.

Image segmentation is a vital procedure to collect the useful and necessary information from the images. It is also an easy procedure to recognize the knowledge out of the accrued images. To find out what is what out of the accrued images is easy. Knowledge is easily recognized when it is presented in the form of images in the following cases: geophysical and environmental data from satellite photos, web pages that contain images, and medical images such as Confuted Tomography (CT), Magnetic Resonance Imaging (MRI), and Ultrasound Imaging (UI). These are the most important sources of useful and much needed information in day to day life of the people. The creation of image signature is shown in Fig.1.7.


The Query Image
LIM of Red Histogram
LIM of Green Histogram
LIM of Blue Histogram
LIM of Hue Histogram
LIM of Saturation Histogram
LIM of Luminance Histogram
DCT Profile of all the histogram
The Image Signature Representing Vector

[1]

Fig. 1.7 Creating the Image Signature

Fig. 1.7 Creating the Image Signature

Nowadays image segmentation and other image analysis techniques are developing very rapidly. With further advent of researchers and increase in computing power, the image mining field will move further rapidly. There are too many challenges against medical image segmentation and many opportunities are present in medical image segmentation. There are new techniques to increase the resolution of the medical images and to identify the features and edges of the medical images. The medical image is preprocessed to reduce the noise element in the first phase. In the first phase of the process itself, the histogram is introduced and a multilevel histogram is generated too. In the second phase, the initial segmentation is obtained which is in grey level contours. After the initial segmentation, the histogram is generated and next, the pixel adjacency graph is constructed. Vincent introduced a fast as well as a flexible algorithm for computing the watersheds in the digital grey scale images. His watershed algorithm can segment an image into many homogeneous regions which will have the same grey levels. To obtain a meaningful segmentation or a meaningful segmentation of the image still, the regions of the different grey levels should be merged if those regions are going to be from the same object.

1.7.1 A multi-resolution segmentation

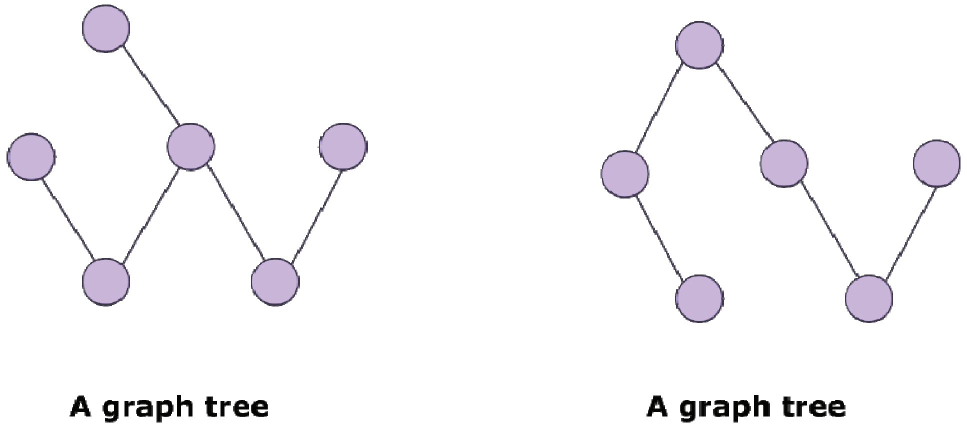
This process of segmentation using histogram techniques reads the input image and obtains very well the grey scale image. The background objects are removed by using the obtained grey scale image. And then the histogram of the image is obtained. The result of the histogram operation will run through the directional filter bank to reduce the noisy signals. Then the segmentation threshold is fixed. The new threshold value is calculated for the segmentation purpose. The following is the result of the medical image segmentation: an image in steps: the first one is a top-left; the 2nd is a top-right; the 3rd one is below- right; and the final image is below-left. The final image is got after the segmentation.

1.8 MINIMUM SPANNING TREE

There are plenty of image segmentation techniques and amidst them is the graph based minimum spanning tree which is very famous for its potentiality in persuasive data representation. The minimum spanning tree is a clustering algorithm which will detect the clusters with irregular boundaries. Clustering is a process of discovering the groups of objects. Objects of the same group are similar and the objects pertaining to different groups are dissimilar. Clustering is an important tool rather to explore the hidden structure of large modern databases. Many techniques have been developed in order to solve the problems arising out of data distribution and these techniques are of four types: hierarchical, partition-based, density-based and model-based. But the clustering algorithm which is based upon the Minimum Spanning Tree solves many problems unlike the above mentioned techniques. Minimum Spanning Tree based clustering approach for image mining can be very well applied to either the grey scale or the color image. There are three phases of the MST partitioning algorithms for image mining. In the first phase optimal number of clusters are created, in the second, discovering of global level outliers with the best number of clusters is performed, and in the third, the segmentation process is held wherein the clusters are segmented accordingly. A number of clustering algorithms are available that can solve the clustering problems but most of these algorithms are very sensitive to the input parameters but the MST based algorithm is indeed very capable of detecting the clusters with irregular boundaries. The MST based clustering algorithm can be used to find out the meta clusters. MST ignores many a possible nexus between the data patterns which cause the cost of clustering to increase. This MST based clustering algorithm is indeed very capable of detecting the clusters of various sizes and shapes. Clustering by Minimum Spanning Tree can be speculated as a hierarchical clustering algorithm which follows a decisive approach. Meta clusters detecting MST based algorithms don't require a predefined cluster number. All the available clustering algorithms require a number of parameters for their inputs and these parameters can significantly affect the quality of the cluster. Divisive Hierarchical Clustering Using Minimum Spanning Tree (DHCMST) based on the hierarchical and divisive approaches can overcome the shortcomings of the classical clustering algorithms.

1.8.1 A Tree

A tree is a connected as well as an undirected graph which will not have simple circuits and it is a simple graph. An undirected graph is a tree if and if ever or if only there is a simple path between any of its vertices. A simple representation of a graph tree is given in Fig. 1.8.



A graph tree

A graph tree

Fig. 1.8 Overview of a graph tree

1.8.2 Spanning Tree

A graph can have many spanning trees. A spanning tree is a sub graph of a graph. It contains all the vertices. A spanning tree is still a tree. An overview of an undirected graph and spanning trees is given in Fig.1.9.

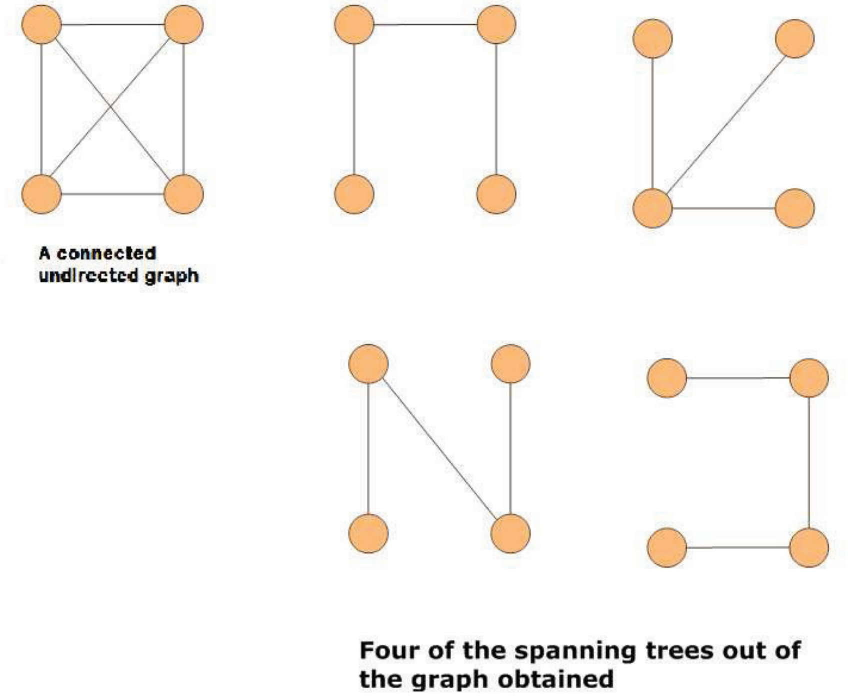

A connected undirected graph
Four of the spanning trees out of the graph obtained

Fig. 1.9 Overview of an Undirected Graph and Spanning Trees

1.8.3 Minimum Spanning Tree

It is a spanning tree of minimum cost for that graph which is called the minimum spanning tree of that given graph. A Minimum Spanning Tree (MST) is visualized in Fig. 1.10.

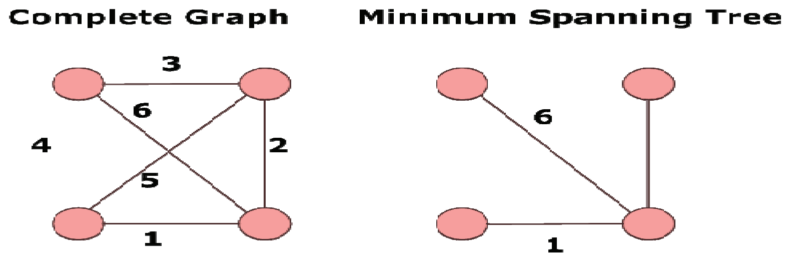


Fig. 1.10 Overview of a Minimum Spanning Tree (MST)

1.8.4 Minimum Weight Spanning Tree

It connects all the vertices through the edges with least weights.

1.8.5 Minimum Spanning Tree Algorithms

Two famous algorithms called Kruskal's and Prim's algorithms are used for constructing MST.

1.8.6 Kruskal's Algorithm

This algorithm works best if the number of edges are kept to a minimum. It only has to check a small fraction of the edges, but in some cases, it would have still to check all the edges. In this algorithm, all the edges are sorted in ascending order by their weights and the MST construction begins with *n* trees. Then for each edge the weight is to be added, and it has to be checked whether the two end points pertain to the same tree and in that case, a cycle will be created. Such cases should be discarded.

1.8.7 Prim's Algorithm

In Prim's Algorithm, the construction of the algorithm begins with some root node *T* and the tree greedily grows outward from the tree *T.* At each step, amidst all the edges between the nodes in the tree *T* and those that are not at all in the tree yet, the smallest weights are added to the nodes and the edge associated with the tree *T*.

1.8.8 Other MST-Based Clustering Algorithms

Other Clustering Algorithms based on minimum and maximum spanning tree were extensively researched. Paivinen designed a Scale Free Minimum Spanning Tree (SFMST) clustering algorithm which constructs scale free networks and output means clusters containing highly connecting vertices. Of late, Wang proposed a new approach hailed as Divide and Conquer mode to facilitate efficient MST-based clustering using the Reverse Delete algorithm. Grygorash proposed two MST-based algorithms known as HEMST and MSDR. The HEMST (Hierarchical Euclidean-distance based MST) requires a certain number of clusters which alone or which number alone is given as an input whereas the MSDR (Maximum Standard Deviation Reduction) doesn't require an input or certain number of clusters. Laszlo and Mukerjee present an MST-based clustering algorithm that was developed for the micro aggregation problem and that puts rather a constraint on the minimum cluster size rather than on the cluster-number. Chowdbury and Murthy designed and proposed an MST-based clustering technique or algorithm that is density-based and that assumes the boundary between any two clusters must pertain to valley regions (regions where the density of the data points is the lowest when compared to the neighboring regions). MST clustering algorithm has been widely used in practical ends. Xu (Ying), Olman and Xu (Dong) use MST as a multidimensional gene expression data. They point out that this MST-based algorithm scarce assumes that the data points are grouped around the center or separated by means of the regular geometric curve and thus, their pointing out denotes that the shape of the cluster boundary has a little impact on the algorithm's performance. A tree is a simple structure to represent any a binary relationship and any connected components of a tree is called the *sub tree* through which the multi-dimensional problem can be converted into a tree partitioning problem. In the tree partitioning problem, we first find the particular tree edges and then cut the relevant tree edges. DHCMST clustering algorithm uses a new cluster validation criterion based on the geometric property of the partitioned regions or the clusters to produce the optimal number of true clusters with a center for each of them. The result shows that DHCMST performs better than K-means algorithm. The result further shows that DHCMST gives the good clusters. The same DHCMST algorithm is programmed in C language. It also generates *dendogram-0* which is used in finding out the inter-cluster relationship between the optimal number clusters. DHCMST algorithm uses both the divisive as well as agglomerative approaches to find out the m*eta clusters.*

1.9 APPLICATION OF MST IN SENSOR NETWORKS

Given are a set of sensor nodes and a data sink in a plane. The transmission power of each sensor node is adjusted in such a manner (using minimum spanning trees in terms of Euclidian distances) that a tree is formed from all the sensor nodes to the sink and the total transmission power of all the sensor nodes is minimum now.

1.10 OUTLIERS

To detect the outliers the Minimum Spanning Tree is used. There is a need to detect the outliers in the database which are detected as unusual objects. The task of detecting the outliers is deemed to be a vital task. The MST based algorithm detects an outlier which is an observation of data that gets diverted or that deviates from the other observations and these outliers are either erroneous or real. Real outliers are those whose real values are very different from those that are observed to be in the rest of the data.

Erroneous outliers are not only the diverted or deviated ones but also the distorted ones. They are distorted owing to the misreporting errors in the data collection. Outliers of both kinds can exert an undue and unwanted influence on the result of the data analysis. Hence both the kinds should be detected absolutely previous to the data analysis. These outliers are individuals or the client-groups and they can behave abnormally. These outliers can be promptly removed or can be viewed separately in regression modeling to improve the accuracy of the data.

1.10.1 The importance of outlier detection

The outliers existing in the data can still translate some intimations into significant intimations in various application domains. And sometimes the outliers existing in the data can translate into critical (of the crisis) intimations. In the field of medicine or public health care, these outlier detection techniques are widely used to detect or recognize the anomalous patterns in the patient's medical records which could be the symptoms of new diseases. In the same way, outliers in the data of the credit cards can even point out to the theft or misuse of the credit cards. These critical translations can even betray an unusual region in the satellite image of the enemy which can indicate the movement of or the encroachment of the enemy troop. Thus in military surveillance it happens to be a phenomenon in national security. Sometimes, the anomalous spacecraft readings can signify the faults in some of the crafts. Thus the detection of outliers has been found to be directly applicable in a large number of domains.

1.10.2 Off from the norm behavior

Many data mining algorithms have the techniques to recognize outliers as the sheer side-product of the clustering algorithms. But the very same techniques define these outliers as the points which do not lie in the clusters. There is a noise feasible in the embedded clusters and the techniques such as these data mining algorithms define the outliers as the noise prevalent in the clusters embedded. This noise only is the background noise. But the techniques other than these data mining techniques define these outliers as neither a part of a cluster nor as the background noise prevalent but as the points which do behave off from the norm.

1.10.3 The principal concern of outlier detection algorithms

The outlier detection algorithms have their chief concern of detecting the outliers. The detected outliers are deemed as the noise that are to be removed in order to make the clustering more reliable. Some noisy points are far off from the data points and some too close to the data points. The far off noisy points would affect the result more significantly, for they are more different from the data points altogether. Hence, it is recommended that the far off noisy points should be recognized and removed from the cluster.

1.10.4 The outlier detection approaches

There is no single universal approach applicable to outlier detection or there is no generic approach to outlier detection. Hence, there are many approaches to outlier detection. These approaches are classified into four major categories. They are distribution-based, distance-based, density-based, and clustering-based categories.

1.10.5 Distribution-based approach

These approaches develop the statistical models from the given data and then a statistical test is applied to determine whether an object belongs to the model or not. Objects that have a low probability depend on or belong to this statistical model and such objects are detected in the outliers. However, this approach is not accessible to the multidimensional data set.

1.10.6 Distance-based approach

In the distance-based approach, the outliers are detected using a given distance measure on feature space. A point *q* in a data set is an outlier with respect to the parameters *M* and *d*. The very same point *q* is an outlier if there are less than *M* points within the distance of *d* from *q* whereas the values of *M* and *d* are determined by the user. This approach is explained in Fig.1.11.

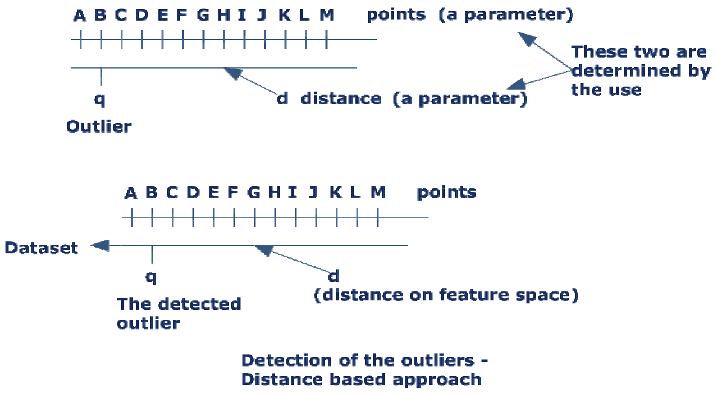

A B C D E F G H I J K L M
points (a parameter)
These two are determined by the use
q
Outlier
d distance (a parameter)
A B C D E F G H I J K L M
points
Dataset
q
The detected outlier
d
(distance on feature space)
Detection of the outliers - Distance based approach

Fig. 1.11 Outline of the Detection of Outliers – Distance-based Approach

1.10.7 Density-based approach

The low density regions in the data are computed and the very same regions will have some object which means the regions will be detected as outliers.

1.10.8 Clustering-based approach

This approach deems the small clusters as outliers. The small clusters significantly contain points less than the other clusters. The advantage of this approach is that it doesn't need to be supervised. Moreover, Custom and Gath present a fuzzy clustering approach. In testing the presence or absence of outliers two hypotheses are used in the fuzzy clustering approach.

1.10.9 Modified k-means algorithm

Jiang proposed a two-phase method for the detection of outliers. In the first phase, using the modified k-means algorithm the clusters are produced and in the second phase, the Outlier Finding Process (OFP) is proposed. The small clusters are selected and deemed as outliers. A small cluster is defined as a cluster with a fewer points than half the average number of points in the k number of clusters. Loureio proposed a method of detecting outliers. In that mode he designed a key idea to use the size of the resulting clusters as indicators of the presence of the outliers. Almedia too used a similar mode to detect outliers. Yoon too designed a mode to detect outliers by using the k-means algorithm. By using Partitioning Around Medoids (PAM), Moh'd Belal Al-Zoubi proposed a clustering-based approach method. The MSTODDN (Minimum Spanning Tree based Outlier Detection using Degree Number) algorithm detects outliers without using any predesigned input parameter. In order to get the desired output, the users don't ever need to select and try any parameter combination. Thus MSTODDN detects outliers from the data set.

1.10.10 Outliers and cluster analysis

Typically outliers are viewed as noise observations which have an extreme influence over cluster analysis. Cluster analysis is otherwise known as the clustering which is the task of grouping a set of objects in such a way that objects in the same group called clusters are more similar certainly to each other than to those in the other clusters. It is the main task in exploratory data mining and it is a common technique in the statistical data mining used in many fields including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

1.10.11 Fuzzy-C means technique

The pre-processed images are classified as low-texture, average-texture, and high-texture detailed images relying on MLE (Maximum Likelihood Examination) or relying on some factors like MLE. These classified images are then subject to the color feature extraction and the resultant retrieval result is pre clustered by Fuzzy-C means technique. This image clustering reduces the search time. The Fuzzy C-means (FCM) is one of the clustering modes which allow one piece of data to pertain to two or more clusters.

1.11 DECISION TREE

Unlike in the conventional decision tree classifiers, in the recent data collection modes, appreciable amount of attributes are uncertain. The conventional decision tree classifiers work with known and apt data values. The uncertainty as regards the recent modes needs to be handled properly and very promptly. Such uncertainty occurs owing to measurement errors, quantization errors, and the data stalemate. A decision tree can always be used to find an answer to a complex problem. The users are allowed to take a problem with multiple possible solutions and to display it in an easy to grasp format which shows the relationship between different events or possible decisions. The possible end results are represented by the furthest branches of the tree. A schematic tree-shaped diagram is used to perform a probability and the diagram is so designed as to determine a course of action. Decision tree learning uses a decision tree as a predictive model which maps observations about an item up to a conclusion about the very same item's target value. Hence, decision trees are powerful in calculation and prediction. These decision trees represent rules which can be understood by humans and as data base these rules are used in the knowledge system. Decision trees help users to obtain both image processing as well as image mining.

1.11.1 Pixel-wise image features

Pixel-wise image features were extracted and transformed into a database-like table which allows a variant data mining algorithm to make explorations on it.

1.11.2 The well-known classification model

The well-known classification model is the decision tree model. Decision trees are simple, real, and fashionable too. From the decision trees, the rules will be mined. Very many algorithms like ID3 and C4.5 have been designed and devised for decision tree creation. These algorithms are widely accepted and applied in image recognition, medical diagnosis, and credit rating of loan applicants, scientific tests, fraud detection, and target marketing. Data certainty takes place due to repeated measurements. Uncertain data management is one of the research themes in recent years. Data uncertainty is generally categorized into existential uncertainty and value uncertainty. Existential uncertainty is an uncertainty about the existence of spatial objects or events. Value uncertainty happens when a data tuple is identified as existing, but its values are not identified exactly. The very renowned k-means clustering algorithm is expanded to the k-means algorithm for clustering the uncertain data.

1.12 VARIOUS ISSUES

Authors have classified the image mining issues into four divisions such as: associations, classifications, sequential patterns, and series patterns. Image mining is more to do than data mining. Wynne Hsu, Mong Lee and Ji Zhang examined the research issues so far as image mining and the development in image mining are concerned. They also designed an information driven framework for image mining and developed four levels of information. They are:

Pixel level

Object level

Semantic concept level

Passion and knowledge level.

1.13 PROBLEM DEFINITION

The parameters can affect the cluster quality but still all the existing clustering algorithms require a number of parameters as inputs. The classical clustering algorithms have their own limitations. But DHCMST can still overcome the limitations of such classical clustering algorithms. The final clustering result was not that apt before, but now by means of the DHCMST the accuracy of the final clustering result can be improved. The DHCMST algorithm optimizes the number of clusters at each hierarchical level with the cluster validation criteria during the minimum spanning tree construction.

There were two kinds of clustering problems: one was that the maximum inter-cluster distance was minimized and the other was that the minimum inter-cluster distance was maximized. Although there are many methodologies available for medical segmentation, the struggles are still there due to noise representation in medical images. Due to the noise in medical images, the features are missing and hence, the struggles persist. A combination of graph cut technique is efficiently used to solve a wide variety of low-level computer vision problem, which is image smoothening as well as k-means clustering with the multi threshold. Image segmentation has been recognized as the chief problem of medical image analysis. A range of computational vision problems can use the principle of segmented images, and if these segmentations are very reliable and efficient in computation, then that range of computational vision problems or the wide range of the same can indeed be solved. For instance, intermediate level vision problems such as stereo and motion estimation do require an appropriate region of support for correspondent operations. Higher level problems such as recognition and image indexing can also make use of segmentation results. Computational approaches should be developed to solve the crisis in image segmentation. There is a problem with the distance based approach too which is high and harsh computational complexity. It is also difficult to determine the *M* and *d* values. And as far as the outliers are concerned, k-means algorithm is sensitive to them and the very sensitive hue affects the effective and apt result. Cluster with irregular boundaries cannot be detected using both the k-means and the PAM algorithms. Further EMST-based methods are there for solving the different problems in detecting the outliers in the dynamic dataset. Feature selection is an important crisis in the object detection and Zsun demonstrates genetic algorithms which provide a sample, general and powerful framework choosing good subsets of the features leading to improved detection rates. The fundamental challenge or problem or crisis in image mining is to reveal how low-level pixel representation enclosed in a raw image or image sequence can be processed to high-level image objects along with the relationships. The problem so to say consists of the content-based image retrieval problem, the image understanding crisis, the data mining problem and the data bases crisis.

1.14 OBJECTIVE OF THE THESIS

The objective of the thesis it is to analyze image mining deeply and to detect the problems arising in image mining and then the sequent salvations by means of such and such algorithms. The objective is to be discussed along with medical imaging, medical imaging process, medical image segmentation, and also the minimum spanning tree and the maximum spanning tree, and further clustering techniques, and finally the decision tree classifiers too.

1.15 ORGANIZATION OF THESIS

The thesis is organized as follows

Chapter 2 provides a deliberation on the related works. Chapter 3 presents a description of proposed work in two phases. The first phase comprises the construction of Meta clusters through MST based clustering. The second phase proposes a data cleaning technique through MST. Chapter 4 discusses the third phase involves the formation of a cluster interfaced objective function for decision tree classification. The fourth phase describes an integrative approach for a graph and user-interface based medical image segmentation and multi-level thresholding using k-means clustering. Chapter 5 discusses the application part of Divisive Hierarchical Clustering using Minimum Spanning Tree (DHCMST) which has been chosen as part of this research work. Eventually experimental results are shown using this algorithm. Chapter 6 discusses the conclusion of this research as well as possible future enhancement work related to the improvement of the proposed algorithm.

CHAPTER II

LITRATURE REVIEW

2.1 INTRODUCTION

There are many methodologies to approach image mining by segmenting the image. The problem is traditionally organized into two main categories as region based and boundary based. Each of the approaches presents its own advantages and drawbacks. They can be used isolated or combined in any convenient manner to explore the complementary properties of each method. Otherwise, they can either be unsupervised without any user intervention or interactive as often required by medical imaging applications [54].

Many issues still remain open in image mining, as there are many different approaches with many different application areas where image segmentation is mandatory along with evaluation of the performance on mining algorithm. The present research will also look at those problems from a different level, trying to identify those contributions where integration, fusion, combination, co-operation or interaction are the major keywords for approaching the segmentation and classification issues. By reviewing the various methods based on the use of different and complementary methodologies, the study explores the advantages and disadvantages of a particular method in order to improve the performance. The specific aim of the thesis is towards image mining and it is achieved by clustering techniques and classification methods and the better solution is obtained by a better decision-tree approach. The literature review provides some methods and also a quantitative evaluation of the problem domain by various authors.

2.1.1 Image Segmentation

A brief overview of the two earlier surveys [30] and [55] is presented here. In [30] the author describes the main ideas of image segmentation methods which are grouped into five major classes as 1) Measurement space guided spatial clustering [further divided into thresholding and measurement space clustering] 2) Region growing [further divided into single linkage, hybrid linkage, centroid linkage schemes] 3) Hybrid linkage combination techniques 4) Spatial clustering, and 5) Split and merge. In Pal and Pal [55], the author reviews some image segmentation methods (distributed by 178 papers) by express fuzzy and non-fuzzy techniques including color image segmentation and neural network based approaches. The author compares some of the methods and also provides some comments on quantitative evaluation of the segmentation results. Specialized surveys in a specific image segmentation topic can be found in [19] for edge detection, [83] for region–based segmentation methods, [65] for Thresholding techniques, [62] for feature and feature extraction methods, [13], [46] for color images, and [3] reviews the use of neural networks for image processing in general and image segmentation in particular.

Active contours constitute a general technique of matching a deformable model onto an image by means of energy minimization. The introduction by kass et al in [42] of deformable models has been used in the application of image segmentation.

2.2 WATERSHED TRANSFORM

Water shed transform is an important paradigm for image segmentation and it is a main step in several hybrid image segmentation frame works. Although watershed is usually considered as a region based approach, De smet et al [20] pointed out that the watershed transform has proven to be a powerful segmentation tool that can hold the attributed properties of both edge detection and region growing techniques which makes it a co-operative approach. Other authors, as Haris et al [32] and Adiga et al [1] used both pre-processing and post-processing steps. Nevertheless the performance of watershed-based image segmentation methods depends largely on the algorithm used to compute the gradient. The advantage is that produces coherent regions where boundaries are always guaranteed to be connected and closed. Since the early 1990's there has been a considerable amount of scientific work on watershed transform that was originally proposed by Beucher and Lantuejoul [7] as an image processing tool.

2.3 FEATURE DOMAIN AND CLUSTERING

A number of approaches to segmentation are based on finding compact clusters in some feature space [17]. In this technique, a vector of local properties (features) is computed at each pixel and then mapped into the feature space. Features such as intensity, texture or motion are the commonly studied parameters. Significant features will be shared by numerous pixels, and thus will form a dense region in feature space. The feature space is then clustered, and each pixel is labeled with the cluster that contains its feature vector. Clusters in feature space can then be used for image segmentation, typically by fitting a parametric model to each cluster and then labeling the pixels whose vectors lie in the cluster with the parameter.

2.4 THRESHOLDING METHODS

Thresholding techniques are based on the assumption that cannot always be modeled as mixture of Gaussian, for example luminance histograms of natural images. An early review of thresholding methods was reported in the highly cited paper of [65] Sahoo.et al surveyed segmentation algorithms based on thresholding and attempted to evaluate the performance of some thresholding techniques using uniformity and shape measure. Cheriet et al [14] presented a general recursive approach for image segmentation by extending Otsu's method. This approach has been implemented in the area of document images. It segments the brightest homogenous object from the given image at each recursion, leaving the darkest homogeneous object.

2.5 CLUSTERING METHODS

Clustering techniques appeared earlier in literature and were used in numerous applications [35]. Adjacent pixels whose value (grey level, color value, texture) lies within a certain range belong to the same class [68]. Those methods achieved reasonable performance when the input was characterized without noise and with a small number of regions. This explains why these methods are mainly used in text segmentation from an image [68]. For review of thresholding techniques, readers are referred to the survey papers [64] and [54]. Among the algorithms proposed for histogram segmentation parametric and non-parametric approaches are treated distinctly. In the first ones, a histogram is considered to be a probability density function of Gaussian [74], [57] and the problem is reformulated as parameter estimation followed by pixel classification. The main drawback of these approaches is that the histogram is obtained from real images following the selection of image features usually based on intensity of color, texture and clustering operator on the feature space in order to capture the global characteristics of the image. Ignoring spatial information and using a specific distance measure, the feature samples are handled as vectors and the objective is to group them into compact but well-separated clusters. After the clustering process is completed, the sample data are mapped onto the image plane typically by fitting a parametric model to each cluster and then labeling the pixel according to each parametric model to produce the final region [58]. If the number of objects is known, optimization algorithm can estimate efficiently the parameters of these distribution. The main drawback of these approaches is that the histogram can be obtained from real images only.

2.5.1 Hard clustering

Currently K-means is among the most popular clustering algorithm due to its simplicity and efficiency in unsupervised classification. It starts with a random initial partition and keeps reassigning the features to clusters based on the similarity between the feature and the cluster centers until a convergence criterion is met. A major problem with this algorithm is that it is sensitive to the selection of the initial partition and may cover only a minimum of the criterion function value if the partition is not properly chosen.

In [58] Pappas indicated two problems with K-means algorithm which are: it uses no spatial constraints and it assumes that each cluster is characterized by a constant intensity. In order to overcome these problems, Pappas introduced a generalization of the K-means clustering algorithm and applied this procedure on grey level images. This approach aims at separating the pixels in the image into clusters based not only on the intensity but also on their relative spatial location. This algorithm considers the segmentation of grey level images as a maximum a posteriori probability (MAP) estimation problem.

Work by Turi [71] described a method of automatic determination of the optimal number of clusters in K-means clustering. It proposes a validity measure using the ratio of intra-cluster and inter-cluster measures incorporated with a Gaussian multiplier. The optimal number of clusters is found by minimizing the validity measure.

The mean-shift algorithm is a non-parametric statistical method that finds the peaks (local maximal) of the histogram without estimating the underlying density function. It has been used for the first time by Fukunaga and Hostetler in [26] with the goal of proposing an intuitive estimation of the gradient probability density of a set of points. Later it has been used extensively for detecting the unseen object in the image [17].

2.5.2 Fuzzy Clustering

In the recent years there has been considerable interest in the use of fuzzy segmentation methods which are able to retain more information from the original image than hard segmentation methods. Fuzzy clustering theory was first introduced by Zadeh [80] to generalize the conventional cluster theory. Based on the definition of a fuzzy event [80] the grey level image can be seen as a fuzzy event modeled by a probability space.

Fuzzy c-mean (FCM) is one of the most well-known methodologies in clustering analysis [08, 72]. The reason for its success is due to the introduction of fuzziness for the belongingness of each image pixel. The FCM algorithm classifies the image by grouping similar data points in the feature space into clusters. This clustering is achieved by iteratively minimizing a cost function that is dependent on the distance of the pixels to the cluster centers in the feature domain. In most situations FCM uses the common Euclidean distance which supposes that each feature has equal importance in FCM. This assumption seriously affects the performance of FCM since in most real world problems, features are not considered to be equally important.

In [74] Wang et al proposed a new robust metric, which is distinguished from Euclidean distance for it improves the robustness of FCM. The feature-weight learning FCM technique [10] assigns various weights to different features to improve the performance of clustering. The spatial function can be estimated at each interaction and incorporated into the membership function which makes the new FCM technique less sensitive to noise.

2.6 IMAGE RETRIEVAL SYSTEMS

A large amount of unstructured image data that resides in distributed multimedia repositories has lured researchers to come up with efficient image querying and retrieval or miming mechanisms.

The early 1980s were motivated with text-based tagging and retrieval methods. However, the approach fails where the user wishes to find a pattern in the image. Also, the user has to essentially rely on the credibility of the tagging process. [10] and [11] are excellent examples of the research motivated by text based retrieval. [40] is a survey that highlights the gap between the user's semantic queries and the shortcomings of text based retrieval systems in meeting these requirements.

The pioneering work in the field of content based image retrieval was done in the early nineties by Kato [43] who developed an automated system for image retrieval using color and shape features. Query By Image Content QBIC[53] was the first commercial CBIR system. It allowed the user to upload an image query, sketches and drawings, and some color and texture patterns etc.

Virage [05] is a content-based image search engine developed at virage Inc. virage goes one step further than QBIC and supports combinations of visual queries based on color, composition, texture and structure.

MIT Media labs photo book [60] presents a set of interactive tools for browsing and searching images. It consists of sub-books from which shape, texture and face features are extracted.

Retrieval ware is a content-based image retrieval engine developed by Excalibur Technologies Corp [63,22]. Visual search [67] is a n internet based text/ image search engine that explores the spatial relationship query of image regions.

Cheriet [14] uses color, texture, shape and spatial location information in the segmented image regions to search and retrieve similar regions from the database.

Though a number of easy systems exist, image mining has become a challenging issue of this modern era. A high volume of research has been done in the area of image mining [11]. In [48], Ma and Manjunath evaluated the texture image annotation by various wavelet transform representations. Finally for shape representation, the available choices were Rui et al's [64] modified Fourier descriptor and Gross and Lateeki's [29] approach to preserve qualitative differential geometry of the object boundary.

However, research methodology is not to define an image by content description format [open source java] for image feature extraction[2]. With rapidly increasing data repositories useful information is retrieved by implementing mean clustering which includes the dynamic features.

2.7 IMAGE MINING

In the context of image mining, the image documents including directly using LDA [24] and Spatial Latent Direichlet Allocation (SLDA) [76] models for the salient points of the image using Lowe's difference-of-Gaussian (DOG) detector [18], is the image patches which includes the 128- dimension SIFT descriptor, by applying K-mean clustering on a large collection of patches from different categories of the image. The categorization of image is essential to perform image mining in an efficient way. A study is made with reference to the paper with emphasis on using saliency models as an intermediate step to interpret the semantic meaning of images. Affine invariant saliency model such as the SIFT descriptor [38], [40], [18] have exhibited very good performance in image retrieval across several well-known databases. More specifically, Belongie and Carson introduce the blob world approach [9], which bring global image features (color, texture) together and represent their spatial distributions as a mixture of Gaussians.

Image annotation was conventionally solved as nearest-neighbor problem [16] [77]. Similar approaches range from studying the relevance between visual similarity and semantic similarity [68] using language entities to construct visual ontologies [34] or jointly modeling images and tags [45]. Most recently, [27] proposed to use Conditional Random Field (CRF) to predict the image-based potential of the object categories, visual attributes and preposition relationship present in the image. Recently, a new imaging technology has been introduced, the Multi-epitode Ligand Cartography MELC, also referred to as Toponome Imaging System (TIS) [33]. This technique allows the spatial location of at least a hundred proteins on the same tissue or cell sample to be imaged in situ, overcoming the spectral limitations of florescence microscopy. In such cases the technique of mining improves the Toponome Imaging System (TIS), to sub-divide the alpha-beta cells, by means of supervised Learning-based Tissue Detection.

2.8 IMAGE REGISTRATION

The basic objective of image registration can be roughly described as to find a suitable transformation, applying one of the images, to make it similar to the other. To measure the similarity between two images, some researchers choose to detect the features in the image first, and then measure the similarity of the two images by the alignment of the two sets of the detected features [Flusser and Sul 1994] [25]. Since the features capture the important information in their image, the alignment is a sensible answer to the alignment to the two images. Besides, these methods have an advantage since the features are sparse representations of the image and thus the computational complexity is not heavy.

However image segmentation relies on the detection of features manually. Instead of explicitly extracting features before registration, some researchers directly measure the difference between the two images. A straight-forward way is to use the LP norm, or cross-correlation of the two images [44].

Such a model has the limitation that it assumes the two images to be of the same modality. To choose this assumption, authors in [73] [75] [49] used mutual information and authors in [30] used the gradient of the intensity. The minimum of the similarity functional does not reflect the visually suitable alignment. Rigid transformation, though simple, is a suitable choice for intra-subject registration. [49]

Similarity [36] and affine transformation are usually considered to be on the “ boundary” of human perception since two objects differing by a similarity or an affine transform may be thought of as having the same shape, while a transform with more degree of freedom (DOF) is considered to be of another shape [56]. All the above schemes only handle the local registration task.

2.9 MARKOV RANDOM FIELD MODEL

Jun Xu, James P. Monaco, and Anat Mada Bhushi[78] present a model named Markov random field driven region-based active contour model (MaRACel) for medical image segmentation. State-of-the-art Region-based Active Contour (RAC) models assume that every spatial location in the image is statically independent of the others, thereby ignoring valuable contextual information related as to medical image segmentation[78]. In probability theory and statistics, the term Marko property refers to the memory-less property of a stochastic process. If a process is going to have this property, then it is called the Markov process. In the domain of physics and probability, a Markov random field (often abbreviated as MRF), Markov network or undirected graphical model is a set of random variables having a Markov property described by an undirected graph. By means of this Markov random field model which is based upon the region based contour model, the field of medicine takes to a random measure of medical images which aid the medical therapy voiding thus the fatal future of the folks.

2.10 MINIMUM SPANNING TREE

Minimum Spanning Tree is a graph-cut technique. If a connected, undirected graph is assumed, the sub-graph of that graph is the spanning tree. Sub-graph is only a tree and it should connect all the vertices together so as to become a spanning tree. A spanning tree is a tree or a sub-graph in which all the vertices are connected together. Thus, a single graph can have many different spanning trees. Hence, a sub-graph is a tree. A sub-graph that connects all the vertices is a spanning tree. An MST is a spanning tree with minimum weight or with weight less than or equal to the weight of every other spanning tree. The vertex is a curve. The vertex of an angle is the endpoint where two line segments or lines come together. A vertex is the highest point, the top or apex, or the crown of the head. Connected graphs without a circuit are called trees and several trees make a forest (Hochstattler and Schliep, 2010). Every connected graph has a spanning tree. A weighted graph is a graph in which each edge has a weight. The weight of a graph is the sum of the weights of all the edges. Adding an edge that connects two vertices in a tree creates a unique cycle. One can make a tree break into partition of its vertices into two disjoint sets. A crossing edge is an edge that connects a vertex in one set with a vertex in another. The weight of a spanning tree is the sum of the weights given to each edge of the spanning tree. A minimum spanning tree has (V-1) edges where V is the number of vertices in the given graph. MST is applied to a problem like the phone network design. If one has many offices, and if that one aims to connect all his offices with a minimum total cost, it should be a spanning tree, and one can remove some edges as to save money to make the cost minimal. Thus in telephone connections, electrical setups, hydraulic bases, TV cable connections, computer links, and road connections MST can be applied.

Zhang Jian et al[82] propose a paper with such an approach for automated segmentation and detection of dendrites and spines from a fluorescence confocal image (Zhang Jian-Mingal et al, 2012). The proposed approach can effectively extract the skeleton of dendrite and achieve a better result on spines detection. The detection of the spines will cause much cure on the back. Otherwise there may never be a cure for the back pain pertaining to the spine and the backbone.

2.11 DECISION TREES

A decision tree is a decision support tool which uses a tree like graph or model of decisions. These decision trees are commonly used in operation research, especially and specifically in decision analysis to analyze a strategy to reach the goal.

Semantic texts on forests(STF) are a form of random decision forest that can be employed to produce powerful but low-level code words for computer vision. Each decision acts directly on image pixels, resulting in a codebook that bypasses the expensive computation of filter-bank responses or load descriptors. Furthermore, the STFs are extremely fast when compared with K-means clustering and nearest neighbor assignment of feature descriptors [66]. Hence, there are decision forests too. Another use of decision trees is to calculate the conditional probabilities.

2.12 RANDOM FORESTS

Random Forests (RF) are an ensemble of multiple decision trees that are trained in a random manner [52]. These forests are each a multiple lot of trees capable of taking multiple decisions. And these decision forests are nothing but the multiple graphs put together to take multiple decisions.

Enhancement of the random forests to segment the 3D objects into different 3D medical imaging modalities. More accurate voxel classification is achieved by intelligently selecting "good" features and neglecting irrelevant ones; this also leads to faster training [52]. Voxel classification is made for the airway tract medical image segmentation. The entangled decision forest is a new discriminative classifier which augments the state-of-the-art of the decision forest, resulting in higher prediction accuracy .

2.13 SUMMARY

In this chapter many techniques related to the proposed mining method have been reviewed. Special emphasis has been placed on the strategy used to carry out image segmentation using clustering and various other strategies and methods used to segment the image for easy mining process.

Based on all the techniques discussed in this chapter, it is clear that image segmentation and mining procedure are complex issues. Another conclusion is that the methods are application dependent and some parameters have to be refined accordingly to suit the type of image. The large number of available methods is an indication that the final solution is still far to come.

Actually it is not feasible to determine the best approach. Need for image mining can however be emphasized so that researchers can quickly and effectively compare their algorithms with well established methods. Evaluation issues and methods are addressed in the coming chapters.

CHAPTER III

DETECTING THE OUTLIERS AND META CLUSTERS IN MINIMUM SPANNING TREE

3.1 INTRODUCTION

The researcher intends to detect the outliers and to find the meta-clusters. Hence, in this thesis, an algorithm is proposed which is a minimum spanning tree based on an algorithm for detecting the outliers as well as for finding the meta-clusters. The outliers are detected by means of a distance between the objects or the points in the data set. To find out the meta clusters, the algorithm finds first the proper number of clusters available at each level by means of the data partition of the data set: the algorithm acts twice in two phases. The first phase of the algorithm creates clusters with guaranteed intra-cluster similarity, and the second phase creates a dendrogram. Outliers still can influence the data analysis unduly and hence they have to be detected prior to the performance of data analysis. The outliers can deviate or divert or even attack the prompt or proper performance of the data analysis. Hence, these outliers will have to be detected and removed. But sometimes they are very beneficial too. There are a few data cleaning algorithms proposed in this thesis. They are: Divisive Hierarchical clustering using the Minimum Spanning Tree (DHCMST), the Minimum Spanning Tree based Outlier Detection using Degree Number (MSTODDN), and the Euclidean Minimum Spanning Tree algorithms. Thus such data cleaning culminates in the image mining process and the data is cleaned or gleaned for the sake of the image segmentation and the means of this image segmentation in turn is used to culminate the segmented image into the medical image segmentation.

3.1.1 Meta clustering

Meta clustering is a new approach to the problem of clustering. It aims at creating a new mode. In this mode, there will be an interaction among the users, the clustering system, and the data. Meta clustering finds many alternative good clustering of the data and allows the user to choose the most useful clusters from these alternative good clusterings. In the meta clustering process, in order to help the user avoid having to evaluate too many clusterings, the base-level clusterings are organized into a meta clustering, and that clustering group the similar base-level clusterings together.

3.1.2 The definition of outliers

An outliers in an observation of data that deviates from the other way round observations so that each does arouse a suspicion that each was indeed generated by means of a different mechanism.

3.2 CLUSTERING IN IMAGE SEGMENTATION

Image segmentation is the decomposition of the gray level image or the color image into homogeneous tiles. Before this process in image segmentation, cluster analysis help us to detect the borders of the objects in an image.

3.2.1 Outliers Detection

Some observations are different. Such observations are defined as outliers. Outlying observations may be the errors, or they could have been recorded under exceptional circumstances, or belong to another population or mass data or massive data. Consequently, they do not fit into the model obtained. It is very important to keep the users able to detect these outliers. In practice, one often tries to detect outliers using diagnostics starting from a classical fitting method. However, classical methods can be affected by outliers so strongly that the resulting fitted model does not allow to defect the deviating observations. The outliers themselves don't let the user detect them as per their suggestion. This effect is called the masking effect. In addition, some good data points might even appear to be outliers, which is known as swamping. To avoid these effects, the goal of robust statistics is to find a fit that is close to the fit the user would have found without the outliers. The user can identify the outliers by their large deviation from the robust fit. When analyzing data, outlying observations cause problems because they may strongly influence the result. Robust statistics aims at detecting the outliers by searching for the model fitted by the majority of the data.

3.3 MINIMUM SPANNING TREE BASED OUTLIER DETECTION USING DEGREE NUMBER (MSTODDN) ALGORITHM

As in metric space such as E^n, a point set S has been given. The points in S are selected and the hierarchical mode starts by constructing a Minimum Spanning

Tree (MST) from those very same points in S. The Euclidean distance between the two end points of the tree is the weight of the edge of the tree. The Euclidean distance between a pair of objects can be surely represented by a corresponding weighted edge. The algorithm is based upon the Minimum Spanning Tree but not limited to the two dimensional points. The data set is represented as MST and to detect the outliers from that dataset represented by MST, a new approach based on the Minimum Spanning Tree is proposed. For any undirected graph *G* with the *degree* of a vertex *v,* which is written as *deg (V),* is equal to the number of the edges in G which consists of *v*, that is, the number of edges which are incidental on *v*.

Definition 1

When MST has been given for a data set S, outlier is a vertex *v*, whose degree is less than THR, with dist (*v, Nearest-Neighbor (V)) is greater than* THR where THR is a threshold value used as a control parameter. The EMST is constructed using the given data set. Then vertices v,which have the degree less than THR, are identified. Then Nearest-Neighbors for the above vertices v are found. The distance between the vertices v and the nearest neighbor vertex is computed. If the computed distance exceeds the threshold value THR, then corresponding vertices are identified as an outlier. When scanning the MST, the edges are ordered from the smaller to the larger length. Then we define the threshold as:

$THR=\max(L_i-L_{i-1})*t$

Where L_i is largest in the order and $t \in [0,1]$ is a user defined parameter.

Algorithm: MSTODDN

Input: S is the point set

Output: O is the set of outliers

```
1. $O = \varphi$
2. Construct an EMST T from S
3. For i=1 to |v| do
4. if deg ($v_i$)< THR and
   dist($v_i$, Nearest-Neighbors($v_i$))> THR.,
    then remove $v_i$; $O=O\cup\{v_i\}$
5. Return outliers O.
```

Fig.1 shows EMST, is constructed from the point set S. The algorithm finds the outlier based on the degree number of the objects or points. Fig.3.1.shows the possible distribution of the points in the MST with their outlier objects. We use the graph of Fig.3.1 as an example to illustrate the algorithm which finds the outliers in the dataset represented by MST. If the outliers are present in the dataset (which is the MST itself), then such are identified based on the degree number of points as in the MST (step 4). Fig.3.1 shows the MST constructed from the point set S.

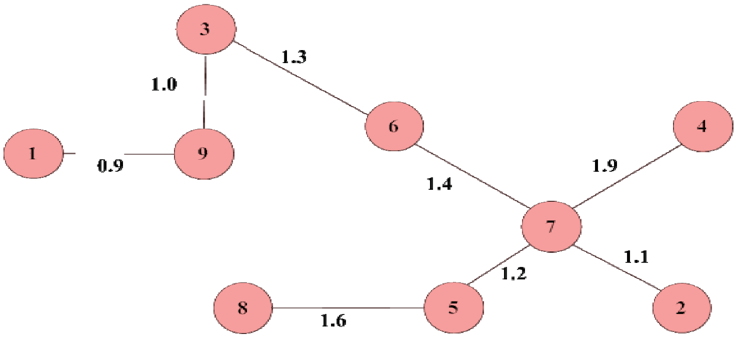


Fig.3.1 Overview of the MST constructed from the point set S

3.4 OUTLIERS AND THE MINIMUM SPANNING TREE

An outlier is certainly an observation of the data that deviates from other observations so that it definitely arouses suspicion that it was generated by a different mechanism from the greater part of data. On the other hand, inliers are defined as an observation that is explained by the underlying probability density function. In clustering, outliers are deemed to be the noise observations that should be removed in order to make more reasonable clustering. Outliers may be erroneous or real as in the following sense: real outliers are observations whose actual values are very different from those of the rest of the data and which violate the tangible relationship among the variables. Outliers can often be individuals or groups of clients exhibiting behavior outside the range of what is deemed normal. Outliers can be removed or considered separately in *regression modeling* to improve the accuracy which can be considered the benefit of the outliers. Identifying them prior to the modeling and analysis is important. In clustering-based methods, an outlier is defined as observation that does not fit to the overall clustering pattern.

Many data-mining algorithms find out outliers not as the main product but as a side-product of clustering algorithms. However these techniques define the outliers as points, which do not lie in the clusters. Thus the techniques deem and define outliers as the background noise in which the clusters are embedded. Another class of techniques define outliers as points, which are neither a part of a cluster nor part of the background noise; they specifically or very significantly the points which behave very differently from the norm or the usual trend. But these outliers are the background noise which distorts the images in the image segmentation or the image mining process. And these noises are of three different kinds and they are as follows:

1. Transmission noise;
2. Gaussian additive noise;
3. Noise due to the multiple sources of the lightings.

These noises do indeed affect the algorithm's performance. The MST algorithm detects the clusters and the resultant image mining or the image segmentation will be distorted due to the effect of the noise.

3.4.1 The outlier detection problem

The outlier detection problem in some cases is the same as the classification problem. Clustering is a popular technique used to group similar data points or objects into groups or clusters. Clustering is an important tool for outlier analysis. When one wants to analyze the presence of the clusters, one should first go for the clustering process. Several clustering-based outlier detection techniques have of late been developed. Most of these techniques rely on the key assumption that normal objects belong to large and dense clusters, while the outliers form very small clusters. These very small clusters are nothing but the metaclusters and hence by using the MST algorithm, the metaclusters are formed for the concrete realization of the outliers within them.

The main concern of *clustering based* outlier detection algorithms is to detect the clusters as well as the outliers, which are often regarded as noise that should be removed in order to make more reliable clustering. Some noisy points may be far away from the data points, whereas the others may be close. The far away noisy points would affect the result more significantly because they are more different from the data points. It is desirable to identify and remove the outliers, which are far away from all the other points in the cluster. The outliers are always found at a far off distance almost off from the cluster. The close noisy points can still be beneficial. Hence, to improve the clustering, such algorithm use the same process and functionality to solve both clustering and outlier discovery.

The tree-partitioning algorithm partitions a MST into sub trees which represent different homogeneous regions and minimize the sum of the gray level variations over all the sub trees under the constraints that each sub tree should at least have a specified number of nodes and two adjacent sub trees should have significantly different average gray levels. MST algorithm can detect the clusters with the irregular boundaries and also detect the highly connected sub trees. These highly connected sub trees are called the clusters.

3.4.2 Novel data cleaning through Minimum Spanning Tree for data mining

Each connected component is considered a cluster, and a connected component with just one vertex is defined as an outlier. *Density-based* approaches compute the density of the regions in the data and declare the objects in low dense regions as outliers. *Clustering based* approaches consider clusters of small sizes as outliers. In these clustering-based approaches, small clusters (clusters containing significantly less points than other clusters) are considered the outliers. The advantage of *clustering-based* approaches is that they do not have to be supervised. Some authors have proposed a two-phase method to detect the outliers. In the first phase, the clusters are a modified *K*-means algorithm. In the second phase, an Outlier-Finding Process (OFP) is proposed. The *small clusters* are selected and regarded as outliers. A s*mall cluster* is defined as a cluster with fewer points than half the average number of points in *k* number of clusters. A method for detecting the outlier known as hierarchical clustering technique. The key idea in the technique is to use the size of the resulting clusters as indicators of the presence of outliers. Yet another author proposes to construct MST of point set and delete the inconsistent edges compulsorily – the edges, whose weights are significantly larger than the average weight of the nearby edges in the tree. The MST clustering algorithm has been very prevalently used in the practice. This thesis however proposes the minimum spanning Tree based algorithm to detect the outliers. The outliers are detected from the data set based on the degree of an object or the point and the distance between the objects in the data set (MST). The outliers can be deemed severally or even can be removed as in *regression modeling* to improve the accuracy of the outliers. The regression modeling consists in finding a dependence of one random variable or a group of variables. The Minimum Spanning Tree based outlier detection using Degree Number (MSTODDN) algorithm is based on Minimum Spanning Tree and does not require a predefined input number of the parameters and the algorithm constructs an EMST of a point set.

3.5 DHCMST

The DHCMST (the Divisive Hierarchical Clustering Using Minimum Spanning Tree) can overcome the classical clustering algorithms. These algorithms optimize the number of clusters at each hierarchical level with the cluster validation criteria during the Minimum Spanning Tree construction process. Then the resultant hierarchy constructed by the DHCMST algorithm can very properly represent the underlying hierarchical structure of the data set. This representation alone improves the accuracy of the eventual result.

This DHCMST clustering algorithm addresses the issue of the undesired clustering structure and produces a large number of clusters. This algorithm does not require a predefined cluster number and the very same algorithm removes the inconsistent edges which are alone and satisfies the inconsistence measure. The particular clustering process is reiterated until the optimum number of clusters or regions are obtained. Finally this algorithm with the means of the dendrogram produces the optimum number of clusters as for the cluster of clusters means. This cluster of clusters is called the meta cluster.

The point set S in E^n has already been given and the hierarchical method starts by constructing a Minimum Spanning Tree (MST**)**. The weight of the edge in the tree is the Euclidean distance between the two end points (vertices/pixels). Given an image, the hierarchical method starts by means of constructing a Minimum Spanning Tree (MST**)**. This MST has been named as EMST1**.** Next the average weight Ŵ of the edges in the entire EMST1 and its standard deviation σ are computed; any edge with $W > \hat{W} + \sigma$ or *current longest edge* is removed from the tree. This leads to a set of disjoint sub trees ST = *{T1, T2 …}(which is a divisive approach)*. Each of these sub trees *Ti* is treated as a cluster or segment. The Minimum Spanning Tree based algorithm generates k number of the clusters. But yet another algorithm is available, the previous one, which generates k clusters with centers, which is used to produce the meta clusters or the meta similarity clusters. Both of these algorithms will have to assume the desired number of the clusters in advance; but, in practice, determining the number of the clusters is often coupled with discovering the structure too. Hence, a new algorithm is proposed, namely Divisive Hierarchical Clustering using Minimum Spanning Tree (DHCMST), which doesn't require a pre-defined cluster number and resultantly the study of the structure too. This particular algorithm works in two phases. The first phase of the algorithm partitions the EMST1 into the subtrees such as clusters or regions or segments. Fig. 3.2 shows the clusters connected through the points to construct

EMST1. The centers of clusters or regions or segments are identified using eccentricity of the points. These points are a representative point for the each sub tree *ST*. A point *ci* is assigned to a cluster / segment *i* if $ci \in Ti$. The group of center points is represented as *C={c1, c2……ck}*. These center points *c1, c2 ….ck* are connected and again minimum spanning tree EMST2 is constructed. The algorithm produces clusters with both intra-cluster and inter-cluster similarity. The intra-cluster category will have the documents within a cluster as very similar and the inter-cluster will have the documents in other clusters too as similar. The second phase of the algorithm converts the Minimum Spanning Tree EMST2 having optimal clusters into a dendrogram. The various levels in the dendrogram will indicate the least sum of similarity between the clusters.

Here, a cluster validation criterion based on the geometric characteristics of the cluster is used in which only the inter-cluster metric is acutely used. The DHCMST algorithm is the nearest centroid-based algorithm, which creates regions or sub trees (clusters / regions / segments) of the data space. The algorithm partitions a set *S* of data *D* in data space into *n* regions or clusters. Each region is represented by a centroid reference vector. Let *p* be the centroid representing a region (cluster / segment), all data within the region (cluster) are closer to the centroid *p* of the region than to any other centroid *q*:

$$R(p)=\{x \in D/dist(x,p) \leq dist(x,q)_q \ \forall q\}$$

Thus, the problem of finding the proper number of clusters of a data set can be transformed into the problem of finding the proper region or clusters of the data set.

Here, the MST is used as a criterion to test the inter-cluster property. Based on this observation, we use a cluster validation criterion, called Cluster Separation (CS) in DHCMST algorithm.

Cluster separation (CS) is defined as the ratio between the minimum and maximum edge of MST. ie., CS = *Emin / Emax* where Emax is the maximum length edge of MST, which represents two centroids that are at maximum separation, and Emin is the minimum length edge in the MST, which represents two centroids that are nearest to each other. Then, the CS represents the relative separation of the centroids. The value of CS ranges from 0 to 1. A low CS value means that the two centroids are too close to each other and the corresponding partition is not valid at all. A high *CS* value means the partition of the data is even and valid enough. In practice, a threshold is predefined to test the *CS*. If the *CS* is greater than the threshold, the partition of the dataset is valid. This process continues until the *CS* is smaller than the threshold. At that point, the proper number of clusters will be the number of clusters minus one. The *CS* criterion finds the proper binary relationship among clusters in the data space. The value setting of the threshold for the *CS* will be practical and is dependent on the dataset. Generally, the value of the threshold will be greater than 0.8. The CS value is less than 0.8, when the number of clusters is 4. Thus, the proper number of clusters for the data set is 3. Furthermore, the computational cost of *CS* is much lighter, because the number of subclusters is small. This makes the *CS* criterion practical for the DHCMST algorithm when it is used for clustering / segmenting large dataset (image) and to detect the outliers too. The goal is both to cluster / segment the graph optimally and to identify and isolate the outliers. Therefore, both connectivity and local structure is used in the definition of the optimal clustering. Here the notion of structure-connected clusters / segments is formulated , which extends that of a density based cluster and can distinguish good clusters / segments and outliers from the image graph. To detect the outliers from EMST, the *degree number* of points (objects / pixels) in the EMST is used. The DHCMST algorithm consists of the following steps:

1.Representation of the data points in the form of Dissimilarity Matrix (**DM**)

2.Construction of the Minimum Spanning Tree (**MST**) using **DM**

3.Generating the optimal number of the clusters using **MST**

4.Generating the Meta Cluster using the optimal number of clusters

Algorithm: **DHCMST**

Input: S, the point set

Output: Dendrogram with optimal number of clusters

Let e1 be an edge in the **EMST1** constructed from S

Let e2 be an edge in the **EMST2** constructed from C

Let W_e be the weight of e1

Let σ be the standard deviation of the edge weights in **EMST1**

Let S_T be the set of disjoint subtrees of **EMST1**

Let n_c be the number of clusters

```
Construct an EMST1 from S using prim's Algorithm
Compute the average weight of W and standard deviation $\sigma$ of the edges from EMST1
$S_T=\varphi$; nc=1; c=$\varphi$;
For each e1 EMST1
    If ($W_e > \hat{W}+\sigma$ or (current longest edge e1)
    Remove e1 from EMST1
$S_T = S_T \cup \{T\}$    // $T$ is a new disjoint subtree
$n_c=n_c+1$
Compute the center $C_i$ of $T_i$ using eccentricity of points
$C=U_{Ti} \in S_T\{C_i\}$
Construct an EMST2 T from C
$E_{min}$ = get-min-edge (T)
$E_{max}$ = get-max-edge (T)
$CS=E_{min}/E_{max}$
Until $CS< 0.8$
Begin with disjoint clusters with level L (0) = 0 and sequence number $m=o$
While (T has some edge)
e2 = get-min-edge (T)   // for least dissimilar pair of clusters
(a,b) = get-vertices (e2)
 Increment the sequence number m=m+1, merge the clusters (a)
 and (b), into a single cluster to form the next clustering $m$ and set the
 level of this cluster to L(m)=e2
Update T by forming new vertex by combining the vertices a, b
Return dendrogram with optimal number of clusters
```

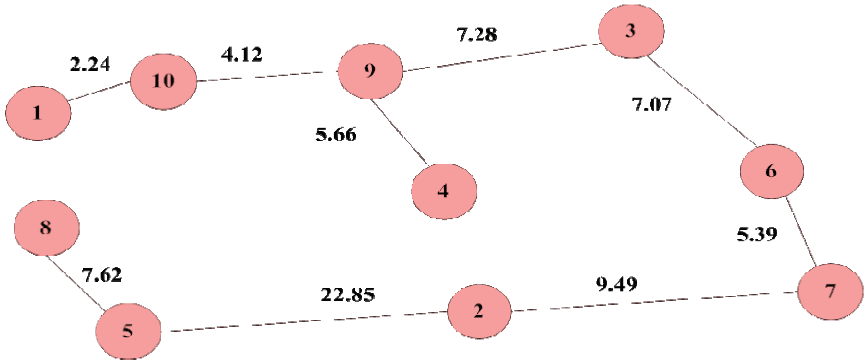


Fig. 3.2 The overview of the clusters connected through the points to construct EMST1

3.5.1 Hierarchical Clustering

These algorithms find successive clusters using previously established clusters. These can be agglomerative (“bottom-up”) or divisive (“top-down”). Agglomerative begins with each element as one cluster and merges them into a larger cluster. Divisive begins with the whole set and divides it into smaller clusters. These small clusters are called as the meta-clusters. Meta is a word which is a prefix and it indicates a concept of an abstraction. Meta comes to mean an alternation or alteration. Meta-clustering aims at grouping similar clusterings out of the input clusterings together. This hierarchical clustering is based upon the core idea of the objects being more related to nearby objects than to the objects farther off. These algorithms do connect the objects enough to form the clusters based on their distance. At different distances, different clusters will form which can still be represented using a dendrogram. These algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of the clusters that merge with each other at certain distances. In a dendrogram, the Y-axis marks the distance at which the clusters merge, while the objects do not ever merge when they are placed along the x-axis. Fig.3.3 depicts the hierarchical clustering wherein the clusters do not mix. Fig.3.4 shows the hierarchical clustering in which the clusters mix along the Y-axis.

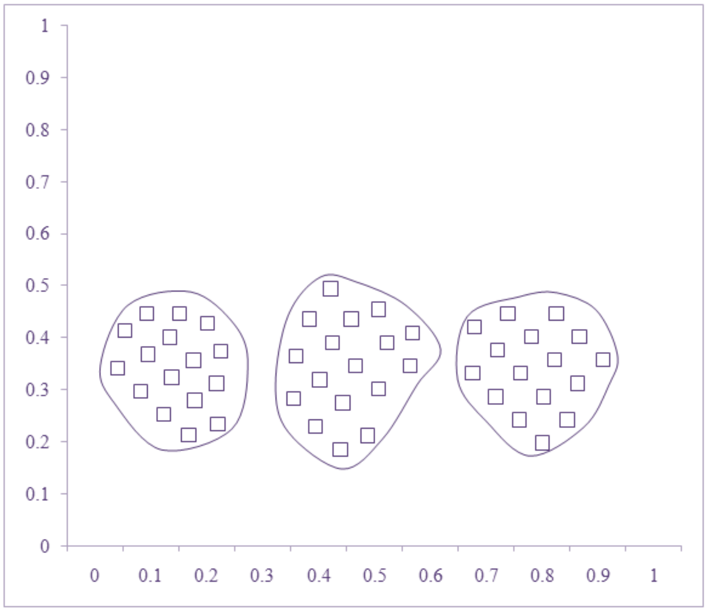


Fig. 3.3 The hierarchical clustering (the clusters don’t mix along the x-axis)

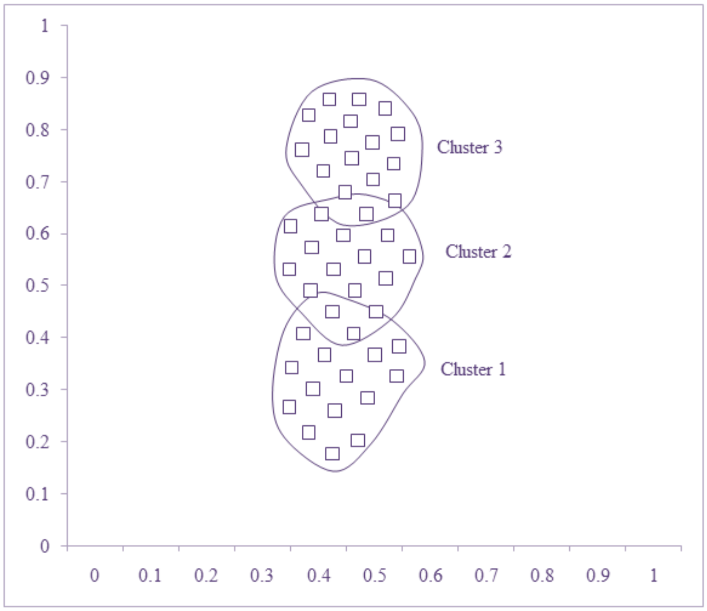


Fig 3.4 The hierarchical clustering (the clusters mix along the y-axis)

3.5.2 Clustering in information retrieval

The cluster hypothesis in information retrieval states that if a document from a cluster is relevant to a search request, it is likely that the other documents from the same cluster are also relevant. This is because clustering puts similar documents in similar clusters with respect to the relevance. The clustering process is said to have inputs of similar objects together as in a single cluster. Clustering is nothing but a process of conglomerating similar objects together into a single gathering as a cluster.

3.5.3 The Computational Complexity of the DHC

Density-based Hierarchical Clustering (DHC) suffers from a certain computational complexity which makes it inefficient for large data sets. Moreover, there is no versatile or an overall algorithm that can handle all types of clustering problems owing to the arbitrary shapes, variable densities and imbalanced sizes. As a result, enormous attention has been paid to design more effective clustering algorithms as in the recent decades. Among various graph-based clustering methods, Minimum Spanning tree (MST) has been paid more attention because of its very far reaching intuitive and effective data representation. MST has been extensively studied for biological data analysis, image processing, pattern recognition and outlier detection. The approach used in these fields is to construct MST over the given dataset and then remove the inconsistent edges to create the connected components. The repeated approach eventually leads to different clusters that are represented or formed by the connected components. Some authors have proposed three approaches, i.e., clustering by removing the longest MST-edges, an iterative clustering and a globally optimal clustering. Although, the methods of these authors are effective, users do not succeed in selecting the inconsistent edges for their removal without any prior knowledge to the structure of the data patterns. The approach used is based upon the maximizing or minimizing the degree of the vertices. But the method is computationally expensive. Wang *et al* have proposed a divide first and then conquer the next algorithm that uses the cut and the cycle property of MST. Some authors have reported a two round Minimum Spanning Tree based clustering algorithm. However, the algorithm is not strong enough to detect the outliers and overlapped clusters. Some other authors have proposed two MST based clustering algorithms. Their first algorithm produces *k*-partition of a set of data points for a given value of *k* which is difficult to predict for unknown data sets. Their second algorithm produces the clusters by removing the inconsistent edges from the MST. Motivated by this, a new clustering algorithm was designed, based on MST in which the coefficient of variation is used as a measure of inconsistency was incorporated. This algorithm does not depend on any prior knowledge of the parameters such as the number of clusters, initial cluster centers and the dimensionality of the data sets. The experimental results on two dimensional data sets are shown for visual purpose.

3.5.4 Euclidean Minimum Spanning Tree-based Algorithms

We propose two Euclidean Minimum Spanning Tree based clustering algorithms such as k-constrained and unconstrained algorithms. The k-constrained clustering algorithm produces neatly a k-partition of a set of points for any given k. The algorithm constructs a Minimum Spanning Tree of a set of representative points and removes edges that satisfy a predefined criterion or condition. The process is repeated until k number of clusters is produced. The unconstrained clustering algorithm partitions a point set into a group of clusters by maximally reducing the overall standard deviation of the edges in the Euclidean Minimum Spanning Tree constructed from a given point set, without prescribing the number of clusters. The experimental results are presented comparing the proposed algorithms with k-means and the Expectation-Maximization (EM) algorithm on both artificial data and benchmark data and then the results are transformed into the possible image color clustering.

The cost of constructing an EMST is O (n^2 log n), where n is the number of points. Such algorithms are capable of detecting clusters with irregular boundaries. The clustering process only detects the clusters with the irregular boundaries. In these algorithms, the points in the metric space are partitioned by the strategic removal of the edges of the EMST, generating sub trees each of which represents a cluster. The inherent cluster structure of a point set is closely related to the objects and the concepts that are embedded within that particular set. In practice, there are two general types of the clustering problems. In the first type, the number of embedded objects can be acquired successfully by means of the help of the application domain experts. Here the input to an algorithm will specify (in addition to the point set) the number of clusters, k, to be formed. In the current context, these are called k-constrained algorithms. In the second type of problem, the information on the number of embedded objects is hidden, and thereby unavailable as an input to the clustering algorithm. We term Such algorithms are termed unconstrained algorithms. In addition, some algorithms may also include as part of their input from a small number of tuning parameters whose values will depend on object characteristics or whose values will vary as per the character of the objects these parameters ie., the values of these parameters are directly proportional to the values of the objective characters or characteristics.

An example application of EMST-based algorithms is color clustering in web image analysis. Analyzing such images-for example in preparation for the extraction for the textual content-may be complicated by the image having complex backgrounds, and often many colors. In this domain, colors are represented as points in a three-dimensional (e.g., RGB, or HSV) color space.

Thus two EMST-based algorithms were presented: a k-constrained one, and also an unconstrained algorithm-address some of the shortcomings of the existing clustering algorithms, such as poor cluster discrimination, or (in the case of unconstrained algorithms) the generation of an excessive number of clusters. In the mid 80's, an O ($n^2 \log^2 n$) algorithm for the min-max diameter 2 clustering problem was found. Later some other authors gave an optimal O (n log n) algorithm using

Maximum Spanning Trees for minimizing the maximum diameter of a bipartition. Yet another author describes a method to remove inconsistent edges- edges, whose weights are significantly larger than the average weight of nearby edges -from the EMST. The inconsistent edges are to be removed is the vital criterion. His definition of inconsistent edges that relies on the concept of *depth-d neighborhoods*. Some other authors use an EMST-based algorithm to represent multidimensional gene expression data. They point out that an EMST-based clustering algorithm does not have to assume those points within a cluster which are grouped around centers or separated by a regular geometric curve. They describe three objective functions, corresponding to the k-constrained algorithms.

3.5.5 Hierarchical EMST-based Algorithm (HEMST) and MSTSSCIMLRO

The simple heuristic is used in some k-constrained EMST clustering algorithms, i.e. removing the longest edges, often does not work well. Image mining is more than just an extension of data mining to the image domain. Image mining is a technique commonly used to extract knowledge directly from the image. Image segmentation is the first step in image mining. Further the authors came forward to propose a novel algorithm called as Minimum Spanning Tree based Structural Similarity Clustering for Image Mining with Local Region Outliers (MSTSSCIMLRO) to segment the given image and to detect anomalous patterns (outliers). In MSTSSCIMLRO algorithm weighted Euclidean distance is used for edges, which is the key element in building the graph. MST-based image segmentation is a fast and an efficient method of generating a set of segments from an image. The algorithm uses a new cluster validation criterion based on the geometric property of data partition of the data set in order to find the proper number of segments which are called the clusters. The algorithm works in two phases. The first phase of the algorithm creates an optimal number of clusters or segments, whereas the second phase of the algorithm further segments the optimal number of clusters or segments and detects local region or the outliers.

3.6 DETECTING THE CLUSTERS WITH IRREGULAR BOUNDARIES

The Minimum Spanning Tree clustering algorithm is capable of detecting the clusters with irregular boundaries. The clusters with irregular boundaries should be detected for medical image segmentation. Otherwise the deep recognition of the attacks of the ailments cannot be viewed and previewed. There are two Minimum Spanning Trees based on clustering algorithms developed. The first algorithm produces k-clusters with center and guaranteed intra-cluster similarity. The radius and diameter of k-clusters are computed to find out the tightness of the k- clusters. The variance of the k-clusters is also computed to find the compactness of the clusters. The second algorithm is proposed to create a dendrogram using the k-clusters as objects with guaranteed inter-cluster similarity. The algorithm also finds out the central cluster from the k number of clusters. The first algorithm uses the divisive approach, whereas the second algorithm uses the agglomerative approach. Both the approaches are used to find out the Informative Meta-similarity clusters.

Given a connected, undirected graph $G=(V,E)$, where V is the set of nodes, E is the set of edges between the pairs of nodes, and weight $w\ (u\ ,\ v)$ is the specifying weight of the edge $(u,\ v)$ for each edge $(u,\ v)\ E$. A spanning tree is an acyclic sub graph of a graph G, which contains all the vertices from G. The Minimum Spanning Tree of a weighted graph is the minimum weight spanning tree of that particular graph. Several well established MST algorithms exist to solve the Minimum Spanning Tree problem. The cost of constructing a Minimum Spanning Tree is $O\ (m \log n)$, where m is the number of edges in the graph and n is the number of vertices. The diameter set of G is Dia (G) = $\{x \in V \mid e(x) = D\ (G)\}$. All existing clustering algorithm require a number of parameters as their inputs and these parameters significantly affect the cluster quality. This MST algorithm produces clusters of n-dimensional points with a given cluster number and a naturally approximate intra-cluster distance. Hierarchical clustering is a sequence of partitions in which each partition is nestled into the next in sequence for want of metaclusters. A Euclidean Minimum Spanning Tree (EMST) is a spanning tree of a set of n points in a metric space (E^n), where the length of an edge is the Euclidean distance between a pair of points in the point set. The hierarchical clustering approaches are related to graph theoretic clustering. Clustering algorithms using Minimal Spanning Tree take the advantages of MST. The MST ignores many possible connections between the data patterns, and so the cost of clustering can be decreased. The MST-based clustering algorithm is known to be capable of detecting clusters of various shapes and size. Unlike traditional clustering algorithms, the MST clustering algorithm does not assume a spherical shape structure of the underlying data. The EMST clustering algorithm uses the Euclidean Minimum Spanning Tree of a graph to produce the structure of point clusters in the *n*-dimensional Euclidean space. Clusters are detected to achieve some measure of optimality, such as minimum intra-cluster distance or maximum inter-cluster distance. The EMST algorithm has been widely used in practice.

Clustering by Minimal Spanning Tree can be viewed as a hierarchical clustering algorithm which follows a divisive approach. Using this method, an MST is constructed for a given input. There are different methods to produce a group of clusters. More efficient algorithms for constructing MST's have also been very extensively researched.

The length of the longest path in the graph G is called the *diameter* of the graph. The hierarchical clustering algorithm being employed dictates how the proximity matrix or proximity graph should be interpreted to merge two or more of these trivial clusters, thus nesting the trivial clusters into second partition. The process is repeated until a sequence of nested clustering is formed in which the number of clusters decrease as the sequence progresses until a single cluster containing all *n* objects, called the *conjoint clustering*, remains. An important objective of hierarchical cluster analysis is to provide the picture of data that can easily be interpreted. A picture of a hierarchical clustering is much easier for a human being to comprehend than a list of abstract symbols. A *dendrogram* is a special type of tree structure that provides a convenient way to represent hierarchical clustering. The EMST-based clustering algorithms are there to address the issue of undesired clustering structures. An EMST of a point set is constructed and the inconsistent edges are removed and then the particular process is repeated to create a hierarchy of clusters is created and thus *k* clusters are obtained.

3.7 FUZZY C-MEANS CLUSTERING

Fuzzy C-means (FCM) Clustering that has been widely used in image segmentation. However, in spite of its computational efficiency and widespread prevalence, the FCM algorithm doesn't take the spatial information of pixels into consideration, and hence may result in low robustness to the noise and may result in less accurate segmentation. The weighted image patch-based FCM (WIPFCM) algorithm has already been proposed for image segmentation. In this algorithm, image patches have been used to replace the pixels or the image patches have been used in the place of the pixels in the fuzzy clustering, and the pixels are placed in the image patches having constructed a weighting scheme to make the pixels able in each image patch to have anisotropic weights. An isotropy is the property of being directionally dependent, as opposed to isotropy, which implies identical properties in all the directions. Thus, this algorithm incorporates local spatial information embedded in the image into the segmentation process, and hence improves its robustness to the noise. This novel algorithm was compared to the several state-of-the-art segmentation approaches in the synthetic images and the clinical brain MRI studies. [Functional magnetic resonance imaging or functional

MRI (FMRI) is an MRI procedure that measures brain activity by detecting associated changes in blood flow. This technique relies on the fact that the cerebral blood flow and neuron activation are coupled. When an area of the brain is in use, blood flow to that region increases]. The results show that the WIPFCM algorithm can effectively overcome the impact of noise and substantially improve the accuracy of image segmentations.

Flat clustering is efficient and conceptually simple but it has a number of drawbacks. The algorithms in quest of flat clustering, return a flat unstructured set of clusters, and they require a pre-specified number of clusters as input and are nondeterministic. Hierarchical clustering (or Hierarchic clustering) outputs a hierarchy, a structure that is far more informative than the unstructured set of clusters returned by flat clustering. Hierarchical clustering does not require the user to pre-specify the number of clusters and most hierarchical algorithms that have been used in information retrieval are deterministic. These advantages of hierarchical clustering come at the cost of lower efficiency. The most common hierarchical clustering algorithms have a complexity that is at least quadratic in the number of documents compared to the linear complexity of *K*-means and EM.

Many researchers believe that hierarchical clustering produces far better clusters than flat clustering. Hierarchical clustering algorithms are either top-down or bottom up. Bottom up algorithms treat each document as a single cluster at the outset and then successively merge (or *agglomerate*) pairs of clusters until all clusters are merged into a single cluster that contains all documents. Bottom up hierarchical clustering is called Hierarchical Agglomerative Clustering (HAC). Top-down clustering requires a method for splitting a cluster.

More formally, a sequence is said to be a hierarchical clustering if 2 samples, c_1 and c_2, which belong in the same cluster at some level *k* remain clustered together at higher levels >*k*.

One example of a hierarchical clustering is a correspondence tree, or a dendrogram which shows how samples are grouped together. The first level shows all samples x_i as singleton clusters. As levels are increased, more and more samples are clustered together in a hierarchical manner. The hierarchical clustering procedures can be divided into two different approaches: agglomerative and divisive. The agglomerative approach for cluster analysis, used by the nearest and farthest neighbor algorithms, is a bottom-up clumping approach in which approach the user begins with *n* singleton clusters and successively merges clusters produce the other ones. The Minimum Spanning Tree, on the other hand, uses the divisive approach which is a top-down approach where the user starts with one cluster and successively splits clusters into producing the others.

In general, all agglomerative algorithms usually yield the same results if the clusters are compact and well separated. However, if the clusters are close to one another or if their shapes are not hyper spherical, different results can be expected. The space and time complexities are $O(n^2)$ and $O(cn^2d^2)$ respectively, where c is the number of clusters and d is the distance between them.

3.8 CONCLUSION

In this chapter an MST based algorithm is proposed for detecting the outliers and the metaclusters. Such outliers are malignant with their attacks though they are beneficial too. The metaclusters are detected to extract the images in image mining. The Decisive Hierarchical Minimum Spanning Tree(DHMST) algorithm, Minimum Spanning Tree-based Outlier Detection using Degree Number (MSTODDN), the Euclidean Minimum Spanning Tree (EMST) algorithm have all been proposed in this thesis for the image segmentation process or for extraction of the images. Such algorithms are used for the sake of data cleaning first which in turn gets culminated in image segmentation and which further gets culminated into medical image segmentation.

CHAPTER IV

CLASSIFICATION OF CLUSTERING ALGORITHMS AND ITS APPLICATION

4.1 INTRODUCTION

In this chapter, the related work done on clustering with minimum spanning tree have been discussed and compared. The existing clustering algorithms and their limitations are discussed. Clustering techniques are used in various fields like medical to remote sensing .density based clustering algorithms are used in remote sensing images. Hierarchal methods are used in medical images. Soft classification of clustering is as shown in the figure below.

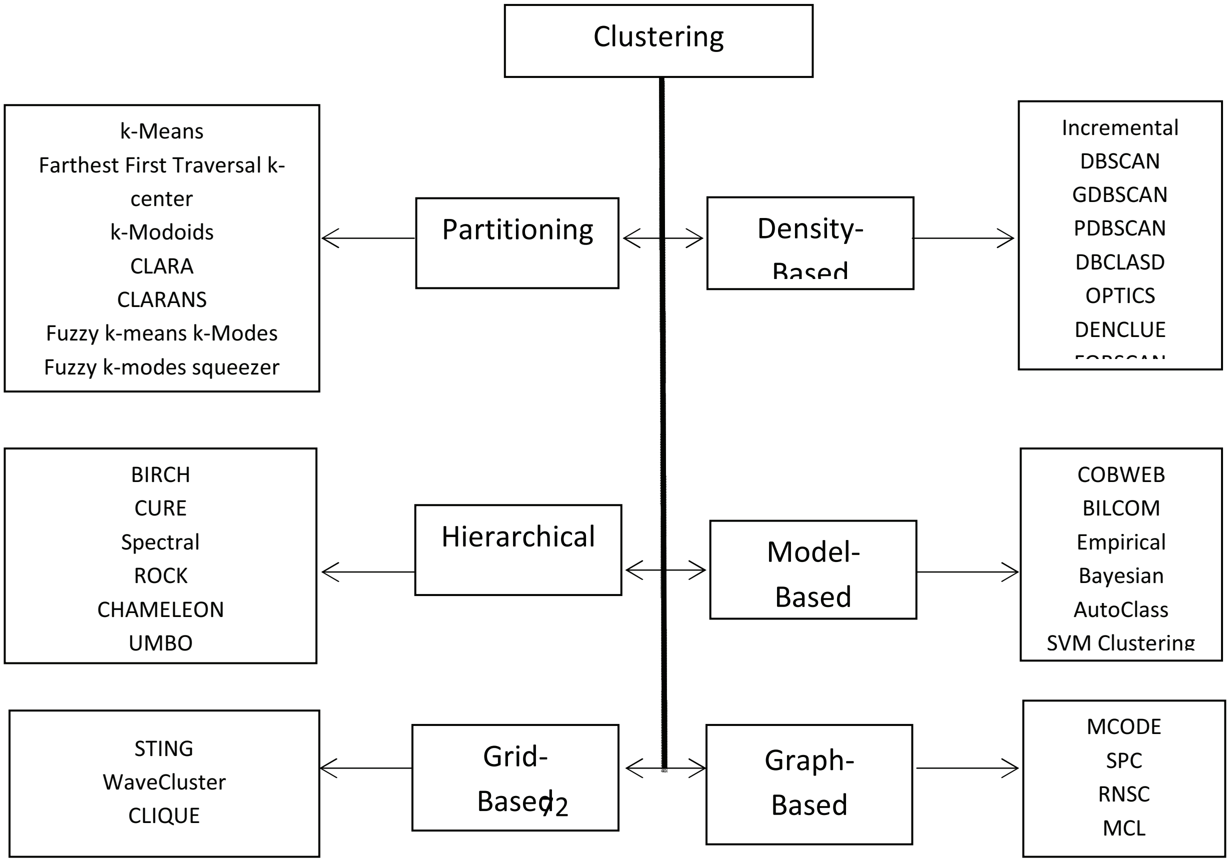


Fig. 4.1 Soft Classification of clustering Algorithms

The simplest k-constrained EMST-based algorithm is to remove k − 1 edges from the EMST, The simplest k-constrained EMST-based algorithm is to remove k − 1 edges from the EMST, resulting in k sub trees. Each cluster is the set of points in each subtree. In the mid 80's, Johnson[40] found an O ($n^2 \log^2 n$) algorithm for the min-max diameter 2 clustering problem. Asano, Bhattacharya, Keil, and Yao [4] later gave an optimal O (n log n) algorithm using maximum spanning trees for minimizing the maximum diameter of a bipartition. The problem becomes NP-complete when the number of partitions is beyond two . Asano also considered the clustering problems in which the goal is to maximize the minimum intercluster distance. They gave an O(n log n) algorithm for computing a k-partition of a point set by removing the k − 1 longest edges from the minimum spanning tree constructed from that point set [4].

Eldershaw and Hegland [23] re-examine the limitations of many 2D clustering algorithms that assume that clusters of a point set are essentially spherical, and provide a broader definition of a cluster based on transitivity: if two points p1 and p2 are close to the same point p0, they are both members of the same cluster. They present an algorithm which constructs a graph using Delaunay triangulation, and remove edges that are longer than a cut-off point. Next, they apply a graph partitioning algorithm to find the isolated connected components in the graph, and each discovered component is treated as a cluster. Unlike Zahn's method in which inconsistency is a locally determined property of an edge, they choose a cut-off point which corresponds to a global minimum. Bansal, Blum and Chawla [6] introduced an unconstrained algorithm called correlation clustering. The clustering problem they consider is a complete graph in which each edge is labeled qualitatively as “+” (similar) or “-” (dissimilar). The objective is to find the clustering that minimizes the number of disagreements with the edge labels. They proved NP-hardness of the fundamental problem, but provided approximation algorithms which have been further improved by others [12,6]. More recently, P¨aivinen [59] proposed a scale-free minimum spanning tree clustering algorithm which constructs a scale-free network and outputs clusters containing highly connected vertices. We will refer to this algorithm as SFMST. Xu Ying, Olman and Xu Dong [79] use an EMST-based algorithm to represent multidimensional gene expression data. They point out that an EMST-based clustering algorithm does not have to assume that points within a cluster are grouped around centers or separated by a regular geometric curve. They describe three objective functions, and corresponding k-constrained algorithms. The first objective function constitutes the implementation of SFMST. The second objective function is defined to minimize the total distance between the center and each data point in a cluster. This algorithm first removes k□1 edges from the tree, creating a k-partition. Next, it repeatedly merges a pair of adjacent partitions and finds its optimal 2-clustering solution. They observe that the algorithm quickly converges to a local minimum. The third objective function is defined to minimize the total distance between a representative point of a cluster and each other point in the cluster. The representatives are selected so that the objective function is optimized. This algorithm has exponential worst-case running time.

4.1.1 Hierarchical EMST-based Algorithm (HEMST)

The simple heuristic used in some k-constrained EMST clustering algorithms, i.e. removing k - 1 longest edge, often does not work well. Fig.4.1 [59] illustrates a typical example of the cases in which simply removing the k - 1 longest edges does not necessarily output the desired cluster structure. This is also a good example where the nearest neighbor or the single linkage [28] method does not work well.

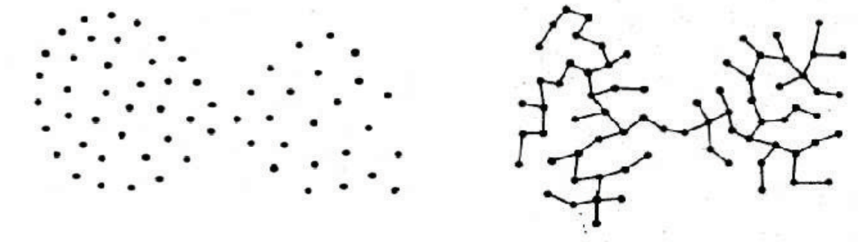


Fig. 4.2 Clusters connected through a point

An effective k-constrained algorithm, which is a hierarchically derived EMST-based clustering algorithm on a point set, it is known HEMST. To simplify discussion, in the current context we will refer to an EMST simply as *tree*. It can be noted that edge removal from such a tree is a closed operation on EMSTs since each of the resulting subtrees is also an EMST.

This is a point in the cluster closest to the geometric centroid of that cluster. For the sake of algorithmic orthogonality, the points in the input point set can be considered as representative points, each representing only themselves, and thus the data contained in tree nodes are representative points. A representative point has associated with it the set of representative points that it represents. The weight of a tree edge is the Euclidean distance between its incident representative points. We furthermore will omit the word *representative*, when the context is clear.

Let the input point set for the algorithm be S. The tree T(S) is initially partitioned in an unconstrained manner by removing all edges whose weight $w >$ AvgWeight(T) + StdDev(T), resulting in a set of subtrees T = {T1, T2, …), each representing an initial cluster. If $|T| \leq k$ then $k - |T|$ additional edges having greatest weight are removed from T(S), generating k subtrees, and k corresponding output clusters. If jTj > k, the centroid of each cluster is used to find a representative point pi for that cluster, i.e., pi = ReprPoint(Ti). At this stage, this set of representative points forms a new point set S′ = { Pi | Pi = ReprPoint(Ti) }. The process is recursively repeated until the number of clusters is k. Finally, the original point set belonging to each of the k clusters is recursively generated by the union of the point sets associated with each representative point in the cluster. The pseudo-code for HEMST is given in Algorithm 1

Function HEMST(**S**,*k*)

Input: Point set **S**, and required number of clusters, k

Output: Set of k point sets (set of clusters), S_K

```
$T \leftarrow CreateEMST(S)$;
$S_k \leftarrow \emptyset$;
$\overline{w} \leftarrow Avgweight(T)$;
$\sigma \rightarrow stdDev(T)$;
$\boldsymbol{E} \rightarrow \emptyset$;                                    /*E is a set of edge */
$\boldsymbol{forall}\ e \in EdgeSet(T)\ \boldsymbol{do}$
     $\boldsymbol{if}\ Weight\ (e) > \overline{w} + \sigma\ \boldsymbol{then}$
          $\boldsymbol{E} \leftarrow \boldsymbol{E} \cup \{\boldsymbol{e}\}$;
     $\boldsymbol{end}$
$\boldsymbol{end}$
$\boldsymbol{T_k} \leftarrow \boldsymbol{ParEMST}(\boldsymbol{T}, \boldsymbol{E})$;                   /*Tk is a set of subtrees */
$\boldsymbol{if}\ |\boldsymbol{T_k}| > k\ then$
    $\boldsymbol{P} \leftarrow \emptyset$;                                 /* P is a set of respresentative points */
    $\boldsymbol{forall}\ \boldsymbol{T} \in \boldsymbol{T_k}\ \boldsymbol{do}$
        $\boldsymbol{P} \leftarrow \boldsymbol{P} \cup \{ReprPoint(T)\}$;
    $\boldsymbol{end}$
    $HEMST(\boldsymbol{P}, \boldsymbol{k})$;                          /* Recursive call to HEMST */
$\boldsymbol{else}$
   /* PQ is a priority queue of edges in Tk in order of weights */
   $\boldsymbol{PQ} \leftarrow PriorityQueue(\boldsymbol{T_k})$;
   $\boldsymbol{for}\ i \leftarrow 1\ to\ k - |\boldsymbol{T_K}|\ \boldsymbol{do}$
        $\boldsymbol{E} \leftarrow \boldsymbol{E} \cup \{Dequeue(\boldsymbol{PQ})\}$;
   $\boldsymbol{end}$
    $\boldsymbol{T_K} \leftarrow ParEMST(\boldsymbol{T}, \boldsymbol{E})$;
   $\boldsymbol{forall}\ \boldsymbol{T} \in \boldsymbol{T_k}\ \boldsymbol{do}$
         $\boldsymbol{S_k} \leftarrow \boldsymbol{S_k} \cup \{\boldsymbol{PointSet}(ReprPoints(T))\}$;
    $\boldsymbol{end}$
    $\boldsymbol{return}\ \boldsymbol{S_k}$;
$\boldsymbol{end}$
```

Given that each iteration requires an amount of work bounded by O(n^2 log n) (finding the minimum spanning tree for the representative points, where n is the number of points in the (sub)tree), the algorithm has a complexity given by the recurrence

$$\begin{cases} T_n = 6T_{n/6} + n^2 log_2 n \\ T_n = 1 \end{cases}$$

Yielding a closed form complexity for HEMST of O(n^2 log n).

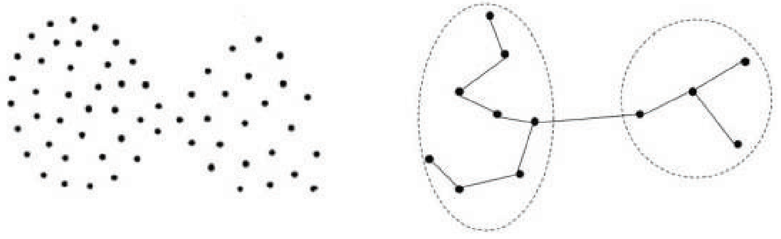


Fig.4.3 The representative points of the two clusters connected through a point

4.1.2 Maximum Standard Deviation Reduction Algorithm (MSDR)

The algorithm described next is an EMST-based unconstrained algorithm which tries to discover the underlying cluster structure in the input point set. It is based, unlike correlation clustering introduced by Bansal et al. [4], on removing tree edges that contribute the most to the tree's overall standard deviation of edge weights. The underlying idea is that those edges will tend to be inter-cluster rather than intra-cluster.

The algorithm first builds an EMST from the point set S, and removes edges from this tree one-by-one, portioning the zero-generation tree into successive generations of sets of subtrees. After i (i ≥ 0) edges removals (i.e. i th generation), the partition (set of subtrees) is denoted

$T^{(i)} = \{T_1^{(i)}, T_2^{(i)}, \dots\ , T_k^{(i)}\}$.

Let $\sigma(T^{(i)})$ represent the average standard deviation of the i^{th} generation partition, defined as,

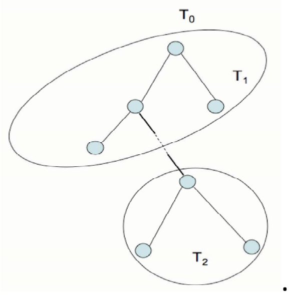
$$\boldsymbol{\sigma(T^{(i)}} = \frac{\sum_{\boldsymbol{j=1}}^{|\boldsymbol{T^{(i)}}|} \left|\boldsymbol{T_j^{(i)}}\right| \cdot \boldsymbol{\sigma(T_j^{(i)}}}{\sum_{\boldsymbol{j=1}}^{|\boldsymbol{T^{(i)}}|} |\boldsymbol{T_j^{(i)}}|}$$

Where, $\left|T_j^{(i)}\right|$ denotes the number of edges in the j th subtree of that generation. The standard deviation of the partition is the weighted average of the standard deviations of the weights of the edges in each subtree, weighted by the number of its edges. The edge to be removed from one of the subtree in $T^{(i-1)}$, is the one that will maximize the reduction of the partition standard deviation, in the i^{th}generation, i.e., $T^{th} = argmax(\Delta\sigma(T^{(i)}$, where $\Delta\sigma(T^{(i)}) = \sigma(T^{(0)}) - \sigma(T^{(i)}\ )$. The iterative edge removal process stops when

$|\Delta\sigma(T^{(i)}) - \Delta\sigma(T^{(i-1)})| < |\in\cdot\ (\Delta\sigma(T^{(i)}) + 1)|$.

```
**Function** MSDR (S)
**Input**: Point set S
**Output**: Set of k point sets (set of clusters), **Sk**
$\mathrm{T} \leftarrow \mathrm{CreateEMST(S)};$
$\mathrm{T}(_0) \leftarrow \{\mathrm{T}\};$
$i \leftarrow 0;$
**Repeat**
    **forall** $T \in \boldsymbol{T}^{(i)}$ **do**
        $\Delta\sigma_{max} \leftarrow 0;$
        **forall** $e \in EdgeSet(T)$ **do**
            $(T_{temp}, \Delta\sigma_{temp})\ ReduceStdDev\ (T^{(i)}, T, e)$ ;
            **if** $\Delta\sigma_{temp} > \Delta\sigma_{max}$ **then**
                $\Delta\sigma_{max} \leftarrow \Delta\sigma_{temp};$
              $T^{(i+1)} \leftarrow T_{temp};$
            **end**
        **end**
    **end**
    $i \leftarrow i + 1;$
**until** $\begin{cases} \sigma\left(T^{(0)}\right) < \sigma(T^{(1)} & for\ i = 1 \\ \left|\Delta\sigma\left(T^{(i)} - \Delta\sigma\left(T^{(i-1)}\right| <\right| \in\cdot \left(\Delta\sigma\left(T^{(i)}\right) + 1\right)\right| & for\ i > 1 \end{cases}$;
$f \leftarrow \mathrm{PolyRegression}\left(\bigcup_{j=1}^{i-1} \Delta\sigma\left(\mathrm{T}^{(j)}\right)\right);$
$k \leftarrow \min(j \in [1, i-1] | f'(j) = 0 \,\&\, f''(j) > 0);$
$S_k \leftarrow \emptyset;$
**forall** $T \in T^{(k)}$ **do**
    $\boldsymbol{S_k} \leftarrow \boldsymbol{S_k} \cup \{\boldsymbol{Points(T)}\};$
**End**
**Return Sk;**
```

Removing edge e from tree T_0, results in its being partitioned into two subtrees, T_1 and T2, as seen in the Fig.4.4



.

Fig.4.4 T_0 is split into T_1 and T_2

Based on the following properties:

$$\sum_{e_i \in T_1} e_i = \sum_{e_j \in T_0} e_j - \sum_{e_k \in T_2} e_k - e_r$$

$$\sum_{e_i \in T_1} e_i^2 = \sum_{e_j \in T_0} e_j^2 - \sum_{e_k \in T_2} e_k^2 - e_r^2$$

$$|T_1| = |T_0| - |T_2| - 1$$

The standard deviations of T1 and T2 can be calculated in the following manner,

$$\sigma(T_1 = \frac{\sum_{e_i \in T_1} {e_i}^2}{|T_1|} - \left(\frac{\sum_{e_i \in T_1} e_i}{|T_1|}\right)^2$$

$$\sigma(T_2 = \frac{\sum_{e_i \in T_2} {e_i}^2}{|T_2|} - \left(\frac{\sum_{e_i \in T_2} e_i}{|T_2|}\right)^2$$

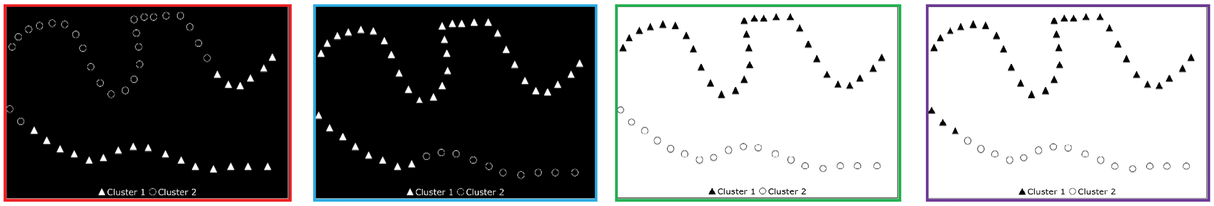
Since all the summations have been pre-computed, the calculation of both standard deviations can be done in constant time. Considering this fact, and the algorithm outlined in the pseudo-code, the complexity of the algorithm is $O(n^2 \log n + nk)$, where n is the cardinality of the point set, and k is the number of clusters generated. The only assumption is that the minimum of the regression $(\sim k)$ is reasonably porportional to the number of iterations actually performed (i.e.

generations generated). The first term in the complexity is from the building of the EMST.

4.1.3 Experimental Results: k-means,HEMST, EM and MSDR

There are three experiments to demonstrate the effectiveness of the proposed clustering algorithms. In the first experiment, three clustering problems are selected and compared the two proposed algorithms to k-means and EM respectively. In the next experiment the unconstrained MSDR algorithm with datasets from the UCI repository [51] are tested.

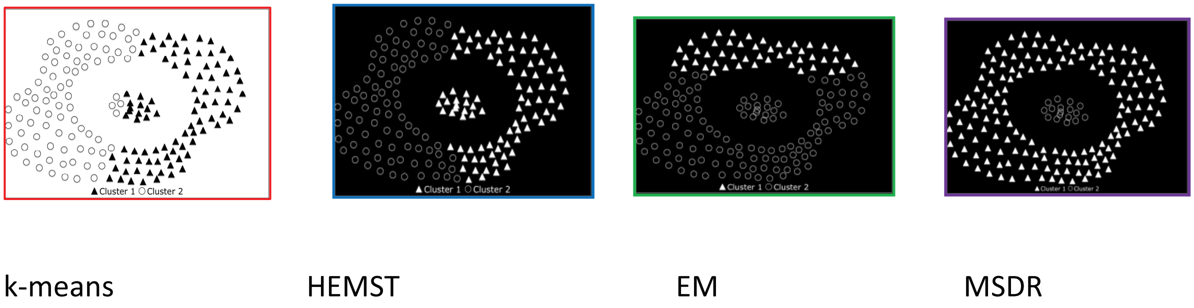
We selected three relatively difficult clustering problems in this experiment. The first problem is presented in Fig.4.5, in which two clusters are desired. Each cluster is formed as a curving irregular shaped line. The second problem shown in Fig.4.6 contains two clusters, with one inside the other. The third problem shown in Fig.4.7 contains two clusters, each with a non homogeneous density.



k-means

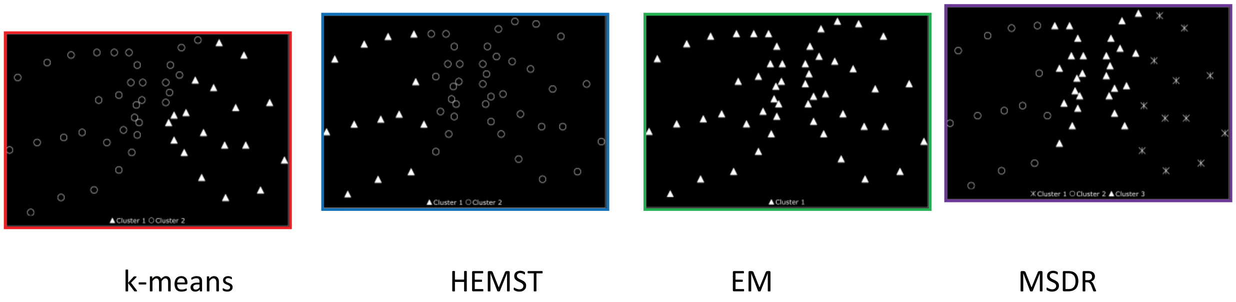
HEMST EM MSDR

Fig. 4.5 Two clusters formed by two lines



k-means HEMST EM MSDR

Fig. 4.6 Two clusters—one inside the other



k-means HEMST EM MSDR

Fig. 4.7 Clusters with non-homogeneous densities

Both HEMST and k-means are k-constrained. The EM algorithm determines the number of clusters through cross validation. Our MSDR algorithm selects the number of clusters with the largest second derivative. As we can see in Fig.4.5, k-means breaks both clusters and mixes them up into two undesired groups. HEMST and EM fragment one of the clusters, and only MSDR successfully identifies the clusters as desired. Similarly, in Fig. 4.6, k-means fragments both the surrounding and the central cluster, while HEMST and EM manage to identify the central cluster, although the surrounding cluster is separated into two. Again, only MSDR successfully outputs the more appropriate cluster structures. In Fig.4.7, both k-means and HEMST tend to group points in a high density region into one cluster. EM only outputs one cluster, while MSDR outputs three clusters,

4.1.4 Comparing MSDR, SFMST and k-means on UCI Datasets

MSDR clustering algorithm is compared with SFMST and k-means on four benchmark datasets from the UCI repository [51]. Table 1 gives the summary of the data sets.

Table 4.1 Benchmark Data Sets from the UCI Repository

Data Set	# of Instances	# of Attributes	# of Classes
Iris Plants	150	4	3
Pima-Indians Diabetes	768	8	2
Thyroid	215	5	3
Image Segmentation	2100	19	7

For the first three data sets: Iris Plants, Pima-Indians Diabetes and Thyroid, we compare the MSDR algorithm to several algorithms that have been reported by Paivinen [59], including the Scale-free MST (SFMST) algorithm, k-means with two different k values, and the standard MST algorithm.. The average entropy of clusters created by each algorithm are reported.. Lower entropy values imply purer clusters. To avoid bias towards a larger number of small clusters, the clusters containing less than 10 instances or 10% of the instances in the data set are ignored, whichever is smaller, when the entropy the entropy is computed..

Table 4.2 shows the results on the Iris Plants data. Our MSDR algorithm produced 3 non-trivial clusters. The first cluster contains the majority of *setosa* and only *setosa*, which conforms to the fact that *setosa* is linearly separable from the other two species. In fact, each algorithm manages to separate *setosa* from the other two, however, our MSDR algorithm is the only one that separates *versicolo* well from *virginica*. This also explains the very low (lowest) average entropy of the clusters produced by our MSDR algorithm.

Table 4.3 shows the results on the Thyroid data. Our MSDR algorithm produced 2 non-trivial clusters, of which the average entropy is significantly lower than that of the clusters produced by other MST based algorithms. The first cluster contains the majority of *Normal*. Clusters 3 to 6 solely contain the *Hypo* instances, while with the other two MST algorithms, *Hypo* instances cannot be identified. k-means, on the other hand, separated instances of different classes quite successfully.

Table 4.4 shows the results on the Pima Indians Diabetes data. The instances in this data set are not well separated using the Euclidean distance measure. Every algorithm failed to successfully separate the instances according to their labels. The MSDR algorithm produced 4 non-trivial clusters, of which the average entropy is significantly lower than that of the clusters produced by all other algorithms.

Table 4.2 Results on the Iris Data Set

	Seto	Vers	Virg	Entropy		Seto	Vers	Virg	Entropy
SFMST					MST				
C_1	44	1	0	0.15	C_1	1	45	30	1.06
C_2	1	35	28	1.09	C_2	36	0	0	0
C_3	0	0	17	0	C_3	0	4	7	0.95
					C_4	13	0	0	0
				Avg:0.41					**Avg:0.50**
k-means (k=5)					k-means (k=3)				
C_1	0	19	2	0.45	C_1	33	0	0	0
C_2	0	2	27	0.36	C_2	0	46	50	0.70
C_3	22	0	0	0	C_3	17	4	0	0.999
C_4	28	0	0	0					
				Avg:0.36					**Avg:0.57**
MSDR									
C_1	42	0	0	0					
C_2	0	40	0	0					
C_3	0	3	39	0.37					
				Avg:0.36					

In each section of the table, the first column is the clusters generated by a clustering algorithm. The last column is the entropy value of each cluster and the average entropy. The rest of the columns give the number of instances in each class.

EMST-based clustering algorithms are very effective when applied to various clustering problems. Our HEMST clustering algorithm is k-constrained. The algorithm gradually finds a set of k representative points that serve as an “attractor” to points nearby, and outputs the inherent cluster structure subsequently. The algorithm works much more reliably than the simple SEMST clustering algorithm. MSDR algorithm automatically determines the desired number of clusters. The objective function is defined to maximize the overall standard deviation reduction. This algorithm does not require the users to select and try various parameter combinations in order to get the desired output.

Table 4.3 Results on the Thyroid Data Set

	Nor	Hpe	Hpo	Entropy		Nor	Hpe	Hpo	Entropy
SFMST					MST				
C_1	37	1	2	0.45	C_1	118	8	30	0.98
C_2	69	20	0	0.66	C_2	7	24	0	0.77
C_3	0	0	12	0	C_3	18	3	0	0.59
B_n	17	5	16	1.43					
				Avg:0.64					**Avg:0.78**
k-means (k=5)					k-means (k=3)				
C_1	64	14	0	0.68	C_1	0	16	0	0
C_2	0	16	0	0	C_2	150	19	8	0.75
C_3	86	1	8	0.50	C_3	0	0	22	0
C_4	0	0	22	0					
				Avg:0.29					**Avg:0.25**
MSDR									
C_1	150	23	6	0.76					
C_2	0	0	13	0					
				Avg:0.38					

The main challenge encountered when using the MSDR algorithm is runtime efficiency .We reduced the runtime of the MSDR algorithm significantly by storing in each node some information that enables the Calculation of the standard deviation of each subtree in constant time. MSDR algorithm outperforms other clustering algorithms, including SFMST, standard MST, and k-means on most of the benchmark data sets from the UCI repository.

Table 4.4 Results on the Pima Indians Diabetes

	Negative	Positive	Entropy		Negative	Positive	Entropy
SFMST				MST			
C_1	9	7	0.99	C_1	224	138	0.96
C_2	15	27	0.94	C_2	9	1	0.47
C_3	26	12	0.90	C_3	9	10	0.998
C_4	133	38	0.76				
C_5	32	37	0.996				
B_n	33	31	0.999				
			Avg:0.93				**Avg:0.81**
k-means				k-means			
(k=9)				(k=5)			
C_1	27	23	0.995	C_1	102	9	0.41
C_2	13	26	0.92	C_2	72	43	0.95
C_3	8	9	0.997	C_3	19	39	0.91
C_4	56	20	0.83	C_4	47	52	0.998
C_5	33	38	0.996	C_5	8	9	0.997
C_6	23	22	0.996				
C_7	80	7	0.40				
C_8	5	6	0.99				
			Avg:0.89				**Avg:0.85**
MSDR							
C_1	215	107	0.92				
C_2	16	17	0.999				
C_3	0	11	0				
C_4	9	2	0.68				
			Avg:0.65				

Image segmentation has been recognized as the major crisis in medical image analysis and remains still a very big challenge in the area of research. Image segmentation is very commonly used in clinical and research applications to analyze the medical image data sets. Many image segmentation algorithms find their foundation in the signal processing theory and methods. Many researchers have proposed multiple modes for medical image segmentation. In data mining, *k*-means clustering is a method of cluster analysis which aims at partitioning *n* number of observations into *k*-clusters in which each observation belongs to the cluster with the nearest mean. This results in partitioning of the data space into Voronoi cells.

To perform image segmentation and edge detection tasks, many modes incorporate both region growing techniques and edge detection techniques. At first the edge detection techniques are applied to obtain the results of the Difference In Strength (DIS) map. Only then, the region growing techniques are employed to work on the map to obtain further results. The image segmentation approach which is solely based upon the multi-resolution edge detection method, region selection method, and the intensity threshold method are applied to detect the white matter structures in the brain. *K*-means clustering algorithm includes spatial constraints which account for the local intensity variations in the image and those variations do usually occur in the regions of the image. The number of clusters *k* is an input parameter. Inappropriate choice of *k* may yield poor results. Hence, when performing *k*-means, it is important to run diagnostic checks to determine the number of clusters in the data set. Convergence to a local minimum may produce such counterintuitive ("wrong") results.

4.2 Limitation of the *k*-means

A key limitation of *k*-means is its cluster model. The concept is usually based on the spherical clusters that are separable in a way so that the mean value converges towards the cluster center. The clusters are expected to be of similar size, so that the assignment to the nearest cluster center is the correct assignment in any other clusters obtained. For example, while applying *k*-means with a value $k = 3$ onto the well-known data set, the result often fails to separate the three Iris species contained in the data set.With $k = 2$, two visible clusters can be discovered, whereas with $k = 3$ one of the two clusters will be split into two even parts. In fact, $k = 2$ is more appropriate for this data set, despite the data set containing 3 *classes*. With any other clustering algorithm, the *k*-means result certainly relies upon the data set to satisfy the assumptions made by the clustering algorithms. It works very well on some data sets, while failing on others.

The result of the *k*-means can also be seen as the Voronoi cells of the cluster means. *K*-means clustering has been used to feature learning or dictionary learning, a step to (semi-)supervised learning. In this usage, clustering is performed on a large dataset, which need not be labeled.

This use of *k*-means has been successfully combined with simple, linear classifiers for semi-supervised learning in NLP (specifically for named entity recognition) and in computer vision. On an object recognition task, it was found to exhibit comparable performance with more sophisticated feature learning approaches to auto encoders and restricted Boltzmann machines.

Statistical supervised learning techniques have already been successful for many natural language processing (NLP) tasks, but they require labeled datasets, which can be expensive to obtain and the labeled data are highly refined, unlike the raw text. On the other hand, unlabeled data (raw text) are often available free in large quantities. Unlabeled data have shown promises in improving the performance of number of tasks, e.g. word sense disambiguation, information extraction, and the natural language parsing. Parsing is the process of analyzing a string of symbols either in a natural language or in the computer languages according to the rules of formal grammar. The term is also used in psycholinguistics when describing language comprehension and it also refers to the way, the human beings analyze a sentence or phrase in the spoken language or text in terms of the grammatical constituents identifying the parts of speech, syntactic relations, etc. In computer science, the term has been used in the analysis of computer languages, referring to the syntactic analysis of the input code into its component parts in order to facilitate the writing of the compilers and the interpreters. The goal of the named entity recognition is to detect the names of the people, organizations, and locations in a sentence. The goal of Chinese word segmentation is to find out the word boundaries in a sentence that has been written as a string of characters without spaces. In a preprocessing step, they use raw text to cluster the words and calculate the mutual information statistics. The output of this step is then used as features in a supervised model, specifically a global linear model trained in using the Perception algorithm. A new mode or methodology for the segmentation process of the medical image is proposed in this thesis, which is a combination of the graph based segmentation technique and the k-means clustering inverse random transformation technique with the threshold set, to increase the efficiency of image segmentation. This threshold is set with the means of user assistance.

In the first phase, the method reads the input image and obtains the gray scale image successfully. The obtained gray scale image is used to remove the background objects and then the histogram of the image will be obtained consequently. The resultant picture will be the input to construct the pixel adjacency graph, which is a construct graph, that is the graph with a set of pixels. The construction of an 8-neighbor pixel adjacency graph and edges will be very promptly assigned to the neighbouring pixels. The weight map will be calculated based on the similarity measure of the neighbouring pixels and according to the weight. The similar pixels are then clustered. These sets of processes will be executed repeatedly with user interaction and the user interaction will provide the threshold value. The threshold value is the limit for the similarity threshold value, which will be used to cluster the similar pixels. The result of the segmentation will be shown to the user and the system will wait for a new threshold value from the user.

4.3 Decision Tree Classifier

Decision trees are used to predict a categorical response known as class which is based on a collection of predictors which are called attributes. Tree based tools offer the following advantages:

- They can operate on both numerical and categorical measurements.
- They do not require any assumptions about either the distribution or the independence of attribute values. This is especially important for the fusion of measurements from different sources like spectral data, DEM data and other ancillary GIS data. DEM data are Digital Elevation Data. GIS is Geographic Information System.
- They are often easy to interpret, even by those with no expertise in statistical knowledge, by creating subgroups of data which the user may graphically analyze promptly.
- They are capable of dealing with the missing data too during both the training phase and the classification phase.

Most algorithms deal with the missing data by ignoring the objects with incomplete measurements. This is quite a waste because the remote sensing of the data is often hard and expensive to obtain. A common technique is to calculate impurities using only the attribute information present so that any object with at least one attribute will participate in training. Another ad hoc solution is to replace the missing attribute by its mean or a randomly generated value using a parametric model estimated using the training objects that are not missing from that attribute.

However, this bias distorts the marginal distributions of the attributes. It is proposed here to handle missing data using surrogate splits and a mimic or a substitute for the primary split which is the original, the surrogate splits being like clones or close approximations. The idea behind surrogate splits is to use the primary decision attribute at a node whenever possible, and to use alternative attributes when the object is missing that attribute. This can be achieved by an ordered set of surrogate splits for each non-leaf node. The first surrogate split maximizes the probability of making the same decision as the primary split, i.e. the number of objects that are sent to the descendant branches by both the primary and the surrogate splits are the same. If an object is missing all the surrogate attributes, the blind rule is used, i.e. it is sent to the descendant node that receives the most of the training objects.

4.3.1 Conversion of trees

At any time of the learning process, decision trees can be automatically converted into decision rules. This can be done by tracing the tree from the root node to each leaf node and forming logical expressions that make the initial set of rules. Occasionally, some of these rules can be very redundant, which means exceeding what is necessary or natural, or superfluous and they can be after all simplified without affecting the classification accuracy. The following schemes were investigated for the rule generalization:

- Lossless generalization where conditions that are completely redundant with respect to the other conditions are removed.

- Loss like generalization where conditions are removed using the greedy elimination.

This is done by comparing the error estimates of the original rule and the resulting rule. However, rules may not stay mutually exclusive even after the rule generalization process. To avoid conflicts, the rules are sorted in descending order of the probability values. If an observation satisfies none of the rules, it is assigned to the default class that appears most frequently in the training set.

4.3.2 Rule-based Classifier

A rule-based classifier has been developed that uses the simplified rule set to classify an image. This classifier also supports surrogate splits in the conditions and can handle very well the missing data. The rules are checked in descending order of the probability (confidence) values. A rule is satisfied if all of its conditions are satisfied. The default rule is used to assign the observations that do not satisfy any rule of the class with the highest frequency in the training data.

The proposed work in this thesis, is intended to improve the pruning of the decision tree classifier algorithm by clustering the distance boundaries and partitioning of uncertain probability distribution values. Clustering techniques do increase the speed of the decision tree construction and minimize the pruning time to a great extent. Distance boundary clustering technique works based on the criteria of the lower and the upper bound distances of the uncertain attribute values. Partitioning is done absolutely with the objective function introduced on probability distribution based on the density levels. Objective function evaluates the discrete value of the uncertain data item. Experiments are then planned to conduct performance evaluation of the heart disease diagnosis and predictions are made from such UCI repository data sets.

4.3.3 Pruning of the Decision Tree Classifier

The pruning of the decision tree classifier algorithm has to be improved and the proposal for the same is being made in this thesis. By means of clustering with the distance boundaries as well as clustering with the partitioning of the uncertain probability distribution values, the pruning of the decision tree classifier algorithm can be improved. The clustering techniques do increase the speed of the construction of the decision tree and the very same speed minimizes the pruning time to a great extent. This very same distance based clustering technique is based upon the criteria of lower as well as upper bound distances of the uncertain attribute values or the uncertain data values. Based on the density levels, the partition is being done and it is done with the objective function introduced on the probability distribution and the probability distribution itself is based upon the density levels. This very objective function which has been already introduced for the sake of partitioning, now evaluates the very discrete value of the uncertain attributes or the uncertain data items. Fig. 4.8 shows the Decision Tree Construction and the Pruning Time; and Fig. 4.9 shows the process of improving the pruning of the decision tree classifier algorithm.

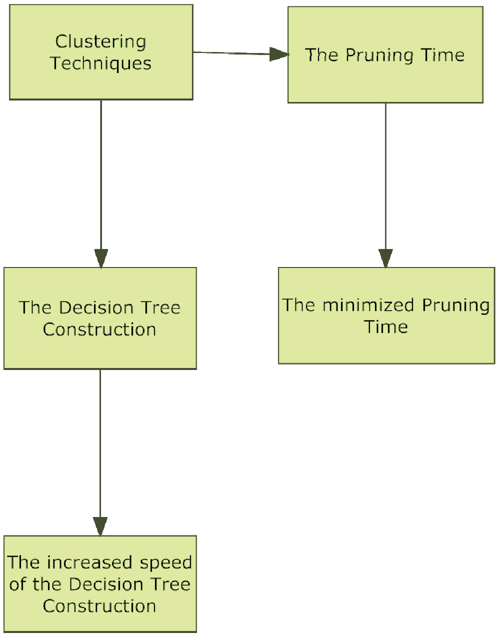

Clustering Techniques
The Pruning Time
The Decision Tree Construction
The minimized Pruning Time
The increased speed of the Decision Tree Construction

Fig. 4.8 The flow chart of the Decision Tree Construction and the Pruning Time

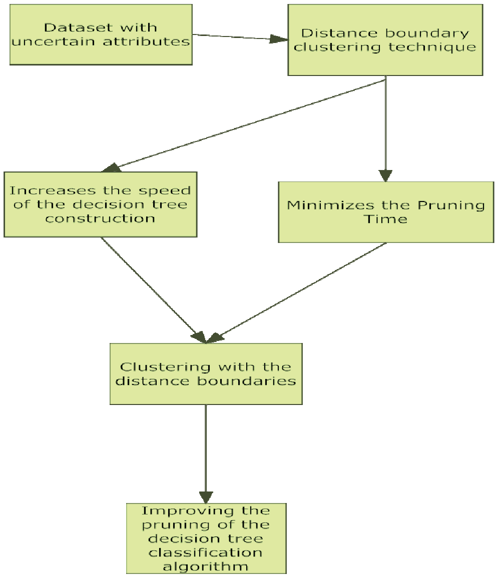

Dataset with uncertain attributes
Distance boundary clustering technique
Increases the speed of the decision tree construction
Minimizes the Pruning Time
Clustering with the distance boundaries
Improving the pruning of the decision tree classification algorithm

Fig. 4.9 The flow chart showing the process of improving the pruning time of the decision tree classifier algorithm

4.3.4 Marker-controlled watershed segmentation algorithm

Segmentation using the watershed transforms works well if the user can identify or mark the foreground objects and the background locations. The gradient magnitude of the primary segmentation is obtained by applying the Sobel filter. The Canny edge detector was also experimented on. But it was found that the results obtained by both the methods were comparable and it was decided that Sobel filter had higher complexity. In addition, Sobel filter had the advantage of providing a smoothing effect. Marker-controlled Watershed Segmentation follows this basic procedure:

- Compute enough a segmentation function. This is an image whose dark regions are the objects sought to be segmented;
- Compute also the foreground markers. These are the connected blobs of pixels within each of the objects;
- Compute the background markers. These are the pixels that are not part of any object;
- Modify the segmentation function so that it only has a minimum at the foreground and background marker locations;
- Compute further and finally get the watershed transformation of the modified segmentation function.

The proposed methodology is a two stage process. The first process uses *k*-means clustering to produce a primary segmentation of the input image, while the second process applies the marker controlled watershed segmentation algorithm to the primary segmentation to obtain the final segmentation map. Fig. 4.10 depicts the flow chart of the K-means clustering mode.

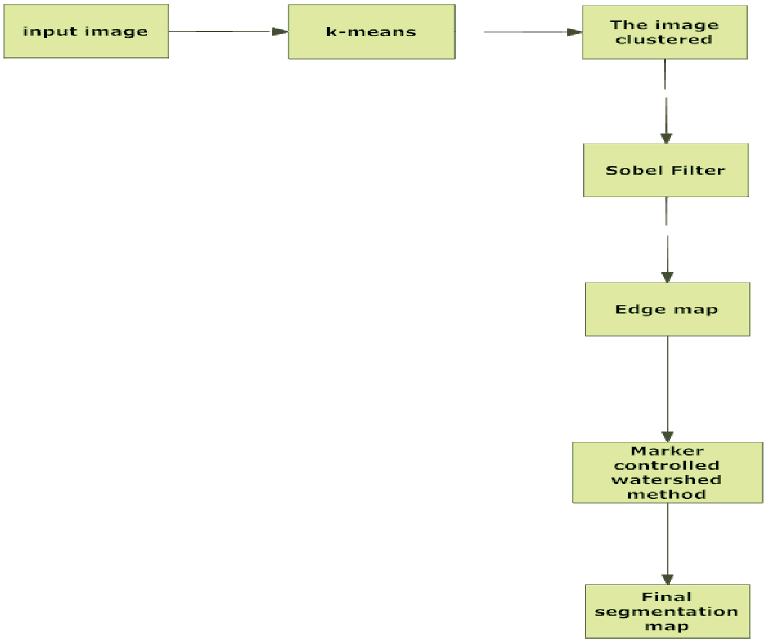


Fig. 4.10 Flow diagram of the *k*-means clustering method

4.3.5 Clustering - a very powerful tool

Clustering is a powerful tool that plays a very specific or a very deep as well as a very significant role for data analysis in various fields such as data mining, computational biology, etc. Gene expression data analysis has been extensively researched over to find the genes with similar expression patterns (co-expressed genes) from DNA micro-array data. A number of clustering algorithms have already been developed or designed or improved: which usually do face some problems for the practical ends. For example, K-means is very simple but robust; however, it requires users to provide or apply a number of clusters, which are not known prevalently and in advance. Any means that is not famous takes the users to a task indeed. K-means is the non-hierarchical method of clustering and this K-means clustering usually takes those many components in the final required number of the clusters. It examines each component meticulously and assigns it to one of the clusters depending on the minimum distance or to one of the clusters that is at a possible close distance. The centroid's position is recalculated every time a component is added to the cluster and this process continues until the components are grouped into the final required number of clusters. In data mining, K-means clustering is a method of cluster analysis which aims promptly at dividing the n number of observations or at partitioning the same n number of observations or objects or data into K clusters in which each observation belongs to the cluster with the nearest mean. In mathematics and physics, the centroid or geometric center of a two-dimensional region is, informally, the point at which a cardboard cut-out of the region could be perfectly balanced on the tip of a pencil when assuming uniform density and a uniform gravitational field. Formally, the centroid of a plane-figure which is in two-dimensional shape is the arithmetic mean or average position of all the points in the shape. The definition extends to any object in *n*- dimensional shape: its centroid is the mean position of all the points in all of the coordinate directions.Fig.4.11shows the Random Centroid Points.

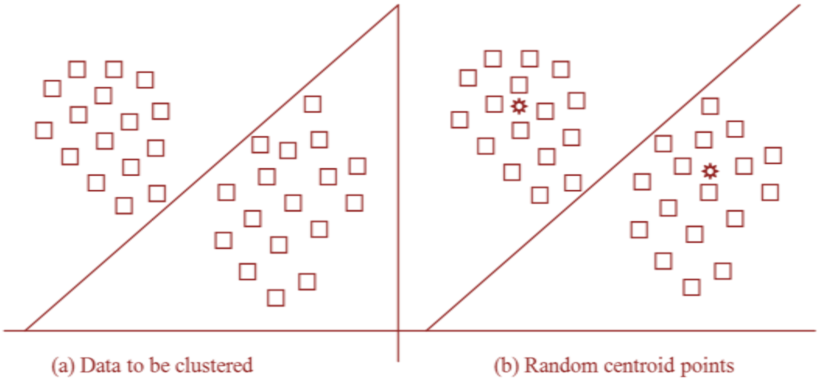


Fig. 4.11 The overview of the Random Centroid Points.

4.3.6 Cluster interfaced objective function for decision tree classifiers for mining data with uncertainty

Under the proposed uncertainty model, namely, Cluster Interfaced Objective Function, a feature value is denoted not by a single value, but by a PDF. For very practical reasons, PDF is deemed to be nonzero only within a bounded interval. A PDF is programmed analytically if it is denoted in the closed form. On the other side, processing a large number of sample points is rather much more expensive. In this work, it is shown that accuracy can be improved by deeming fit the uncertainty information. It is also proposed to prune the decision tree classifier algorithm that can greatly or highly reduce the computational effort. A decision tree under the uncertainty model resembles that of the point data model. The difference lies in the way the tree is employed to classify unseen test tuples whereas the point data model classifies the seen test tuples.

Decision classifier algorithm developed by clustering means with the distance boundaries and the partitioning of the uncertain probability distribution values is employed in this thesis. Clustering technique splits the data points into k partition, where each partition represents a cluster. That means, the data points are split into k number of the clusters. The partition is done based on a certain objective function. One such criterion function is to minimize the square error criterion which is computed as,

$$E = \Sigma\, \Sigma \,||\, p - mi \,||^{2} \qquad \text{------------------->}\ (1)$$

where p is the point in a cluster and mi is the mean of the cluster. The cluster should exhibit two properties, and they are (1) each group must consist of at least one object (2) each object must pertain to exactly one group.

1. Objective function: It defines objective function (F) for varied data types:

$$F = \sum_{x=1}^{K} V(d_x, c_y) \qquad \text{----------------->}\ (2)$$

where distance of data is $d_x$from the closest center $c_{y.}$

2. Selection: It calculates the probability distribution function (P_f) using the fitness value $F(S_f)$.

$$P_f = \frac{F(S_f)}{\sum_{f=1}^{F} F(S_f)} \qquad \text{------------------>}\ (3)$$

Where S_f denotes the solution.

3. Mutation: It is performed enough toward achieving a global optimum using the probability distribution function Pn.

$$P_n = \frac{1.5 \bullet d_{\max}(X_m) - d(X_n, C_k) + 0.5}{\sum_{k=1}^{K}(1.5 \bullet d_{\max}(X_n) - d(X_n, C_k) + 0.5} \text{ --------->} \ (4)$$

where d(Xn, Ck) is the Euclidean distance between pattern Xn and the centroid Ck, representing the maximum distance for the pattern Xn.

4. Finally convergence is well obtained by applying the *k*-means operator.

As discussed earlier, to identify the best attribute and split point for a node, Cluster Interfaced Objective Function has to inspect over the split points, where n = number of attributes, and k = number of tuples. For every such candidate attribute and split point f, an entropy E has to be computed.

Entropy measurements are the most computation-intensive fraction of the Cluster Interfaced Objective Function. The approach to implement more proficient algorithms is to move toward with strategies for pruning candidate split points and entropy calculations. The user is only pruning away the candidate split points which provide suboptimal entropy values. After pruning, one can still identify the optimal split points. It only eliminates suboptimal candidates from the deeming, thereby speeding up the tree building process.

Cluster Interfaced Objective Function for Decision Tree Classifiers attempts to prune heterogeneous intervals through a bounding technique. First, the entropy for all the end points is computed. Next, for each heterogeneous interval, a lower bound (Ln) is computed and the whole interval can be pruned. It may be noted that the number of end points is much smaller than the total number of the candidate split points. If a lot of heterogeneous intervals are pruned in this manner, many entropy calculations can be eliminated. So, the key to this pruning technique is to find out a lower bound that is not much too costly to compute, and yet that is reasonably tight for the pruning to be effective. Such a bound Ln can be derived as given below. Before that, a few symbols may be initiated to make the expression of the bound more compact and manageable:

$$y_c = F_{c,n}(-\infty, p); x_c = F_{c,n}(q,+\infty); k_c = F_{c,n}(p.q);$$

$$y = \sum_{c\varepsilon c} y_c ; x = \sum_{c\varepsilon c} x_c ; Y = y + (\sum_{c\varepsilon c} k_c) + x; \quad \text{-------->}\ (5)$$

$$V_c = \frac{y_c + k_c}{y + k_c}; and, g_c = \frac{x_c + k_c}{x + k_c}$$

where p and q denote bounded interval variables, c is a class label and $F_{c,x}$is a tuple count. By using equation (5) the lower bound value is computed in equation (6) which is given below.

$$L_n = -\frac{1}{Y} \sum_{c\varepsilon c} [y_c \log_2 V_c + x_c \log_2 g_c + k_c \log_2 (\max\{V_c, g_c\})] \quad \text{-------->}\ (6)$$

The calculation of the lower bound is related to the entropy calculation. It costs about the same as the computation of a split point’s entropy. Combining this heterogeneous interval pruning technique with those for empty and homogeneous intervals gives the user efficient Local Pruning algorithm of Cluster Interfaced Objective Function.

4.4 DECISION TREE CLASSIFICATION

In decision tree classification, a tuple's feature or attribute is either definite or statistical and the latter means that an apt point value will be typically implicit. However, the uncertainty of the data value or the data is general in most of the common applications.

In this thesis, the crisis of having created decision tree classifiers on unsure data certainty or unsure statistical attributes is revised. For that the following four steps are proposed:

1. An algorithm is to be devised to build the decision trees from the uncertain attributes or data using the clustering with the distance boundaries as well as with the partitioning of the uncertain probability distribution values.

2. Partitioning with the objective function is to be investigated. Partitioning is done with the objective function have been introduced on the probability distribution, based on the density levels. Objective function means the function done for a purpose.

3. Distance boundary clustering techniques based on the criteria of the lower and the upper bound distances have to be developed;

4. Such a theoretical base on which the pruning techniques are extracted for improving the decision classifier algorithm has to be implemented.

Uncertain data management is one of the research interests or the study interests in the recent years. There are two sorts of uncertainty: the existential uncertainty and the value uncertainty. Existential uncertainty is due to the uncertainty of either the object or the data tuple whether the availability or the existence of one or the other is possible, that is, one is not at all certain whether the particular object or the data tuple exists. Value uncertainty happens when an object or the data tuple is identified in existence but not with the expected certain value. Data uncertainty is typically captured by PDF's by means of their sets of the example values. A data item with value uncertainty is very typically denoted by means of a PDF over a limited and a bounded region of the promising values. The famous *k*-means clustering algorithm is expanded to the *uk*-means algorithm for the sake of extracting or clustering the uncertain data. k-means algorithm minimizes the sum of the squared errors whereas the UK-means minimizes the expected sum of the squared errors.

For decades, the decision tree classification upon the uncertain data was discussed strenuously. It was so discussed in guise or form of the absent values which were the uncertain values too. These absent values are due to the absence of the attributes or the attribute values which are absent because they don't exist. They don't exist either due to the data collected works which exist already and which may not still have the expected values or due to the data admission errors. Solutions comprise approximating absent values and mass values or inferring absent values either by means of the correct or probabilistic values using the classifier on the attribute. A plain method of filling in the absent values might even be adopted. By deeming the uncertain data gravely, the accuracy of the extracted data can be obtained with greater certainly. It is proposed to prune the decision tree classifier algorithm that can reduce the computational effort. In the decision classifier algorithm, the clustering technique splits the data points into k-partitions, where each partition represents a cluster and the partition is being done by a certain objective function. One such criterion function is to minimize the square error criterion. Finally convergence is obtained by means of applying the *k*-means operator. Entropy measurements are the most needed computation in the cluster interfaced objective function. If the user wants to implement more proficient algorithms, then they should move forward with the strategies for pruning the candidate split points and the entropy calculations. By pruning the end points, the cluster interfaced objective function minimizes the number of entropy calculations and increases the pruning efficiency and thus the very same function obtains a pruning effectiveness ranging from 83% to as much as 99%. The entropy calculations control the execution time of cluster interfaced objective functions, and such calculations are the only effective pruning techniques which can reduce indeed the tree construction time.

In many applications, data contains inherent uncertainty. A number of factors contribute to the uncertainty such as the random nature of the physical data generation and the collection process of the same, measurement error, and data stalling. The importance of deeming the data uncertainty explicitly in clustering with a focus on the quality of the clustering results is always recognized. An algorithm called *uk*-means has been proposed in this thesis. The algorithm was applied to the moving object uncertainty and was found to improve the accuracy of the clusters formed. Uniform distribution was assumed for the uncertainty associated with the data used. While the method can be very well generalized to the other distributions, there are significant efficiency issues that have not been addressed well. This thesis, however, focuses on the efficiency issues. Given a fixed clustering process, the ways to reduce the running time are studied. Efficiency is of particular importance in real time applications such as the one about mobile devices and so on. It is also important when a large amount of object data is involved. Traditional clustering processes often require the definition of a distance metric. For example, in K-means clustering, an object o is assigned to a cluster c such that the distance between o and a representative of c is the smallest among all the clusters. With multi-dimensional uncertainty, an object is no longer a single point feasible in the space but is represented by a PDF over an uncertainty region. The definition of the distance thus has to be deemed again. In this thesis, the choice is to use the expected distance as the distance measure. The expected distance metric not only provides an intuitive way of handling uncertainty, but also enables the development of efficient clustering algorithms.

For arbitrary PDFs, computing the expected distance of an uncertain object and a cluster representative requires very expensive integration operations, which are often hundreds or even thousands of times more expensive than a simple distance computation. Traditional clustering techniques have to be refined so that the algorithms become computationally feasible. One of the major contributions of this thesis is certainly the pruning techniques that can significantly reduce the number of expected distance calculations in the clustering process. There has been significant research interest in data uncertainty management in recent years. For value uncertainty, most work has been devoted to "the not precise queries," which provide probabilistic guarantees over the correctness of answers. Indexing solutions for the range queries over uncertain data have been already proposed by some authors. The same authors have also proposed solutions for aggregate queries such as the nearest neighbor queries. All these works of various authors have applied the study of uncertain data management to the simple database queries, instead of to the relatively more complicated data analysis and mining problems. Data clustering is one of the most studied areas in data mining research. Depending on application, there are several goals of clustering: to identify the (locally) most probable values of the model parameters (e.g., means of Gaussian mixtures), to minimize a certain cost function, or to identify high-density connected regions (e.g., areas with high population density). Expectation Maximization (EM) algorithm is expected to provide the maximum data distinguishing the accurate from the uncertain data.

4.5 CONCLUSION

The segmentation methodology such as k-means clustering uses the graph based user assisted image segmentation and inverse random transformation to improve the quality of the image meticulously. The proposed methodology can be used in various fields of vision technology and can be further extended to adapt the various number of filters at the last stage of the process after the inverse radon transformation. The algorithm proposed in this thesis can be further reviewed for medical image segmentation process.

Thus this thesis presents Cluster Interfaced Objective Function for Decision Tree Classifiers for mining uncertainty data which will in turn yield the medical image. Herein, the pruning of the Decision Tree Classifier algorithm has been improved by clustering with the distance boundaries and the partitioning of the uncertain probability distribution values. The clustering is achieved by the distance boundary clustering technique, based on the criteria of the lower and the upper bound distances of the uncertain attributes values. Partitioning and estimating the discrete value of uncertain data is done by the objective function. Relative entropy measure also is made on the lower and the upper bounded distances on the attribute characteristics related to the other certainty attributes in the data set. Experimental results would be carried out with the metrics such as Pruning Effectiveness and a number of entropy calculations and the execution time would certainly achieve much better classification accuracy.

CHAPTER V

THE PROPOSED CLUSTERING TECHNIQUE IN IMAGE SEGMENTATION THROUGH THE MINIMUM SPANNING TREE

5.1 INTRODUCTION

5.1.1 Fuzzy Clustering

Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" but "fuzzy" and “fuzzy” means soft. In the section 5.1 the introduction to the algorithm has been discussed, In 5.2 comparison of our proposed method with K-mean is described. 5.3 give the experimental results and methods 5.4 and 5.5 discusses about the comparative analysis using entropy. Application of clustering in image segmentation was discussed in the section 5.6. In this chapter with the proposed method the dissimilarity matrix is obtained. With DHCMST algorithm the cluster points are arrived and the dataset with 14 attributes are taken for execution and results are tabulated with execution time and a comparison is made with other methods.

The algorithm minimizes intra-cluster variance as well, but has the same problems as *k*-means. Using a mixture of Gaussians along with the expectation-maximization algorithm, a more statistically formalized method is formed which includes some of the partial membership in lasses. The graph of a Gaussian is a characteristic symmetric curve shaped bell that quickly falls to or toward zero. The parameter *a* is the height of the curve’s peak, *b* is the position of the center of the peak, and *c* the standard deviation which controls the width of the bell. The algorithm certainly minimizes the intra-cluster variation.

Another algorithm closely related to fuzzy c-means is soft k-means.

Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of the clustering process even under noise. Noise means an attack against the perfection of image of segmentation. Noise detection is an important aspect in the course of image segmentation. It only irritates the image segmentation process. The presence of noise hides the originality of the image segmentation process and the originality of the segmented images.

Image segmentation plays a crucial role in many medical imaging applications. A novel algorithm for fuzzy segmentation of magnetic resonance imaging (MRI) data is observed. The algorithm is realized by modifying the objective function in the conventional fuzzy c-means (FCM) algorithm using a kernel-induced distance metric and a spatial penalty on the membership functions. Initially, the original Euclidean distance in the FCM is replaced by a kernel-induced distance, and thus the corresponding algorithm is derived and called the kernelized fuzzy c-means (KFCM) algorithm, which is shown to be more robust than FCM. Then a spatial penalty is added to the objective function in KFCM to compensate for the intensity in homogeneities of MR image and to allow the labeling of a pixel to be influenced by its neighbors in the image. The penalty term regularizes and has a coefficient ranging from zero to one. Experimental results on both synthetic and real MR images show that the proposed algorithms have better performance when noise and other artifacts are present in the standard algorithms.

With the increasing size and number of medical images, the use of computers in facilitating their processing and analyzing has become very necessary. In particular, as a task of delineating anatomical structures and other regions of interest, image segmentation algorithms play a very principal role in numerous biomedical imaging applications such as the quantification of tissue volumes, diagnosis, study of anatomical structure, and computer-integrated surgery and the like.

Image segmentation is defined as the partitioning of an image into non overlapping, constituent regions which are homogeneous or similar with respect to some characteristics such as intensity or texture. Because of the advantages of magnetic resonance imaging (MRI) over other diagnostic imaging, the majority of researches in medical image segmentation pertain to its use for MR images, and there are a lot of a methods available for MR image segmentation. Among them, fuzzy segmentation methods are quite beneficial, because they can retain much more information from the original image than hard segmentation methods can. In particular, the fuzzy c-means (FCM) algorithm, assigns pixels to fuzzy clusters without labels (with labels means choice ones). The clusters without labels are the clusters void of a selection and they are not the choice clusters. Unlike the hard clustering methods which force pixels to belong exclusively or particularly to one class, FCM allows pixels to belong to multiple clusters with varying degrees of membership. Because of the additional flexibility, FCM has been very prevalently used in MR image segmentation applications recently.

However, because of the spatial intensity in homogeneity which is induced by the radio-frequency coil in MR image, conventional intensity-based FCM algorithms have proved to be problematic, even when the advanced techniques such as non-parametric, multi-channel methods are used. To deal with the homogeneity in problem, many algorithms have already been proposed by adding correction steps before segmenting the image or by modeling the image as the product of the original image and with a smooth varying multiplier field. Recently, many researchers have incorporated the spatial information into the original FCM algorithm to better the way of segmenting the images. Some authors have already proposed a fuzzy rule-based system to impose spatial continuity on FCM, and in another proposal, they have used a small positive constant to modify the membership of the center pixel in a window. Some other authors have modified the objective function in the FCM algorithm to include or to incorporate a multiplier field containing the first and second order information of the image very well.

Similarly, some authors have proposed an algorithm to compensate for the intensity of homogeneity and to label a pixel by deeming thickly to its immediate neighborhood. A recent approach proposed by yet another author is to penalize the FCM objective function to constrain the behavior of the membership functions - a method, very similar to the methods used in the regularization. In the domain of physics and probability in the Markov random field (MRF), the Markov network or an undirected graphical model is a set of random variables having a Markov property described by an undirected graph. A Markov random field is similar to a Bayesian network in its representation of the dependencies; the differences being that Bayesian networks are directed and acyclic (not cyclic), whereas Markov networks are undirected and may be cyclic. Thus, a Markov network can represent certain dependencies that a Bayesian network cannot (such as cyclic dependencies); on the other hand, it cannot represent certain dependencies that a Bayesian network can (such as induced dependencies). Induced dependencies (IDs) can be loosely defined as the functional dependencies that are consistent with the data currently held in the database but have not been defined within the database scheme. The induced dependencies are not necessarily true; they merely appear to be true with the data.

5.1.2 Particle swarm optimization

Particle Swarm Optimization (PSO), an evolutionary algorithm for optimization is extended to determine if natural selection, or survival-of-the-fittest, can enhance the ability of the PSO algorithm to escape from the local optima. To simulate selection many simultaneous parallel PSO algorithms, each one a swarm, operate on a test problem. Simple rules are developed to implement the selection. The ability of Darwinian PSO to escape local optima is evaluated by comparing a single swarm and a similar set of swarms, differing primarily in the absence of the selection mechanism, operating on the same test problem. The selection process is shown to be capable of evolving the best type of particle velocity control, which is a problem in the specific design choice of the PSO algorithm.

5.1.3 Darwinian PSO

The basic assumptions made to implement Darwinian PSO are:

- The longer a swarm lives, the more it has the chance of possessing the offspring. This is achieved by giving each swarm a constant, small chance of spawning a new swarm.
- A swarm will have its life-time extended (be rewarded) by finding a more fit state.
- A swarm will have its life-time reduced (be punished) for failing to find a more fit state.

These simple ideas implement an algorithm imitating natural selection. In nature, individuals or groups that possess a favorable adaptation are more likely to thrive and procreate. The favorable adaptation is assumed to prolong the lifetime of the individual. Unfavorable adaptations shorten the lifespan of an individual or group.

However, a general problem with the PSO and other optimization algorithms is that of becoming trapped in a local optimum. It may work in some problems but may fail in others.

In search of a better model of natural selection using the PSO algorithm, researchers have formulated what is called a Darwinian PSO, in which many swarms of the test solutions may exist at any time. Each swarm individually performs just like an ordinary PSO algorithm with some rules governing the collection of swarms that are designed to simulate natural selection. The selection process implemented is a selection of swarms within a constantly changing collection of swarms. The Pso algorithm is used for better classification compared with genetic algorithm.

5.1.4 The concept of entropy

In information theory, entropy is a measure of the uncertainty in a random variable. This uncertainty is measured in the guise of a measure called entropy. In this context, the term usually refers to a quantity which quantifies the expected value of the information contained in a message. In information theory, the concept of entropy is used to quantify the amount of information necessary to describe the macro state of a system. If a system presents a high value of entropy, it means that much information is necessary to describe its state. Depending on the specific application, entropy can be defined in different ways. Entropy can provide a good level of information to describe a given image. In this case, if all pixels in an image have the same gray level or the same intensity of color components, this image will present a minimal entropy value. On the other hand, when each pixel of an image presents a specific gray level or color intensity, then this image will exhibit maximum entropy. The pixel intensities are related to texture, because different textures tend to result in different distribution of gray level or color intensity and the entropy can be used for texture characterization. The texture approach is often based on this assumption about texture analysis.

Images are divided into the square windows with a fixed size L, the entropy is calculated for each window, and then a classification methodology is applied for identifying the category of the respective windows. The classification approach can be supervised or non-supervised. Supervised classification needs a training set composed by windows whose classes are previously known (prototypes), such as rural and urban areas. Here, a segmentation methodology based on supervised classification is focused on. Initially, the training is done by selecting samples. Each of these sample windows are selected in order to present pixels of only one class. The entropy is calculated for each color component based on the texture.

5.1.5 DHCMST

The DHCMST (Divisive Hierarchical Clustering Using Minimum Spanning Tree) can overcome many a shortcoming of the classical clustering algorithms. This algorithm optimizes the number of clusters at each hierarchical level with the cluster validation criteria during the Minimum Spanning Tree construction process. The resultant hierarchy constructed by the DHCMST algorithm can properly represent the underlying dataset’s hierarchical structure. The representation can improve the accuracy of the eventual result.

This DHCMST clustering algorithm addresses the issue of undesired clustering structures and unnecessarily large number of clusters. This algorithm does not require certain predefined cluster number but it removes inconsistent edges. The edges satisfy the inconsistence measure. The particular clustering process is reiterated until the optimum number of clusters or regions are obtained. Finally this algorithm with the means of the dendrogram produces the optimum number of clusters for the cluster of clusters. This cluster of clusters is called a metacluster.

The point set S in E^n has already been given and the hierarchical method starts by constructing a Minimum Spanning Tree (MST). The weight of the edge in the tree is the Euclidean distance between the two end points (vertices/pixels). Given an image, the hierarchical method starts by means of constructing a Minimum Spanning Tree (MST). This MST is named EMST1. Next the average weight $\hat{W}$ of the edges in the entire EMST1 and its standard deviation σ are computed. Any edge with $W > \hat{W} + \sigma$ or current longest edge is removed from the tree. This leads to a set of disjoint sub trees ST = $\{T1, T2 \ldots\}$(which is a divisive approach). Each of these sub trees Ti is treated as a cluster or segment. The Minimum Spanning Tree based algorithm generates k number of clusters. But there is yet another algorithm, the previous one, which generates k clusters with centers. This algorithm produces meta clusters or meta similarity clusters. Both of these algorithms will have to assume a desired number of clusters. But in practice, determining the number of the clusters is often coupled with discovering and structuring. Hence a new algorithm is proposed, namely Divisive Hierarchical Clustering Using Minimum Spanning Tree (DHCMST), which doesn't require a predefined cluster number. The same study is carried out for structure also. This particular algorithm works in two phases. The first phase of the algorithm partitions the EMST1 into sub trees such as clusters, regions or segments. Fig. 5.2 shows the clusters connected through the points to construct EMST1. The centers of clusters or regions or segments are identified using the eccentricity of the points. These points are a representative point for the each subtree *ST*. A point *ci* is assigned to a cluster/segment *i* if *ci* ∊ *Ti*. The group of center points is represented as *C = {c1, c2……ck}*. These center points *c1, c2, ...,ck* are connected and again the minimum spanning tree EMST2 is constructed. The algorithm produces clusters with both intra-cluster and inter-cluster similarity. The intra-cluster will have the documents within a cluster which are very similar and the inter-cluster will have the documents in other clusters which are also similar. The second phase of the algorithm converts the minimum spanning tree EMST2 having optimal clusters into a dendrogram. The very levels in the dendrogram will show that the least sum of similarity and the points between clusters differ.

Here, a cluster validation criterion based on the geometric characteristics of the clusters is used and only the inter cluster metric is acutely used. The DHCMST algorithm is the nearest centroid-based algorithm, which creates a region or sub trees (clusters/regions/segments) of the data space. The algorithm partitions a set *S* of data, data *D* in data space into *n* regions or clusters. Each region is represented by a centroid reference vector. If *p* is the centroid representing a region (cluster/segment), all data within the region (cluster) are closer to the centroid *p* of the region than to any other centroid *q*:

$$R(p)=\{x \in D/dist(x,p) \leq dist(x,q)_q \ \forall q\}$$

Thus, the problem of finding the proper number of clusters of a dataset can be transformed into the problem of finding the proper region or clusters of the dataset. Here, the MST is used as a criterion to test the inter-cluster property. Based on this observation, a cluster validation criterion, called Cluster Separation (CS) in DHCMST algorithm is used.

Cluster separation (CS) is defined as the ratio between minimum and maximum edge of MST ie., *CS = Emin / Emax* where Emax is the maximum length edge of MST, which represents two centroids that are at maximum separation, and Emin is the minimum length edge in the MST, which represents two centroids that are nearest to each other. CS represents the relative separation of the centroids. The value of CS ranges from 0 to 1. A low value of CS means that the two centroids are too close to each other and the corresponding partition is not valid at all. A high CS value means the partitions of the data are even and valid enough. In practice, a threshold is predefined to test the CS. If the CS is greater than the threshold, the partition of the dataset is valid. This again partitions the data set by creating a subtree or cluster or region. This process continues until the CS is smaller than the threshold. At that point, the proper number of clusters will be the number of clusters minus one. The CS criterion finds the proper binary relationship among the clusters in the data space. The value setting of the threshold for the CS will be practical and is dependent on the dataset. The higher the value of the threshold the smaller the number of clusters would be. Generally, the value of the threshold will be > 0.8. Fig 5.3 shows the CS value versus the number of clusters in hierarchical clustering. The CS value < 0.8 when the number of clusters is 4. Thus, the proper number of clusters for the data set is 3. Furthermore, the computational cost of CS is much lighter because the number of sub clusters is small. This makes the CS criterion practical for the DHCMST algorithm when it is used for clustering/segmenting large datasets (image) and to detect the outliers too. The goal is to cluster/segment graph optimally and to identify and isolate the outliers. Therefore both connectivity and local structure are used in the definition of optimal clustering. Here the notion of structure-connected cluster/segments is formulated, which extends that of a density based cluster and can distinguish good clusters/segments and outliers from the image graph. To detect the outliers from EMST, the degree number of points (objects/pixels) in the EMST is used.

5.2 COMPARISON OF FUZZY C-MEANS, PARTICLE SWARM OPTIMIZATION AND DARWINIAN PSO

Image segmentation generally refers to the process that partitions an image into mutually exclusive regions that cover the image. Among the various image segmentation techniques, traditional image segmentation methods like edge detection, region based, watershed transformation etc. are widely used but these image segmentation techniques have certain drawbacks, which cannot be used for any accurate result. In this thesis, clustering based techniques are employed on the images which result in the segmentation of the images. The performance of Fuzzy C-Means (FCM) integrated with the Particle Swarm optimization (PSO) technique and its variations are analyzed in the different application fields. To analyze and grade the performance and computational and time complexity of the techniques in different fields, several metrics are used namely the global consistency error, probabilistic rand index and variation of information. The rand index or rand measure in statistics, and in particular in the data clustering process is a measure of the similarity between two data clusterings. A form of the rand index may be defined as the adjustment for the grouping of the elements, and the very same can be called the adjusted rand index. From a mathematical standpoint, rand index is related to the accuracy, but is applicable even when class labels are not used.

This experimental performance analysis shows that FCM along with fractional order Darwinian PSO gives a better performance in terms of the classification accuracy as compared to other variations of other techniques used. The integrated algorithm tested on images proves to give better results visually as well as objectively. Finally, it is concluded that fractional order Darwinian PSO along with neighborhood fuzzy c-means and partial differential equation based level set method is an effective image segmentation technique to study the intricate contours provided the time complexity is as less as possible to make it more real time compatible.

A new multilevel thresholding method is proposed for segmentation of hyper spectral images into different homogenous regions. The new method is based on the Fractional-Order Darwinian Particle Swarm Optimization (FODPSO) which exploits the many swarms of test solutions that may exist at any time. In addition, the concept of fractional derivative is used to control the convergence rate of particles. The FODPSO is used to solve the so-called Otsu problem for each channel of the hyper spectral data as a grayscale image that indicates the spectral response to a particular frequency in the electromagnetic spectrum. In other words, the problem of n-level thresholding is reduced to an optimization problem in order to search for the thresholds that maximize the between-class variance. Experimental results successfully compare the FODPSO with the traditional PSO for multi-level segmentation of hyper spectral images. The FODPSO acts better than the other method in terms of both CPU time and fitness, thus enabling to find the optimal set of thresholds with a larger between-class variance in less computational time.

The novel algorithm was successfully compared with both the fractional-order PSO and the traditional DPSO. Significant progress has been made in the creative inspiration of bio-inspired computer algorithms applied to optimization, estimation, and control through the application of principles derived from the study of biology and it is presented as a swarm based model for trail detection in real-time. Experimental results on a large dataset revealed the ability of the model to produce a success rate of 91% using a 20 hz camera with a resolution of 480 - 640 that was carried through a scenario at an approximate speed of 1 m s_1. The authors compared the PSO and Bacteria Foraging algorithm (BF) with the Otsu method to determine the optimal threshold level for the deployment of sensor nodes. It should be noted that all methods were run offline and the PSO presented a superior performance when compared to the Otsu and the BF. Yet another author presented the application of the PSO to the field of pattern recognition and image processing. He introduced a clustering algorithm based on PSO. Further, he developed a dynamic clustering algorithm that could find the ‘‘optimum’’ number of clusters in a dataset with minimum user interference. Some other authors proposed a multilevel thresholding method based on PSO and compared their method with GA-based thresholding method. Results showed that the PSO-based image segmentation executed faster. PSO based image was more stable than GA. Few authors have applied FODPSO and DPSO to multilevel segmentation. A theoretical comparison is made with the fuzzy C-mean and PSO. The issues of various methods are discussed in this part.

5.3 RESULT AND DISCUSSION

5.3.1 Experimental result 1: Dissimilarity Matrix Representation

The students sample data with student id, semester marks are collected and used for the experiment. By using the distance measure data collected is clustered with the proposed method.

Most of the algorithm data points can very well be represented as a Dissimilarity Matrix (DM) representation. It contains the distance values between the data points and distance values are represented as the lower or the upper triangular matrix. The distance calculation measure is the Euclidean distance between the data points.

$$d(i,j) = \sqrt{\left(\left|X_{i,1} - X_{j,1}\right|^2 + \left|X_{i,2} - X_{j,2}\right|^2 + \cdots + \left|X_{i,n} - X_{j,n}\right|^2\right)}$$

where *i,j* are *n*-dimensional data points.

The sample data about the semester percentage marks of some post graduate students as shown in table 5.1may now be considered.

Table 5.1 Sample Data for semester marks(%)

Student ID	Semester I Mark	Semester II Mark
1	65	70
2	62	65
3	72	74
4	61	80
5	71	86
6	73	67
7	71	62
8	78	83
9	65	76
10	64	72

The DM for the above data is shown in the table 5.2.

Table 5.2 Dissimilarity matrix

	1	2	3	4	5	6	7	8	9	10
1	0	5.83	8.06	10.77	17.09	8.54	10.00	18.38	6.00	2.24
2		0	13.45	15.03	22.85	11.18	9.49	24.08	11.40	7.28
3			0	12.53	12.04	7.07	12.04	10.82	7.28	8.25
4				0	11.66	17.69	20.59	17.26	5.66	8.54
5					0	19.10	24.00	7.62	11.66	15.65
6						0	5.39	16.76	12.04	10.30
7							0	22.14	15.23	12.21
8								0	14.76	17.80
9									0	4.12
10										0

Table 5.3 Minimum spanning tree edges

Edge	Euclidean Distance/ Weight
[1,10]	2.24
[10,9]	4.12
[9,4]	5.66
[9,3]	12.04
[3,6]	7.07
[6,7]	5.39
[7,2]	9.49
[2,5]	22.85
[5,8]	7.62

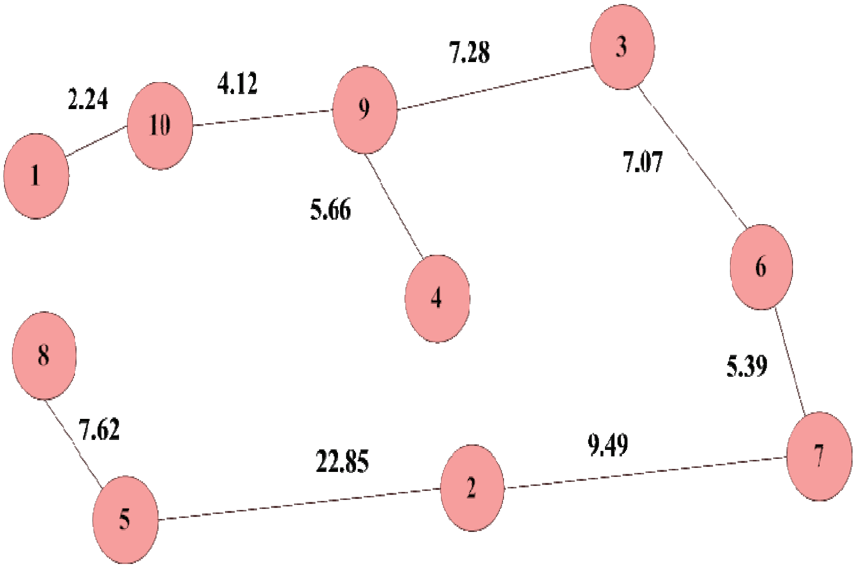


Fig. 5.1 Clusters connected through points-EMST1

5.3.2 Construction of EMST1

The DHCMST algorithm constructs EMST1 from the dissimilarity matrix as shown in Fig. 5.1. The mean $\widehat{W}$ and the standard deviation σ of the edges from EMST1 are computed as 7.986 and 5.966 respectively. The sum of the mean $\widehat{W}$ and the standard deviation σ is computed as 13.936. This value is used to identify the inconsistent edges in EMST1 to generate the clusters or the sub trees. Based upon the same value, the edge weighing 22.85 connecting the vertices 2 and 5 is found to be an inconsistent one. And by removing this inconsistent edge from EMST1, vertices or data points in EMST1 are partitioned into two sets or two sub trees or two clusters. And these two sets namely T1 and T2 are represented as T_1={1,10,9.4,3,6,7,2} and T_2={5,8} and are shown in Fig. 5.2.

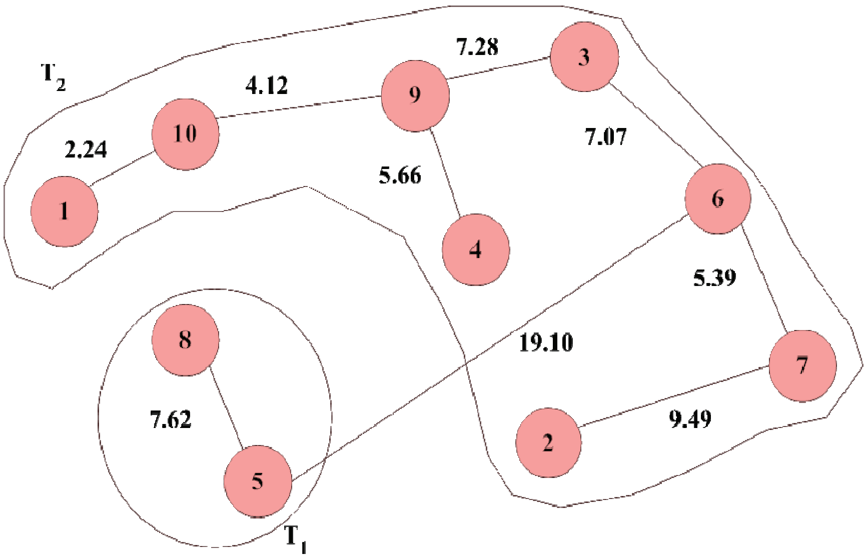


Fig. 5.2 EMST2 from three clusters using center points 5 and 6

Center point (vertex) for each of the subtree is found by using the eccentricity of the points (vertices). These center points or vertices are connected and again another Minimum Spanning Tree EMST2 (Fig. 5.2) is well constructed. Using EMST2, the minimum edge (E_{min}=19.1) and the maximum edge (E_{max}=19.1) are found to compute the cluster separation value (CS=1). If the CS is > 0.8, then it is concluded that the sub trees or the clusters created are well separated. Next the inconsistent edges are identified based on the inconsistency condition to create the sub trees or the clusters again. Center point for each of the subtree or cluster is found using the eccentricity of the points. Once again by connecting the vertices EMST2 is constructed. And the new CS value is also computed. Thus the process is repeated so as to create the optimum number of clusters as shown in Fig. 5.3.

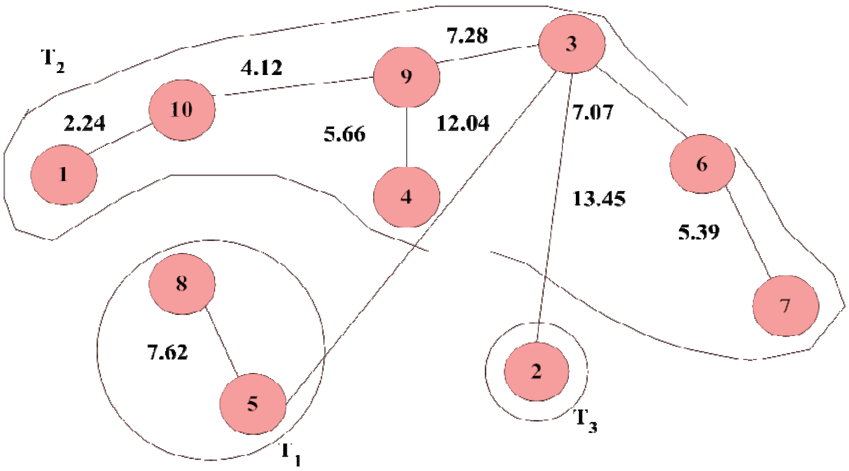


Fig. 5.3 EMST2 from 3 clusters using center points 5, 3 and 2

For the given data, the above process terminates when E_{min}=7.62, E_{max}=13.45 and CS=0.566 as shown in Fig. 5.4.

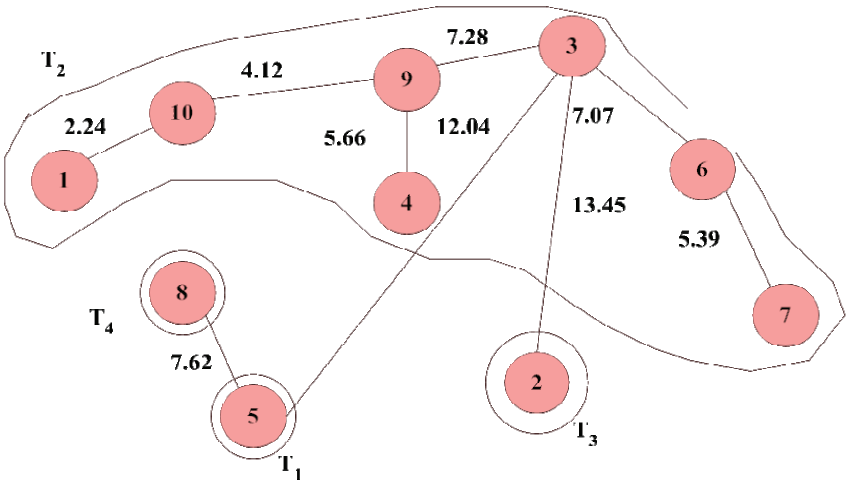


Fig. 5.4 EMST2 from 4 cluster center points 8, 5, 3 and 2

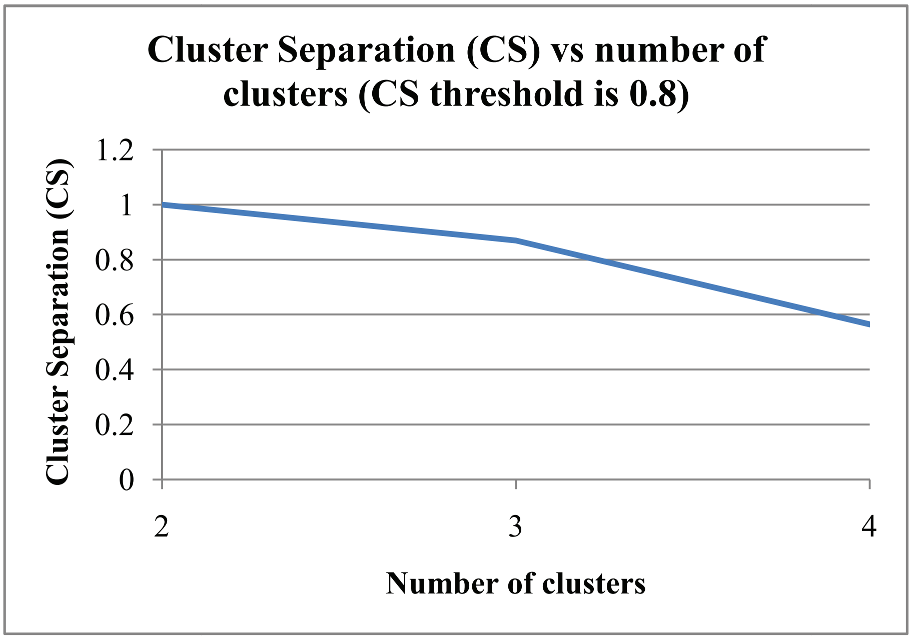


Fig. 5.5 Number of clusters Vs clusters separation

The DHCMST algorithm creates three well separated disjoint sub trees or clusters for the given data namely T_1={5,8},T_2={1,10,9,4,3,6,7} and T_3{2}. Each of the subtrees Ti is deemed as the cluster and based on the above subtrees or clusters, the DHCMST algorithm creates a cluster of clusters using a hierarchical mode and this cluster of clusters is called metacluster. The metaclusters are represented in the form of a dendrogram as shown in Fig. 5.5.

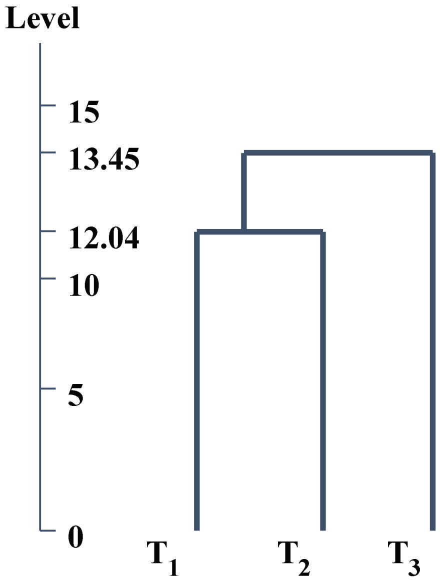


Fig. 5.6. Dendrogram for optimal metacluster

From the outcome of the clustering, it is clear that the students bearing ID numbers 5 and 8 are very good in the first two semester examinations whereas the student with IS number 2 is average in the first two semester examinations. The remaining students are neither too good nor average in the two semester examinations.

5.3.3 Experiemental result 2: Cluster Interfaced Objective Function

The experiment is carried out for the second set of data of patient with heart disease and 14 attributes are taken in account for execution and the results are plotted. To explore the potentiality of achieving superior classification accuracy by considering the data uncertainty, Cluster Interfaced Objective Function has been applied to the heart disease diagnosis and prediction taken from the UCI Machine Learning Repository. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are very regularly used by the machine learning community for the empirical analysis of machine learning algorithms. It is practically used by students, educators, and researchers all over the world as a primary source of machine learning data sets. The data set is chosen because it contains mostly numerical attributes generated from measurements. For the sake of experiments, classifiers are constructed on the numerical attributes and their "class label" attributes. The attributes of the dataset used are given in table 5.4.

Table 5.4: Sample dataset of heart disease diagnosis

age	Sex	cp	Trest Bps	chol	fbs	Restecg	thalach	exang	Old peak	slope	c a	thal	num
63	M	typ_ angina	145	233	t	left_vent_ hyper	150	no	2.3	down	0	fixed_ defect	'<50'
67	M	asympt	160	286	f	left_vent_ hyper	108	yes	1.5	flat	3	normal	'>50_1'
67	M	asympt	120	229	f	left_vent_ hyper	129	yes	2.6	flat	2	reversable_ defect	'>50_1'
37	M	non_ anginal	130	250	f	normal	187	no	3.5	down	0	normal	'<50'
41	F	atyp_ angina	130	204	f	left_vent_ hyper	172	no	1.4	up	0	normal	'<50'
56	M	atyp_ angina	120	236	f	normal	178	no	0.8	up	0	normal	'<50'
62	F	asympt	140	268	f	left_vent_ hyper	160	no	3.6	down	2	normal	'>50_1'
57	F	asympt	120	354	f	normal	163	yes	0.6	up	0	normal	'<50'
63	M	asympt	130	254	f	left_vent_ hyper	147	no	1.4	flat	1	reversable_ defect	'>50_1'
53	M	asympt	140	203	t	left_vent_ hyper	155	yes	3.1	down	0	reversable_ defect	'>50_1'
57	M	asympt	140	192	f	normal	148	no	0.4	flat	0	fixed_ defect	'<50'
56	F	atyp_ angina	140	294	f	left_vent_ hyper	153	no	1.3	flat	0	normal	'<50'
56	M	non_ anginal	130	256	t	left_vent_ hyper	142	yes	0.6	flat	1	fixed_ defect	'>50_1'
44	M	atyp_ angina	120	263	f	normal	173	no	0	up	0	reversable_ defect	'<50'
52	M	non_ anginal	172	199	t	normal	162	no	0.5	up	0	reversable_ defect	'<50'
57	M	non_ anginal	150	168	f	normal	174	no	1.6	up	0	normal	'<50'
48	M	atyp_ angina	110	229	f	normal	168	no	1	down	0	reversable_ defect	'>50_1'
54	M	asympt	140	239	f	normal	160	no	1.2	up	0	normal	'<50'
48	F	non_ anginal	130	275	f	normal	139	no	0.2	up	0	normal	'<50'
49	M	atyp_ angina	130	266	f	normal	171	no	0.6	up	0	normal	'<50'

In our experiment, we use 14 attributes which are described below:

1. Age: age in years
2. Sex: sex (1 = male; 0 = female)
3. Cp: chest pain type

-- Value 1: typical angina

-- Value 2: atypical angina

-- Value 3: non-anginal pain

-- Value 4: asymptomatic

4. Restbps: resting blood pressure (in mm Hg on admission to the hospital)

5. Chol: serum cholesterol in mg/dl

6. Fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

7. Restecg: resting electrocardiographic results

-- Value 0: normal

-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

8. Thalach: maximum heart rate achieved

9. Exang: exercise induced angina (1 = yes; 0 = no)

10. Oldpeak = ST depression induced by exercise relative to rest

11. Slope: the slope of the peak exercise ST segment

-- Value 1: up sloping

-- Value 2: flat

-- Value 3: down sloping

12. ca: number of major vessels (0-3) colored by fluoroscopy

13. Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

14. Num: diagnosis of heart disease (angiographic disease status)

-- Value 0: < 50% diameter narrowing

-- Value 1: > 50% diameter narrowing

To build a decision tree on uncertain data with a combination of the numerical and the categorical attributes, the tree is built recursively in a top-down manner, starting from the root. At each node, all the possible attributes (numerical or categorical) are considered, studied and scrutinized. For each attribute, the entropy of the split is calculated and the attribute giving the highest information gain is selected. The node is assigned that attribute and split point, (if it is a numerical attribute) and the tuples are (fractionally) propagated to the child nodes. Each child node is then further processed successfully.

To evaluate the entropy of a categorical attribute, the tuples in question are split into a set of buckets. Each tuple is copied as a new tuple in which the PDFs are inherited, except for the attribute. The entropy for the split on attributes is calculated using all the buckets. As a heuristic, a categorical attribute that has already been chosen for splitting in an ancestor node of the tree need not be reconsidered, for it will not give any information gain if the tuples in question are split based on that categorical attribute again.

Extensive experiments are conducted to show that the resulting classifiers are more accurate than those using the value averages i.e., the statistical derivatives. Processing probability density functions is computationally more expensive than processing the single values (e.g., averages). Decision tree construction on uncertain data consumes more CPU time. Pruning techniques are adapted to improve the decision tree construction efficiency definitely.

Table 5.5 Pruning effectiveness

Number of Entropy Calculation	Pruning Effectiveness (%)	
	UDT-ES	Cluster Interfaced Objective Function
100	69	99
200	68	98
300	60	96
400	64	97
500	67	84
600	62	85
700	54	83

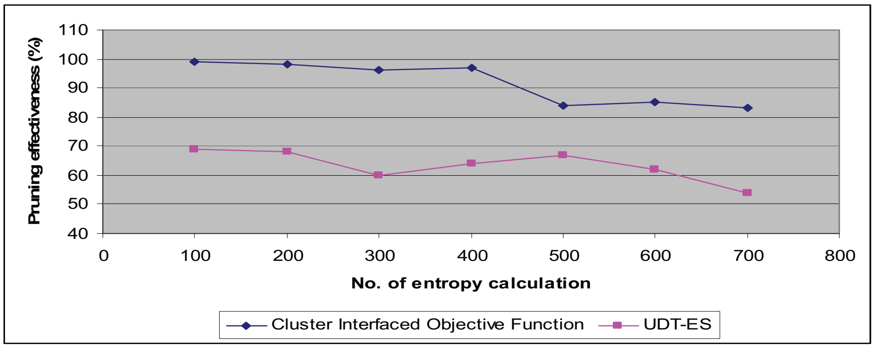


Fig 5.7 Pruning effectiveness

This section deals with the pruning effectiveness of the Cluster Interfaced Objective Function. Fig. 5.7 depicts the number of entropy calculations performed by the Cluster Interfaced Objective Function and UDT-ES. As already explained, the computation time of the lower bound of an interval is comparable to that of the computing entropy. Therefore, for the Cluster Interfaced Objective Function, the number of entropy calculations includes the number of the lower bounds computed.

Figure 5.1 illustrates that the pruning techniques introduced are highly effective. Comparing the techniques in resultant graph against that of Cluster Interfaced Objective Function, it is clear that a lot of entropy calculations are avoided by the bounding techniques. By pruning end points, Cluster Interfaced Objective Function minimizes the number of entropy calculations and increases pruning efficiency. Thus it achieves a pruning effectiveness ranging from 83 % up to as much as 99 %. As the entropy calculations control the execution time of the Cluster Interfaced Objective Function, such effective pruning techniques significantly reduce the tree construction time.

Table 5.6 Execution time

Number of Samples Per PDF	**Execution time (seconds)**	
	UDT-ES	**Cluster Interfaced Objective Function**
50	15	25
100	18	34
150	25	37
200	37	67
250	45	56
300	52	61
350	51	82

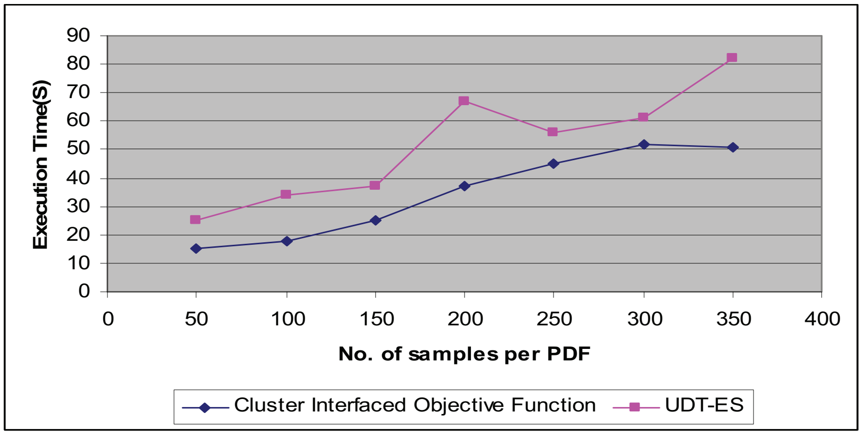


Fig 5.8 Execution time

The execution time of the Cluster Interfaced Objective Function is presented here. Cluster Interfaced Objective Function builds different decision trees by distance boundary clustering technique and objective function. It may be noted that in the experiment, each PDF is represented by 100 sample points. For the data set heart disease diagnosis and prediction, the pruning techniques are so effective that the execution time of the Cluster Interfaced Objective Function is <1.7 times that of UDT-ES, while much better classification accuracy is achieved too.

In this thesis, pruning of the decision tree classifier algorithm has been improved by means of the clustering process with the distance boundaries and the partitioning of uncertain probability distribution values. Clustering is achieved by distance boundary clustering technique, based on the criteria of the lower and the upper bounds distances of the uncertain attributes values. Partitioning and estimating the discrete value of uncertain data is done by the objective function. Relative entropy measure is made on the lower and upper bounded distances on the attribute characteristics related to other certainty attributes in the data set. Experimental results carried out with the metrics such as pruning effectiveness, the number of entropy calculations and the execution time show much better classification accuracy.

5.4 COMPARATIVE ANALYSIS USING ENTROPY

Documents have provided some improvements for the traditional maximum entropy segmentation algorithm which has a range of shortcomings , such as the low accuracy calculation and poor segmentation results. However, the two dimensional (2-D) maximum entropy algorithm assumes practically the whole of the background region. The object region occupies most regions of the two dimensional histogram, and it ignores the impact of the boundary region information on the segmentation results. Hence, in many situations, the segmentation effect is not good. On account of this crisis, the computational domain is re-divided with the method of the two-dimensional maximum entropy algorithm and the image pixels in the segmented images are re-classified with the method of minimum fuzzy entropy, and then its advantage is analysed by means of experimental comparison.

Fuzzy c-means algorithm is one of the most popular algorithms for fuzzy clustering. It could yield compact clusters but might not be able to generate the distinct clusters. On the other hand, entropy-based algorithm could obtain the distinct clusters, which might not be compact. However, the clusters need to be both distinct as well as compact.

Partition coefficient and partition entropy are a class of validation functions that use only the membership function to evaluate the partitioning of the clusters. The following are the disadvantages:

1. It does not take into account the geometrical property of the data.

2. Each depends monotonously on the number of clusters.

3. It decreases monotonously when the number of clusters is very large.

5.4.1 Information entropy

The traditional FCM (Fuzzy c-means) clustering algorithm is very sensitive to the initial center values. Requirements on the data set are too high, and cannot handle noisy data. So, it has already been proposed by the method of using information entropy to initialize cluster centers and introduce the weighing parameter to adjust the location of cluster centers and noise crisis in order to reduce the algorithm's dependence on the initial cluster centers and data sets. The algorithm which uses the information entropy to determine the number of cluster centers. This reduces the algorithm’s dependence on the initial cluster centers and improves the clustering performance greatly. In the same clustering process, the algorithm is simultaneously combined with the ideas of the merger. Then, the weighing parameter is introduced into the fuzzy clustering, which finds the clustering center position, closer to the actual position. In the mean time, the introduction of weighing factor makes the algorithm deal with certain noisy data problems, which broadens the scope of application of the algorithm and reduces the error.

As the FCM algorithm is very sensitive to the number of cluster centers, initialization of cluster centers often artificially gets significant errors, and even gets the actual opposite results. FCM algorithm is hard on data sets too, so the data sets must be quite regular. In order to solve these problems, first of all, information entropy is used to initialize the cluster centers to determine the number of cluster centers. It can reduce some errors, and also can improve the algorithm's efficiency. In order to handle noisy data and adjust the clustering center position, the new algorithm introduces weighing parameters. Then various small clusters are merged so that one can solve the irregular data set clustering. Thus information entropy is adopted to initialize the cluster centers, and introduce the weighing parameter to adjust the position of clustering centers and to handle the noisy data. Then the data sets are divided. Finally, the divided datasets are combined according to certain rules. Experiments prove that the improved algorithm has a strong advantage, since it is able to identify the clusters of arbitrary shape and handle noisy data to some extent. It also can reduce the algorithm's dependence on the initial cluster centers. Simultaneously, it improves the algorithm's efficiency.

5.4.2 Tsallis entropy

Tsallis entropy is a generalization of Shannon entropy and can describe physical systems with long range interactions, long time memories and fractal-type structures. A novel threshold selection technique for image segmentation has been proposed by combining Tsallis entropy and fuzzy c-partition. The image to be segmented is first transformed into fuzzy (soft) domain using membership function. Then, the fuzzy Tsallis entropies for the object and the background are defined. The threshold is selected by finding a proper parameter combination of membership function such that the total fuzzy Tsallis entropy is maximized. To reduce the computational complexity, particle swarm optimization (PSO) is used to search for the optimal parameter combination. The main advantage of the proposed method is that it considers not only the information of the object and the background but also interactions between them in the threshold selection procedure. Experimental results show that the proposed method can give better segmentation performance than methods based on traditional Shannon entropy.

Image analysis usually refers to the processing of the images with the goal of finding objects presented in the image. Image segmentation is a crucial step in image analysis. Thresholding is a widely employed technique for image segmentation. The approaches are based on the assumption that the object and the background can be distinguished by their intensity values. Many techniques for automatically selecting the threshold have been developed over the past years. Among these methods, Shannon entropy based methods and Kapur's and Reny’s entropy methods were intensively studied and proven to be effective. The ambiguity or uncertainty in the image segmentation process, the fuzzy entropy methods, which integrate fuzzy set theory and Shannon entropy, attracted the attention of many researchers. Some other authors introduced the concept of fuzzy c-partition and defined the fuzzy entropy for each fuzzy partition. First, fuzzy sets for the object and the background were defined using membership function with parameters. An optimal threshold was determined by finding a proper parameter combination such that the total fuzzy entropy of each fuzzy partition was maximized. Then, the same method was extended to two-dimensional cases. Few authors proposed another entropy function using fuzzy c-partition and probability partition. Further, it was designed as a three-level maximum entropy method, and also as a minimum cross-entropy method. Further it was all investigated for the performance of the fuzzy entropy approach when it was applied to the segmentation of infrared objects. The basic characteristic of Shannon entropy is extensive or additive, which implies that if a physical system is decomposed into two statistical independent subsystems *A* and *B,* the Shannon entropy of the system is the sum of the Shannon entropy of the subsystems *A* and *B* . Although Shannon entropy is simple in mathematical operation, it ignores the interactions between subsystems. When Shannon entropy is applied to select a threshold for image segmentation, the interactions between the object and the background is ignored, which may lead to inaccurate segmentation results. Recently, Tsallis entropy was proposed as a generalization of Shannon entropy. Compared to Shannon entropy, Tsallis entropy is non-extensive and non-additive, which is suitable to describe systems with long range interactions, long time memories and fractal type structures. Owing to this property of Tsallis entropy, some authors applied maximum Tsallis entropy to select the threshold for image segmentation and got ideal segmentation results.

5.5 COMPARING THE PROPOSED DHCMST WITH K-MEANS CLUSTERING

- In data mining, *k*-means clustering is a method of cluster analysis which aims at partitioning *n* number of observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.
- DHCMST clustering process is reiterated until the optimum number of clusters or regions are obtained.
- The point set S in E^n has already been given and the hierarchical method starts by constructing a Minimum Spanning Tree (MST). The weight of the edge in the tree is the Euclidean distance between the two end points (vertices/pixels). Given an image, the hierarchical method starts by means of constructing a Minimum Spanning Tree (MST). This MST is named as EMST1**.** Next the average weight Ŵ of the edges in the entire EMST1 and its standard deviation σ are computed; any edge with $W > \hat{W} + \sigma$ or current longest edge is removed from the tree. This leads to a set of disjoint sub trees ST = {*T1, T2 …*}(which is a divisive approach).
- To perform image segmentation and edge detection tasks, many modes include or incorporate both region growing techniques as well as edge detection techniques. First, edge detection techniques are applied to obtain the results in the Difference In Strength (DIS) map. Then region growing techniques are employed to work on the map to obtain further results. The image segmentation approach which is solely based upon multi-resolution edge detection method,

region selection method, and intensity threshold method is applied to detect the white matter structure in the brain. K-means clustering algorithm includes the spatial constraints to account for the local intensity variations in the image and the intensity variations usually occur in the regions of the image. The number of clusters *k* is an input parameter: an inappropriate choice of *k* may yield poor results. When performing k-means, it is important to run diagnostic checks for determining the number of clusters in the data set. Convergence to a local minimum may produce counterintuitive ("wrong") results. This is said to be the k-means approach.

- Similar pixels are identified and clustered into a group. These similar pixels are clustered as a group in k-means.
- Cluster separation (CS) is employed to separate the clusters. After having separated the clusters, the separated clusters are grouped or clustered into groups as in DHCMST.
- The threshold value provided by the user is used to identify similar pixels in k-means. Thus the identified similar pixels are grouped into clusters as in k-means. Finally, the threshold value is used to identify the clusters in k-means.
- In DHCMST, a cluster validation criterion called CS is used to identify the clusters.
- The distance of neighboring pixels which are below the threshold is identified as the similar pixel and these similar pixels are clustered into a group in k-means. Such clusters are said to be k-means.

- In DHCMST, the very same algorithm is a nearest centroid-based clustering algorithm, which creates rather regions or sub trees which are the clusters or regions of the data space.
- This particular algorithm works in two phases. The first phase of the algorithm partitioned the EMST1 into the sub trees such as clusters, regions or segments. The clusters are connected through the points to construct EMST1. The centers of clusters or regions or segments are identified using the eccentricity of the points. These points are a representative point for each subtree *ST*. A point *ci* is assigned to a cluster/segment *i* if *ci* ∊ *Ti*. The group of center points is represented as *C = {c1, c2……ck}*. These center points *c1, c2 ….ck* are connected and again minimum spanning tree EMST2 is constructed. The algorithm produces clusters with both intra-cluster and inter-cluster similarity. The intra-cluster will have the documents within a cluster which are very similar and the inter-cluster will have the documents in other clusters which are very similar. The second phase of the algorithm converts the minimum spanning tree EMST2 which has optimal clusters into a dendrogram. This process which is similar to that of DHCMST.
- Based on the result of k-means clustering process the image is segmented and reconstructed to provide the resultant image. The resultant image is displayed to the user and will be allowed to provide a new threshold. Based on the new threshold, the clustering process and segmentation process will be repeated to

provide new results. This methodology will be repeated until the user gets satisfied. This is the process in k-means.

- Here, as in DHCMST, we use a cluster validation criterion based on the geometric characteristics of the clusters is used, in which only the inter cluster metric is acutely used. The DHCMST algorithm is the nearest centroid-based algorithm, which creates region or sub trees (clusters/regions/segments) of the data space. The algorithm partitions a set *S* of data, of data *D* in data space into *n* regions or clusters.
- As in data mining process, *k*-means clustering is a method of cluster analysis which aims at partitioning *n* number of observations into *k* clusters (in which each observation belongs to the cluster with the nearest mean).
- With k-means clustering, the gray scale image is focused on predominently to be converted into the histogram image. The k-means clustering inverse radon transform technique with the threshold is set to increase the efficiency of image segmentation. This threshold is set with the means of user assistance. In the first phase, the method reads the input image and obtains the gray scale image successfully. The obtained gray scale image is used to remove the background objects. Thus a histogram like image will be obtained. The resultant picture will be the input to construct the pixel adjacency graph, which is a construct graph, the graph with a set of pixels. A construct 8-neighbor pixel adjacency graph and edges will be very promptly assigned to the neighboring pixels. Weight map will be calculated based on the similarity measure of the neighboring pixels and

according to the weight values. These sets of processes will be executed repeatedly with user interaction and the user interaction provides the threshold value. The threshold value is the limit for the similarity threshold value, which will be used to cluster the similar pixels. The result of the segmentation will be shown to the user and the system will wait for a new threshold value from the user.

- With the DHCMST, cluster separation alone is chiefly focused on as for the required clusters. *Cluster separation (CS)* is defined as the ratio between minimum and maximum edge of MST ie., *CS = Emin / Emax (2)* where Emax is the maximum length edge of MST, which represents two centroids that are at maximum separation, and Emin is the minimum length edge in the MST, which represents two centroids that are nearest to each other. Then, the CS represents the relative separation of the centroids. The value of *CS* ranges from 0 to 1. A low value of *CS* means that the two centroids are too close to each other and the corresponding partition is not valid at all. A high CS value means the partition of the data is even and valid enough. In practice, a threshold to test the CS is predefined. If the *CS* is greater than the threshold, the partition of the dataset is valid. This process continues until the *CS* is smaller than the threshold. At that point, the proper number of clusters will be the number of clusters minus one. The *CS* criterion finds the proper binary relationship among clusters in the data space. The value setting of the threshold for the *CS* will be practical and is dependent on the dataset. The higher the value of the threshold the smaller the number of clusters would be. Generally, the value of the threshold will be > 0.8. Fig 5.3 shows the CS value versus the number of clusters in hierarchical clustering. The CS value is < 0.8 when the number of clusters is 4. Thus, the proper number of clusters for the data set is 3. Furthermore, the computational cost of *CS* is much lighter because the number of sub clusters is small. This makes the *CS* criterion practical for the DHCMST algorithm when it is used for clustering/segmenting large datasets (image) and to detect the outliers too.

- A key limitation of *k*-means is its cluster model. The concept is usually based on the spherical clusters that are separable in a way so that the mean value converges towards the cluster center. The clusters are expected to be of similar size, so that the assignment to the nearest cluster center is the correct assignment. For example by applying *k*-means with a value of k=3 onto the well-known iris flower data set, the result often fails to separate the three Iris species contained in the data set. With k=2, the two visible clusters are discovered, whereas with k=3 one of the two clusters will be split into two even parts. In fact, k=2 is more appropriate for this data set, despite the data set containing 3 classes. As with any other clustering algorithm, the *k*-means result certainly relies on the data set to satisfy the assumptions made by the clustering algorithms. It works very well on some data sets, while failing on others.
- DHCMST is not at all a limited one. It does not require a predefined cluster number.

5.6 APPLICATION OF CLUSTERING IN MEDICAL IMAGE SEGMENTATION

As a part of the work a segmentation of medical image is carried out with the clustering technique and the output is displayed. A combination of Graph cut technique and k-means clustering with multi threshold for medical image segmentation is presented. The main objective of medical image segmentation is to extract the anatomic structures and its characteristics with respect to some input features. There exist various methodologies for medical image segmentation but struggles with missing features due to the noise presence in the medical images. In propose a new technique to increase the resolution of the medical images to identify the features and edges of the medical images. The medical image is preprocessed to reduce the noise, and then multi-level histogram is generated in the first phase of the process. In the second stage, with the initial segmentation obtained with gray level contours, and generate the histogram then the construct the pixel adjacency graph in which each nodes represents the set of pixels of the image and edges links the neighbor pixels. A calculate the similarity of neighboring pixels; based on the distance value, cluster the similar pixels to a class. In use k-means clustering to group similar pixels and used Euclidean distance measure to calculate the similarity between pixels. The proposed method is carried out iteratively until the user gets satisfied. The method will be carried out repeatedly with the user defined threshold value. The threshold is used to group pixels with in the distance. With the proposed technique the features are maintained and the resolution of the image enhanced. The time complexity of the process is reduced. In propose a new methodology for the segmentation of medical images. Our method uses combination of graph based segmentation technique and clustering inverse radon transform technique with user assistance to set threshold to increase the efficiency of image segmentation. In the first phase our method reads the input image and obtains he gray scale image. The obtained gray scale image is used to remove the background objects then the histogram of the image will be obtained. The resultant picture will be the input to construct the pixel adjacency graph, a construct the graph with set of pixels. A construct 8-neighbor pixel adjacency graph and edges will be assigned to the neighboring pixels. Weight map will be calculated based on the similarity measure of the neighboring pixels and according to the weight values to apply k-means clustering technique to cluster the similar pixels. These set of processes will be executed repeatedly with the user interaction, user interaction is assisted to provide the threshold value. The threshold value is the limit for the similarity threshold value, which will be used to cluster the similar pixels. The result of segmentation will be shown to the user and the system will wait for new threshold value from the user.

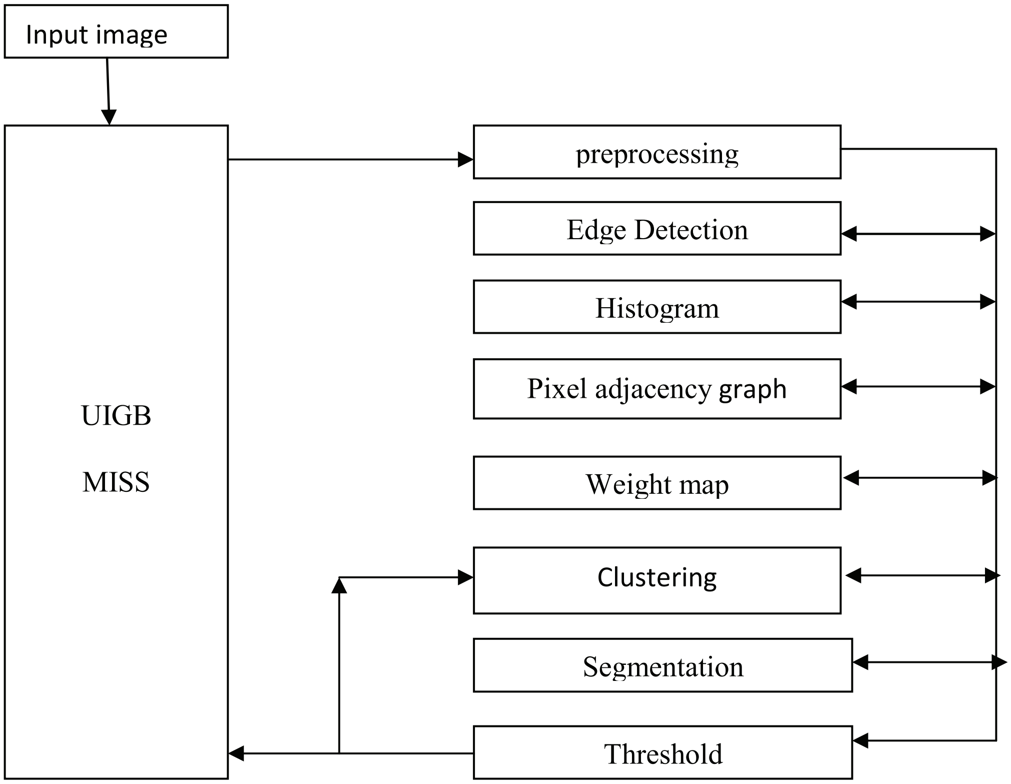


Fig. 5.9 **Overall block diagram of the clustering in Medical Image Segmentation**

5.6.1 Preprocessing

At the preprocessing stage to generate the gray scale values of the original image. The generated gray scale values are then passed to the edge detection and used sobel edge detector for the purpose of edge detection. The output image of the edge detection process is used to increase the intensity values of the original image. An increase the intensity of the detected edges in the original image.

5.6.2 Histogram Generation

Here we generate the histogram value of the image, to used 64 bit histogram for our purpose. The resultant value is used for the segmentation process.

5.6.3 PAG Construction

The pixel adjacency graph is constructed with the set of pixels from the image. The construct 8-neighbor pixel adjacency graph and edges will be assigned to the neighboring pixels. Set of pixels are assigned as nodes and links are assigned with the neighbor pixels of the image. The maintained separate graph for each set of pixels.

5.6.4 Weight map construction

An edge weight is computed with the grey level contrast of the pixel. Given a pixel adjacency graph G = (V;E;W), consider the following edges weights mapping:

$$n \in \mathbb{N}, \forall e_{i,j} \in E, \;\; w_{i,j}^{n} = \left(\frac{|p_i - p_j|}{d(i,j)} + 1 \right)^{n}$$

where i and j are two nodes of the graph, pi and pj are the grey level values of neighbor pixels of the image and d(i; j) is the Euclidean distance between the two neighbor pixels. wi;j plays the role of a dissimilarity measure between neighbors pixels and can be seen as a local estimate of the image's gradient modulus. wi;j is a strictly positive increasing function of $|pi\text{-}p_j|$. A small edge weight means that pixels i and j have similar values, whereas wi;j takes large values when pixels i and j have significant different values.

5.6.5 Clustering

The similar pixels are identified and clustered to a group. The threshold value provided by the user is used to identify the similar pixels. The distance of neighboring pixels which are below the threshold is identified as similar pixel and will be assigned to a class.

5.6.6 Segmentation

Based on the result of k-means clustering process the image is segmented and reconstructed to provide the resultant image. The resultant image is displayed to the user and will be allowed to provide new threshold. Based on new threshold entered the clustering process and segmentation process will be repeated to provide new result. This methodology will be repeated until the user gets satisfied.

5.6.7 Resulting images

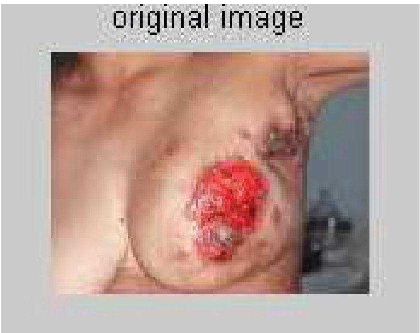


Fig. 5.10 Input Image for Segmentation

The following Fig displays the result of detected edges for the input image.

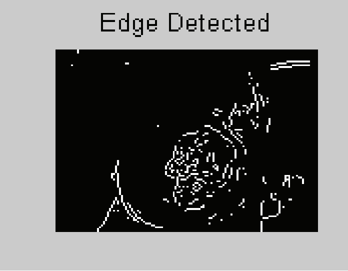


Fig. 5.11 Edge detected Image

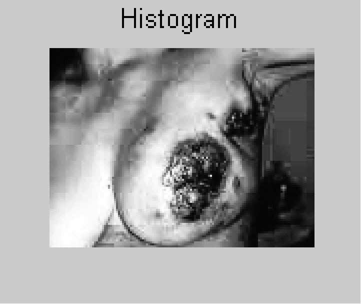


Fig. 5.12 Histogram image

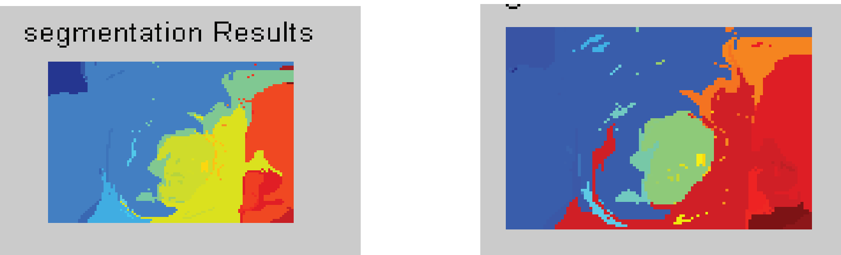


Fig. 5.13 & 5.14 Segmented images

5.7 CONCLUSION

The MSTODDN algorithm finds the outliers without using any predefined input parameter. The algorithm does not require the users to choose and try variant parameter combinations in order to get the desired output. A further study and research toward the rich properties of the EMST-based methods to solve the different problems in the detection of the outliers in the dynamic data set is necessary. The experimental results carried out with the metrics such as pruning effectiveness, number of entropy calculations and executive time will all achieve much better classification accuracy. The proposed method uses the graph based on user assisted image segmentation and inverse radon transformation to improve the quality of the image. The proposed method can be used in variant fields of vision technology. The problem of clustering the data item by using the MST graph model has been considered. Since the approach has been based upon MST, the running time of the algorithm is much reduced. DHCMST clustering algorithm uses a new cluster validation criterion and based upon the geometric property of the partitioned regions or clusters an optimal number of true clusters with center for each of them is produced. In this thesis, the experimental result of a synthetic data set namely students' semester marks has been presented. The experimental result shows that the proposed algorithm performs better than the k-means algorithm. The results also prove that the DHCMST algorithm gives good clusters. This algorithm is programmed in C language and uses both the divisive and agglomerative approach to find out the metaclusters.

CHAPTER VI

CONCLUSION AND FUTURE ENHANCEMENT

The MSTODDN algorithm finds the outliers without using any predefined input parameter. The outliers should at any cost detect the virus in the system, otherwise these outliers to the image segmentation can be harmed. But these outliers can be erroneous and these erroneous outliers are used to misreport the errors. The MSTODDN algorithm detects these outliers. From a theoretical point of view all these research solutions look very good. Practically an improvement of the running time of the algorithm should be made and can be made also. In course of time, the running time will be very much improved. Also the algorithm does not require the users to choose and try variant parameter combinations in order to get the desired output. The researcher intends to do further study and research toward the rich properties of the EMST-based methods to solve different or variant problems for the detection of the outliers in the dynamic data set.

The proposed segmentation mode uses the graph based on image segmentation and inverse radon transformation to improve the quality of the image and the proposed method can be used in various fields of vision technology. In the k-means clustering method, similar pixels are identified as clusters and are grouped as clusters. The proposed mode can be extended to adapt with the variant filters at the last stage of the process after the inverse radon transformation. The proposed algorithm will be further researched and reviewed for the sake of medical image segmentation. The proposed segmentation methodology uses the graph based user assisted image segmentation and inverse radon transformation to improve the quality of the image. The proposed methodology will be further extended to adapt various filters at the last stage of the process after the inverse radon transformation and the algorithm proposed in this thesis may be further reviewed for medical image segmentation.

The problem of clustering the data item is processed by using the MST graph model. The proposed mode promotes good preliminary results to promote the study further. Since the present approach has been based upon MST, the running time of the algorithm is much reduced. The DHCMST clustering algorithm uses a new cluster validation criterion and is based upon the geometric property of the partitioned regions or clusters in order to produce an optimal number of true clusters with a center for each of them. In this thesis, the experimental result on some synthetic data sets, namely students' semester marks, has been presented. The experimental result shows that the proposed algorithm performs better than the k-means algorithm. The results also prove that the DHCMST algorithm gives good clusters. The algorithm is programmed in C language and uses both the divisive and agglomerative approaches to find the metaclusters. In future, experiments may be conducted on computing time and the method may be tried on the real world data obtained from the user's access on a web file in order to embed it all with the web page recommendation model.

In this thesis, the Cluster Interfaced Objective Function is presented for the Decision Tree Classifiers for Mining Uncertainty data. Pruning of decision tree classifier algorithm has been improved by clustering with distance boundaries and partitioning of the uncertain probability distribution values. Clustering is achieved by the distance boundary clustering technique, based on the criteria of lower and upper bound distances of the uncertain attribute values. Partitioning and estimating the discrete value of the uncertain data is done by the objective function. Relative entropy measure is made on the lower and upper bounded distances on the attribute characteristics related to the other certainty attributes in the data set. Experimental results obtained from the metrics such as pruning effectiveness, the number of entropy calculations and execution time has shown much better classification accuracy.

The proposed segmentation methodology uses the graph based user assisted image segmentation and inverse radon transformation to improve the quality of the image. The proposed methodology can be used in various fields of vision technology. The proposed methodology can be further extended to adapt various filters at the last stage of the process after the inverse radon transformation. The algorithm proposed in this paper can be further reviewed for medical image segmentation process.

In future work, the performance analysis may be based on some newer optimization methods and also the algorithm comparisons may be prolonged to a wide range of applications. Furthermore, this application may be used to categorize the medical images in order to diagnose the right disease verified earlier.

REFERENCES

[1] Adiga, P.S. & Chandhuri, B.B. An efficient method based on watershed and rule based imaging for segmentation of 3D Histo- pathological images pattern recognition, Vol.34, No.7, pp 1449-1458, 2001.

[2] An open source java content based image retrieval library, http:// www.semantic metadata, March 2010.

[3] Archip, N., Erard, P.J., Egmount Petersen, M., Haeflinger, J.M. & Germond, J.K. Knowledge-based approach: its automatic detection of the spinal card in CT images, IEEE Transaction on medical Imaging, Vol.21, No.12, pp 1504-1516, 2002.

[4] Asano,T., Bhattacharya, B., Keil, M. & Yao, F. Clustering algorithms based on minimum and maximum spanning trees, *Proceedings of the 4th Annual Symposium on Computational Geometry*, pp 252–257, 1988.

[5] Bach, J.R., Fuller, C., Gupta, A., Hamppur, A., Horowitz, B., Humphrey, R. , Jain, R. & Sha, C.F. The virage image search engine: An open frame work for image managementation, Proceedings SPIE storage and Retrieval for still image and video Data bases lV, Vol.2670, No.1, pp 76-87, 1996.

[6] Bansal, N., Blum, A., & Chawla, S. Correlation clustering. In *Proceedings of the 43rd FOCS*, pp 238–247, 2002.

[7] Beucher, S. & Lantuejoul, C. Use of watersheds in contour detection. In International Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation, Rennes, France, September 1979.

[8] Bezdek, J., Hall, L. & Clarke, L. Review of MR image segmentation techniques using pattern recognition. Med. Physics, Vol.20, No.4, pp 1033-1048, 1993.

[9] Carson, C. & Belongie, B. Color-and texture-based image segmentation using EM and its application to content-based image retrieval. ICCV'98, pp 675-682, 1998.

[10] Chang, N.S., & Fu, K.S. Picture Query Languages for Pictorial data base systems, Computer., Vol.25, No.11, pp 13-21, 1981.

[11] Chang, N.S. & Fu, K.S. Query-by-Pictorial-example, IEEE Transactions on Software Engineering, No.6, pp 519-924, November 1980.

[12] Charikar, M., Guruswami, V. & Wirth, A. Clustering with qualitative information, *Journal of Computer and System Sciences*, Vol.71, pp 360–383, 2005.

[13] Cheng, H.D., Jiang, X.H., Sun, Y & Wang, J. Color image segmentation: advances and prospects, Elsevier: Pattern Recognition, Vol.34, No.12, pp 2259-2281, 2001.

[14] Cheriet, M. , Said, J.N. & Suen, C.Y. A recursive thresholding technique for image segmentation, IEEE Transaction on Image Processing, Vol.7, No.6, pp 918-920, 1998.

[15] Christ, M.C.J. & Parvathi, R.M.S. Medical Image Segmentation using Fuzzy C-means clustering and Marker Controlled Watershed algorithm, International Journal of Modern Engineering Research, Vol.2, No.1, pp 408-411, 2012

[16] Christoph H. Lampert, Hannes Nickisch, Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer, IEEE Xplore : Computer Vision and Pattern Recognition, pp 951-958, 2009.

[17] Comanicia, D. & Meer, P. Meanshift : a robust approach toward feature space analysis, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.24, No.5, pp 603-619, 2002.

[18] David G. Lowe. Distinctive image features from scale-invariant key points, International Journal of Computer Vision, Vol.60, No.2, pp 91-110, 2004.

[19] Davis, L.S. A survey of edge detection techniques, Elsevier: Computer Graphics and Image Processing, Vol.4, No.3, pp 248-270, 1975.

[20] De Smet, P. Pires, V.P.M. Implementation and analysis of optimized rain falling watershed algorithm, Journal of Electronic imaging, Vol.8, No.3, pp 270-278, 1999.

[21] Demaine, E. & Immorlica, N. Correlation clustering with partial information. In *Proceedings of the 6th RANDOM-APPROX*, pp 1–13, 2003.

[22] Dowe, J. Content-based retrieval in multimedia Imaging, Proc. SPIE conference on storage and retrieval for image and video data bases, Vol.1908,January 1993.

[23] Eldershaw,C & M. Hegland. Cluster analysis using triangulation, Computational Techniques and Applications: CTAC97, World Scientific, pp 201–208, 1997.

[24] Fei- Fei, L, & Perona, P. A Bayesian hierarchical model for learning natural scene categories, IEEE Explore:Computer Vision and Pattern Recognition, Vol.2, pp 524-531, 2005.

[25] Flusser, J. & Suk, T. A moment based approach to registration of images with affine geometric distortion, IEEE Transaction: Geoscience and Remote Sensing, Vol.32, No.2, pp 382-387, 1994.

[26] Fukunaga, K. & Hostetler, L.D. The estimation of the gradient of a density function, with applications in pattern recognition, IEEE Transaction on information Theory, Vol. 21, No. 1, pp 32-40,1975.

[27] Girish Kulkarni, Premraj,V., Dhar,S., Siming Li, Yejin chei, Alexander, C., Berj, & Tamara L. Bers, Baby talk: Understanding and generating simple image descriptions, IEEE Transaction on Computer Vision and Pattern Recognition, Vol.10, pp 1601-1608, 2011.

[28] Gower, J. & Ross, G. Minimum spanning trees and single linkage cluster analysis, *Applied Statistics*, Vol.18 pp 54–64, 1969.

[29] Gross, A. & Latecki, L. Digital geometric invariance and shape representation, IEEE Transaction on Compute Vision, Vol.62, No.3, pp.121-126, 1995.

[30] Haker, S., Sapiro, G. & Tannenbaum, A. Knowledge-Based segmentation of SAR Data with learned priors, IEEE Transation on Image Processing Vol.9, No.2, pp 299-301, 2000.

[31] Haralick, R.M. & Shapiro, L,G. Image segmentation Techniques, Elsevier: Computer vision, Graphics and image Processing, Vol.29, No.1, pp 100-132, 1985.

[32] Haris,K., Efstratiadis, S.N., Maglaveras, N. & Katsaggelos, A.K. Hybrid image segmentation using watersheds and fast region merging, IEEE Transaction on Image Processing. Vol.7, No.12, pp 1684-1699, 1999.

[33] Herold, J., Schubert, W. & Nattakemper, T.W. Automated detection and quantification of fluorescently labeled synapses in murine brain tissue sections for high throughput applications, Elsevier:Journal of Biotechnology, Vol.149, No.4, pp 299-309, 2010.

[34] Hu.F,Z., Ehrlich, G. & Hiller, N.L. What makes pathogens pathogenic, Genome Biology, Vol.9, No.6,pp.225-231, 2008.

[35] Jain, A.K., Murty, M.N. & Flynn, P.J. Data clustering: a review. ACM Computing surveys, Vol.31, No.3, pp 264-323, 1999.

[36] Jenkinson, M. & Smith, S. A global Optimizations method for robust affine registration of brain images, Elsevier:Medical Image Analysis, Vol.5, No.2, pp 143-156, 2001.

[37] Jiawei Han & Michieline Kamber. Data mining for image video processing: a promising research frontier, Proceedings of the international conference on content-based image and video retrieval, ACM, 2008

[38] Jirimatas, Ondrej Chum, Martin Urban, Toms Pajdla. Robust wide-base line stereo from maximally stable extremal regions. Elsevier: British Mission Vision Computing, Vol.22, No.10, pp 761-767, 2002.

[39] Johnson, D,S. The NP-completeness column: an ongoing guide, *Journal of Algorithms*, Vol.3, No.4, pp 381–395, 1982.

[40] Joset sivic & Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos, Proceedings of IEEE International conference on Computer Vision, Vol.2, pp 1470-1477, 2003.

[41] Kapur,D., Lakshman, Y.N., & Saxena,T. Computing invariants using elimination methods, Proceedings of IEEE International Symposium on Computer Vision, pp 97-102, 1995.

[42] Kass, M., Witkin, A. & Terzopoulos, D. Snakes: Active contour models, International Journal of Computer Vision, Vol.1, No.4, pp 321-331, 1988.

[43] Kato, T. Data base architecture for content-based image retrieval, Proceedings of SPIE, Vol.1662, pp 112-123, 1992.

[44] Kempf, G,R. Computing invariants, Springer, Vol.1278, pp 81-94, 1987.

[45] Li- Jia, Chong wang, Yongwhan Lim, David Blei, Fei-Fei Li. Building and using a semantivisual image hierarchy, IEEE transaction on Computer Vision and Pattern Recognition, pp 3336-3343, 2010.

[46] Lucchese, L. & Mitra,S.K. color image segmentation : A state of the art-survey Processing of the Indian National science Academy, Vol.67, No.2, pp 207-221,2001.

[47] Luxburg, U. A tutorial on Spectral Clustering, Springer, Vol.17, No.4, pp395-416, 2007

[48] Ma, W.Y. & Manjunath, B.S. A Comparison of wavelet transform features for texture image annotation, Proc. Second International conference on Image Processing (ILIP' 95), Vol.2, pp 256-259, 1995.

[49] Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., & Suetens,P. Multimodality image registration by maximization of mutual information, IEEE Transaction on medical Imaging, Vol.16, No.2, pp 187-198, 1997.

[50] Makrogiannis, S., Economou, G., Fotopoulos, S. & Bourbakis, N.G. Segmentation of color images using multiscale clustering and graph theoretic region synthesis. IEEE Transaction on System, Man, and Cybernetics- Part A: Systems and Humans, Vol.35, No. 2, pp 224-238, 2005.

[51] Merz, C.J. & Murphy, P.M. UCI repository of machine learning databases, 1996

[52] Mohammad Yaqub, M. Kassim Javaid, Cyrus Cooper, Alison Noble, J. Improving the classification accuracy of the classic RF method by intelligent feature selection and weighted voting of trees with application to medical image segmentation, MLMI11, Proceedings of the second international conference on machine learning in medical imaging, Springer, Vol.7009, pp184-192, 2011.

[53] Myron Flickner, Harpeet Sawhney, wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, David Steele, Peter Yanker. Query by Image and Video Content:the QBIC system, IEEE Transaction on Computer, Vol.28, No.9, pp 23-32, September 1995.

[54] Olabarriagar, S.D. & Smeulders, A.M. Interaction in the segmentation of medical images: A survey, Elsevier: Medical Image Analysis, Vol.2, No.5, pp 127-142, 2001.

[55] Pal, N.R. & Pal, S.K. A Review on Image Segmentation Techniques, Pattern Recognition, Vol.26, No.9, pp 1277-1294, 1993.

[56] Palmer, S. Vision Science : Photons to Phenomenology, MIT Press, 1999

[57] Papamorkos, N. & Gatos, B. A new approach for multi-level Threshold Selection, Computer vision, Graphics and Image Processing, Vol.56, No.5, pp 357-370, 1994.

[58] Pappas, T.N. An adaptive clustering algorithm for image segmentation, IEEE Transaction on Signal Processing, Vol. 40, No.4, pp 901-914, 1992.

[59] Paivinen, N. Clustering with a minimum spanning tree of scale-free-like structure, Pattern Recognition Letters, Vol.26, No.7, pp 921–930, 2005.

[60] Pentland, A., Picard, R.W. & Sclaroff, S. Photo book: Content- based manipulation of image data bases, International Journal of Computer Vision, Vol.18, No.3, June 1996, pp 233-254,1996.

[61] Rajendran, P, & Madheswaran, M. Hybrid Medical Image Classification Using Association Rule Mining With Decision Tree Algorithm, Journal Of Computing, Vol.2, No.1, pp 1-10, 2010.

[62] Reed, T. & Du Buf, J. A review of recent texture segmentation and feature extraction techniques, Elsivier: CVGIP Image understanding, Vol.57, No.3, pp 359-372, 1993.

[63] Retrieval Ware, demo page. http.//en.Wikipedia.org/wiki/ Retrieval ware, March 2010.

[64] Rui, Y., She, A.C. & Huang, T.S. Modified Fourier Descriptors for shape Representation - - a practical approach, Proceedings of the First International workshop on Image Databases and Multimedia search, pp 22-30, August 1996.

[65] Sahoo, P.K., Soltani, S. & Wong, A.K.C. A survey of thresholding techniques, Elsevier:ComputerVision, Graphics and Image Processing, Vol.41, No.2, pp 233-260, 1988.

[66] Shotton, J., Johnson, M. & Cipolla, R. Semantic texton forests for image categorization and segmentation, IEEE conference on Computer Vision and Pattern Recognition, pp 1-8, 2008.

[67] Smith, J.R. & Chang, S.F. Visually Searching the web for content, IEEE Multimedia magazine 4(3), 1997. Columbia, U CU/CTR Technical Report 459-96-25, pp 1220.

[68] Solihin, Y. & Leedham, C.G. Integral ratio: A new class of Global Thresholding Techniques for Handwriting Images, IEEE Transaction on Pattern Analysis and Mechine Intelligence, Vol.21, No.8, pp 761-768, 1999.

[69] Thomas Deselears & Vittorio Ferrars. Visual and semantic similarity in image net, Computer Vision Laboratory, pp 1777-1784, 2011.

[70] Toga, A. Brain Warping, Academic Press, Digital image Processing, wiley-Interscience, 2007

[71] Turi, R.H. Clustering-based colour image segmentation, Ph.D thesis, Monash University, Australia, 2001.

[72] Udupa, J. & Samarasekara, S. Fuzzy Connectedness and Object Definition: Theory, Algorithms and Applications in Image Segmentation, Elsivier:Graphical models and image processing, Vol.58, No.3, pp 246-261, 1996.

[73] Viola, P. & Wells III, W,M. Alignment by maximization of mutual information, International Journal of Computer Vision, Vol.24, No.2, pp 137-154, 1997.

[74] Wang, J.Y.A. & Adelson, E,H. Representing moving images with layers, IEEE Transaction on image processing, Vol.3, No.5, pp 625-638, 1994.

[75] Wells III, W.M., Viola, P., Atsumi, H., Nakajima, S & Kikinis, R. Mulli-Model volume registration by maximization of mutual information, Medical Image Analysis, Vol.1, No.1, pp 35-51, 1996.

[76] Xiagogang Wang, & Eric Grimson. Spatial Latent dirichlet Alocation, In Proceedings of Neural Information Processing Systems Conference(NIPS), pp 1-8, 2007.

[77] Xiaoyu Wang, Tony, X. Han & Shuicheng Yan. An HOG-LBP human detector with partial oculusion handling, IEEE Conference on Computer Vision, pp 32-39, 2009.

[78] Xu, J., Monaco,J.P. & Madabushi, A. Markov random field driven region-based active contour model (MaRACel): application to medical image segmentation, Medical Image Computing and Computer-Assisted Intervention, Vol.13,No.3 3, pp 197-204, 2010.

[79] Xu, Y., Olman, V. & Xu, D. Minimum spanning trees for gene expression data clustering. Genome Informatics, Vol.12, pp 24–33, 2001.

[80] Zadeh, L.A. Fuzzy Sets. Information and control, Vol.8, pp 338-353, 1965.

[81] Zahn, C. Graph-theoretical methods for detecting and describing gestalt clusters, IEEE Transactions on Computers, Vol.C-20, No.1, pp 68–86, 1971.

[82] Zhang Jian-Minga1, Ren Xin-Bo1, Zhang Fan-Biao2, Zhang Xiao-Li1. Approach for automated segmentation and detection of dendrites and spines from fluorescence confocal image, 2012.

[83] Zucker, S.W. Region growing: Childhood and adolescence computer graph and its image processing. Vol.5, pp 382-399, 1976.

PUBLICATIONS

[1] Karthigeyan, T & Chidambaranathan,S. Graph based User Interacted Medical Image segmentation with multi threshold using K-means clustering, Advances in Computational Sciences and Technology (ACST), Vol.2, No.1, PP 25-31, 2013.

[2] Karthigeyan, T & Chidambaranathan, S. A CCBIR System using associative mining techniques for semantic search engines, International Journal of Computer Sciences and Applications, Vol.1, No.10, PP 19-24, 2013.

[3] Karthigeyan, T., John peter, S & Chidambaranathan, S. An Efficient Detection of Outliers and Hubs Using Minimum Spanning Tree, Vol.3, Issue.7, PP 91-96, 2011.

[4] Karthigeyan, T., John peter, S & Chidambaranathan, S. Meta clusters through minimum spanning tree based clustering for performance analysis of students, Journal of Discrete Mathematical Sciences & Cryptography, Vol.14, No.4, PP 349-367, 2011

Papers presented in National / International Conferences

S.No	Name of the Institution	Title of the Paper	Year
1	PSG College of Arts and Science, Coimbatore	An enhanced model for achieving high throughput using variant buffer size methodology	National conference on Web Science and Tech, 10th OCT 2012.
2	Einsteen College of Engg., Tirunelveli	An efficient data cleaning through minimum spanning tree for data mining	National conference on Crypt Analysis Techniques in Computer Hacking, 18th & 19th Feb. 2011
3	Aarupadai veedu Institute of tech., Chennai	Discrete Wavelet Transformation in dual Water Marking	National conference on VLSI, Embedded systems, Signal processing and communication,110-116, April 2010.
4	Noorul Islam University, Thucklay	Adaptive wavelet thresholding for Image Denoising in digital Mammographic Images	International Conference on ETES 26th and 27th March 2010; 56-62.
5	Sun Engg. College, Nagercoil	Effect of Median Filter in Interest point detection from noised image	International Conference on Intelligent science and tech, 18th and 19th Feb 2010.
6	Thiagarajar School of Management, Madurai	An Analytical study on tools of Network Security and Firewalls	National Conference of ETBSIS, 12th Feb. 2010.

Seminars/ Conferences/Workshops attended

S.No	Title	Venue	Date	National/International
1	ICT enabled Teaching-Learning through Smart Class Room Facility	St. Xavier's College, Palayamkottai	27the Sep. 2011	National Workshop
2	Recent Trends in Communication Engineering	Sastra University, Kumbakonam	2nd April 2010	National Conference
3	Information and Software Engineering	Aarupadai Veedu Institute of Technology, Vinayaka Missions University,Chennai	26 & 27th Feb.2010	National Conference

4	Computing, Communication and Information system	Sri Krishna College of Engineering and Technology, Coimbatore	12-13, Feb, 2010	National Conference
5	VLSI Design and Communication systems	Meenakshi Sundararajan Engineering College, Chennai	8-10 January 2010	International Conference
6	Wireless Communication and Sensor Computing	SSN College of Engineering, Chennai	02-04 Jan.2010	International Conference
7	Advanced Intelligent system	Dr. Sivanthi Aditanar College of Engineering College, Tiruchendur	15 & 16th Dec. 2009	National Conference
8	Campus to Corporate	Manonmaniam Sundarnar University, Tirunelveli	14&15th Dec.2009	National Workshop
9	Quality Initiatives for student support and progression	St. Xavier's College, Palayankottai	4 & 5th Dec.2009	National Conference
10	IDL Programming	St. Xavier's College, Palayankottai	22 & 23rd February 2008	National Seminar
11	Emerging trends in Algorithms	St. Xavier's College, Palayankottai	2nd & 3rd March 2006	National Conference

Publication of Books

S.No	**Title**	**Authors**	**Year & Page No.**	**Publisher**
1.	A Programmer's Introduction to PHP	T.Karthigeyan & S.Chidambaranathan	2013, 142	Thozalamai Publications, Chennai
2.	Everything in HTML	T.Karthigeyan & S.Chidambaranathan	2012, 109	Einstein Publications, Chennai
3	XML – An Practical approach	S.Chidambaranathan	2011, 170	Thozalamai Publications, Chennai

Achievements/Awards won:

1. **Best Researcher Award(March 2012),** St. Xavier's College, Palayamkottai
2. **Best Researcher Award(March 2012),** St. Xavier's College, Palayamkottai
3. **Best Researcher Award(March 2011),** St. Xavier's College, Palayamkottai
4. **Award of Excellence(March 2010)** in the International Conference at Noorul Islam University, Thucklay
5. **Best paper Award(Feb.2010)** in the International Conference, Sun Engineering College, Nagercoil
6. **Best Performance award(April 2007)** for Students Training and Action for Neighbourhood Development (**STAND**) Program, St. Xavier's College(Autonomous), Palayamkottai.

MST: A CONTEMPORARY APPROACH FOR DATA MINING BASED ON CLUSTERING METHOD

THESIS *submitted to*

MANONMANIAM SUNDARANAR UNIVERSITY

in partial fulfillment of the requirement

for the award of the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

by

S. CHIDAMBARANATHAN

Reg.No. 5997

RESEARCH DEPARTMENT OF COMPUTER SCIENCE

ST. XAVIER'S COLLEGE (AUTONOMOUS)

PALAYAMKOTTAI-627 002

DECEMBER 2013