DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DATA MINING SUPPORTED HOSPITAL INFORMATION SYSTEMS SOLUTIONS

by

Tuba EŞİYOK

September, 2011

İZMİR

DATA MINING SUPPORTED HOSPITAL INFORMATION SYSTEMS SOLUTIONS

A Thesis Submitted to the

Graduate School of Natural and Applied Sciences of Dokuz Eylül University

In Partial Fulfillment of the Requirements for the Degree of Master of Science

in Computer Engineering, Computer Engineering Program

by

Tuba EŞİYOK

September, 2011

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"DATA MINING SUPPORTED HOSPITAL INFORMATION SYSTEMS SOLUTIONS"** completed by **TUBA EŞİYOK** under supervision of **PROF. DR. ALP KUT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Alp KUT

Supervisor

Y.Doç.Dr. Reyat YILMAZ

(Jury Member)

Yrd. Doç. Dr. Canan E. ATAY

(Jury Member)

Prof.Dr. Mustafa SABUNCU
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to thank to my supervisor, Prof. Dr. Alp KUT, for his help, useful suggestions and guidance.

I am also highly thankful to my mother Yüksel EŞİYOK for being near me all time and my father Hasan EŞİYOK for supporting me unconditionally. This thesis would not have been possible without their love and support.

Finally, my brother Mesut Onur EŞİYOK and my sister Gülistan EŞİYOK, you are my family. Thanks for yours care and love.

Tuba EŞİYOK

DATA MINING SUPPORTED HOSPITAL INFORMATION SYSTEMS SOLUTIONS

ABSTRACT

Data mining is the process of obtaining meaningful information from large-scale datasets. It attempts to obtain meaningful results from these data by analyzing the data with the help of a variety of methods and techniques. Data mining is used in many areas such as machine learning, pattern recognition, statistics and medicine in order to analyze and interpret the information.

The purpose of this study is to examine DBSCAN clustering algorithm which is a data mining technique and to allocate appropriate number of clusters by analysing mammography data with a software developed using this algorithm.

In the study, mammography data were divided into optimal number of clusters with DBSCAN algorithm. The noisy data that not included in any cluster were added in the appropriate clusters with the K-NN classification algorithm. Thus, it is aimed to be analysing mammography data's easier.

In a conclusion, parameter values (Eps, MinPts) which give optimal result on MIAS database were determined by examining the values of these parameters. The specifications of clusters which created by this parameter value were defined. For noisy data, k parameter value that produces the best result for the K-NN, was detected by examining all k nearest-neighbors parameter values.

Keywords: data mining, clustering, DBSCAN, mammography, classification, K-NN algorithm

VERİ MADENCİLİĞİ DESTEKLİ HASTANE OTOMASYONU SİSTEMLERİ ÇÖZÜMLERİ

ÖZ

Veri madenciliği, büyük ölçekli veri setlerinden anlamlı bilgiler elde etme işlemidir. Çeşitli yöntemler ve teknikler yardımı ile verilerin analizi yapılarak, bu verilerden anlamlı sonuçlar elde edilmeye çalışılır. Veri madenciliği; bilgilerin analiz ve yorumlanması için makine öğrenmesi, örüntü tanıma, istatistik ve tıp gibi birçok alanda kullanılmaktadır.

Bu çalışmanın amacı, veri madenciliğinde bir kümeleme tekniği olan DBSCAN algoritmasını incelemek ve bu algoritmayı kullanarak geliştirilen bir yazılım aracılığıyla kanserli hücrelere sahip hastalara ait mamografi görüntülerinin analizini yaparak verileri en uygun sayıda kümelere ayırmaktır.

Çalışmada, mamografi verileri DBSCAN algoritması ile optimum sayıda kümeye ayrıldı. Hiçbir kümede yer almayan gürültülü veriler ise K-NN sınıflandırma algoritması ile en uygun kümelere dahil edildi. Böylece, mamografi verilerinin daha kolay analiz edilmesi hedeflendi.

Sonuç olarak, mamografi verilerini kümelere ayıran parametre (Eps, Minpts) değerleri incelenerek en uygun sonucu veren parametre değeri belirlendi. Bu parametre değeri ile oluşturulan kümelerin özellikleri tanımlandı. Gürültülü veriler için ise k-en yakın komşu parametre değerleri incelenerek, hangi parametre değeri için K-NN algoritmasının en iyi sonuç ürettiği tespit edildi.

Anahtar sözcükler: veri madenciliği, kümeleme, DBSCAN, mamografi, sınıflandırma, K-NN algoritması

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

The goal of this study is to achieve meaningful results about analysing mammography images. Clustering and classification algorithms were used together in the project. DBSCAN algorithm for clustering and K-NN algorithm for classification were implemented. This thesis is organized as follows: chapter 2 introduces the major characteristics of the DBSCAN algorithm, chapter 3 presents detailed information about K-NN algorithm. And chapter 4 describes our implementation. Finally, chapter 5 concludes with some final remarks.

1.1 Data Mining

There are some different definitions about data mining, for example, according to the Gartner Group, "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques."(The Gartner Group).

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" (Hand, Mannila, & Smyth, 2001).

"Data mining is an interdisciplinary field bringing togther techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases" (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998).

Data;
Data are any facts, numbers, information or text that can be processed by a computer.

- Information;

The patterns, associations, or relationships among all this data can provide information. For instance, analysis of retail point of sale transaction data can obtain information on which products are selling and when.

Knowledge;

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a producer or retail merchant could determine which items are most susceptible to promotional efforts.

Data Warehouses;

The term data warehouse was coined by Bill Inmon in 1990, which he defined as "A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". Data warehousing is defined as a process of centralized data management and retrieval.

Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining (Han & Kamber, 2000).

1.1.1 Datamining Tasks

The following list shows the most common data mining tasks.

_ Description

_ Estimation

_ Prediction

_ Classification

_ Clustering

1.1.2 Some Basic Operations

Predictive:

- Regression
- Classification
- Collaborative Filtering

Descriptive:

- Clustering / similarity matching
- Association rules and variants
- Deviation detection

1.1.3 Application Areas of Data Mining

- Marketing Industry
  Target marketing, CRM, market basket analysis, cross selling, market segmentation
- Biology, Medicine and Genetic
- Risk analysis and management
  Forecasting, Customer Retention, Improved Underwriting, Quality Control, Competitive Analysis
- Chemistry
- Image Recognition and Robot Vision Systems
- Text Mining
- Web Mining

1.1.4 Data Mining and Other Disciplines

Data mining works together with the disciplines like machine learning, pattern recognization, database technologies, statistics, artificial intelligence, expert systems, data visualization.

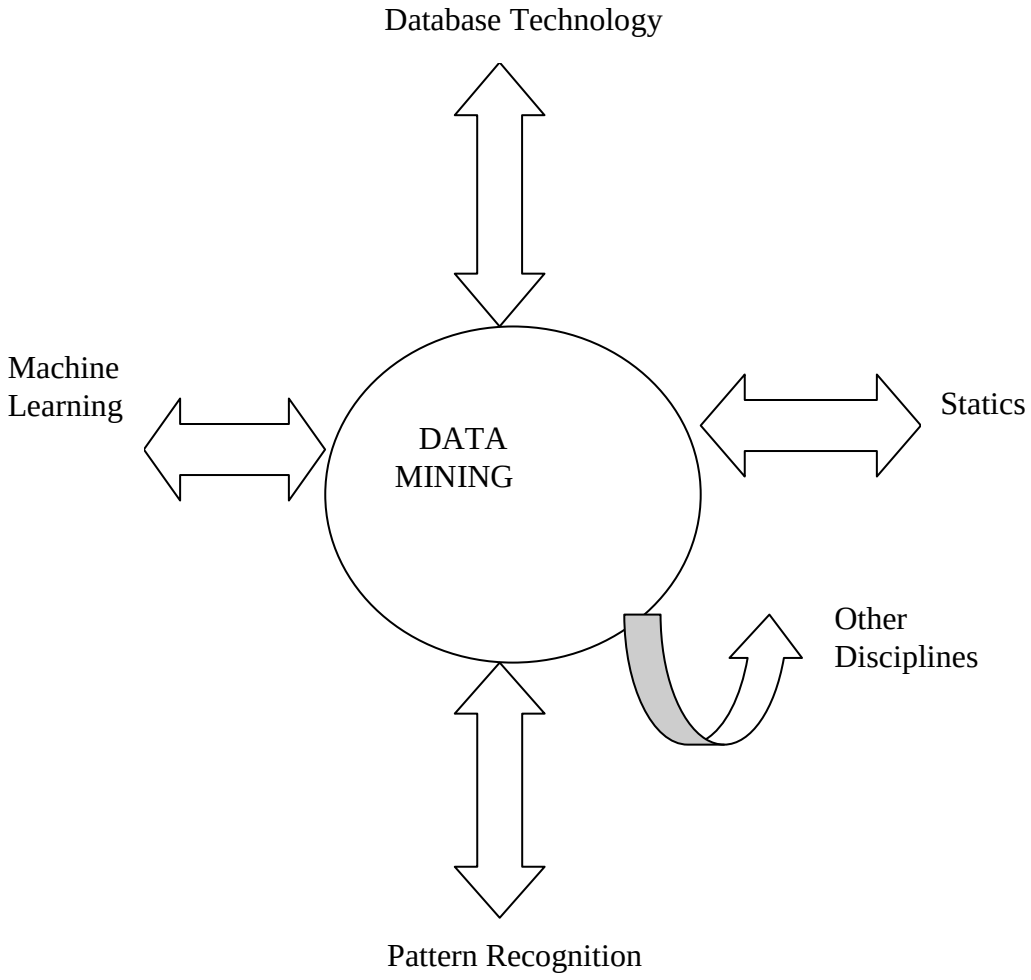

Figure 1.1 Data mining with other disciplines

1.1.5 Sample Data Mining Algorithms

Most Popular Data Mining Algorithms;

- Classification

» C4.5: Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann., 1993.

» CART: L. Breiman, J. Friedman, R. Olshen, & C. Stone. Classification and Regression Trees. Wadsworth, 1984.

» K Nearest Neighbours (kNN): Hastie & Tibshirani, 1996. Discriminant Adaptive Nearest Neighbor Classification.

» Naive Bayes Hand, D.J., Yu, K., 2001.

- Statistical Learning

»SVM: Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag.

»EM: McLachlan, G. & Peel, D. (2000). Finite Mixture Models. J. Wiley, New York. Association Analysis

»Apriori: Rakesh Agrawal & Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In VLDB '94.

»FP-Tree: Han, J., Pei, J., & Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00

- Clustering

»K-Means: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967.

»BIRCH: Zhang, T., Ramakrishnan, R., & Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD '96.

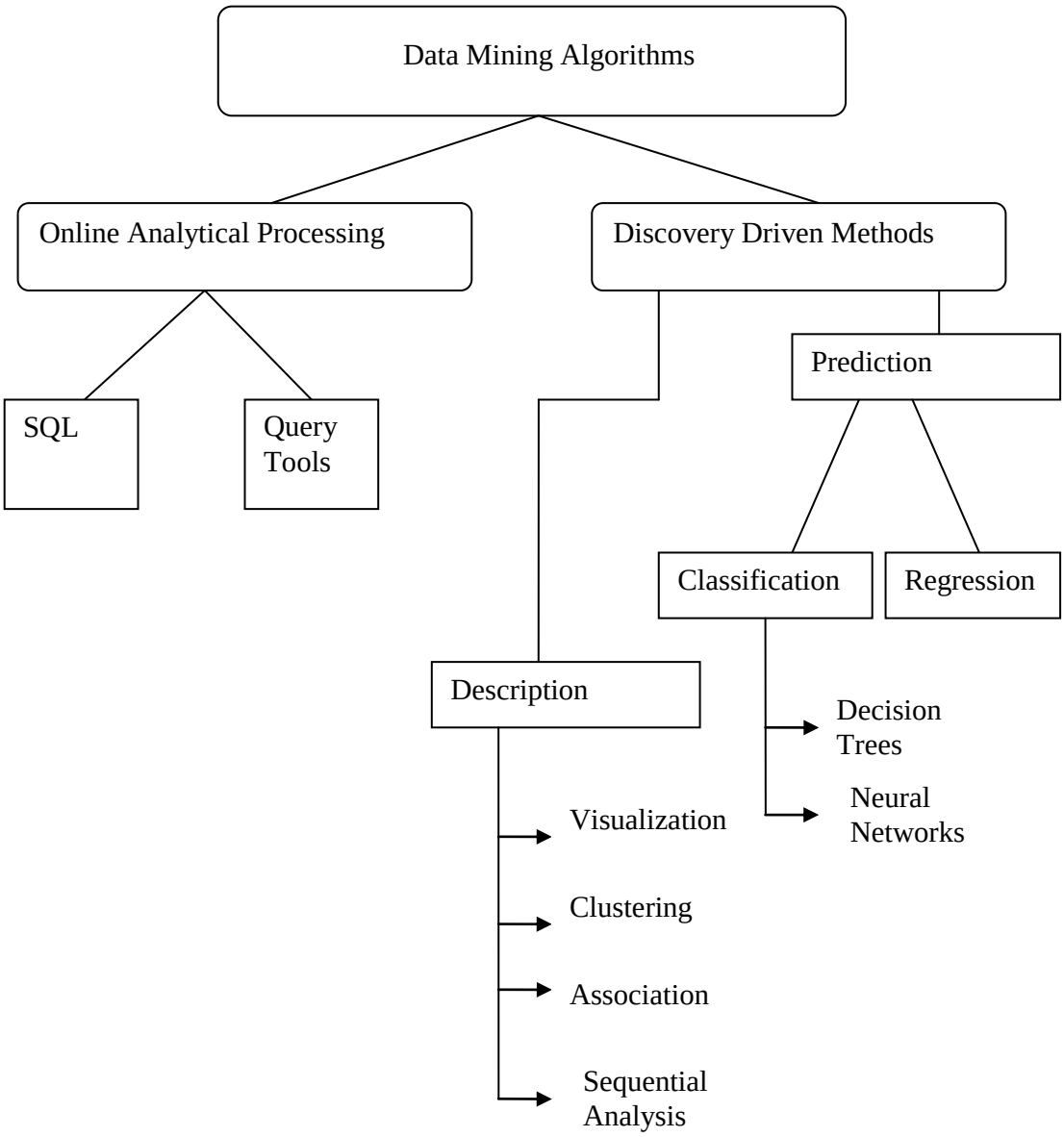

Figure 1.2 Grouping of data mining algorithms

1.1.6 Major Challenges in Data Mining

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, stream, and incremental mining methods
- Handling high-dimensionality

- Handling noise, uncertainty, and incompleteness of data
- Incorporation of constraints, expert knowledge, and background knowledge in data mining
- Pattern evaluation and knowledge integration
- Mining diverse and heterogeneous kinds of data: e.g., bioinformatics, Web, software/system engineering, information networks
- Application-oriented and domain-specific data mining
- Invisible data mining (embedded in other functional modules)
- Protection of security, integrity, and privacy in data mining

1.1.7 Knowledge Discovery (KDD) Process

- Data mining is core of knowledge discovery process

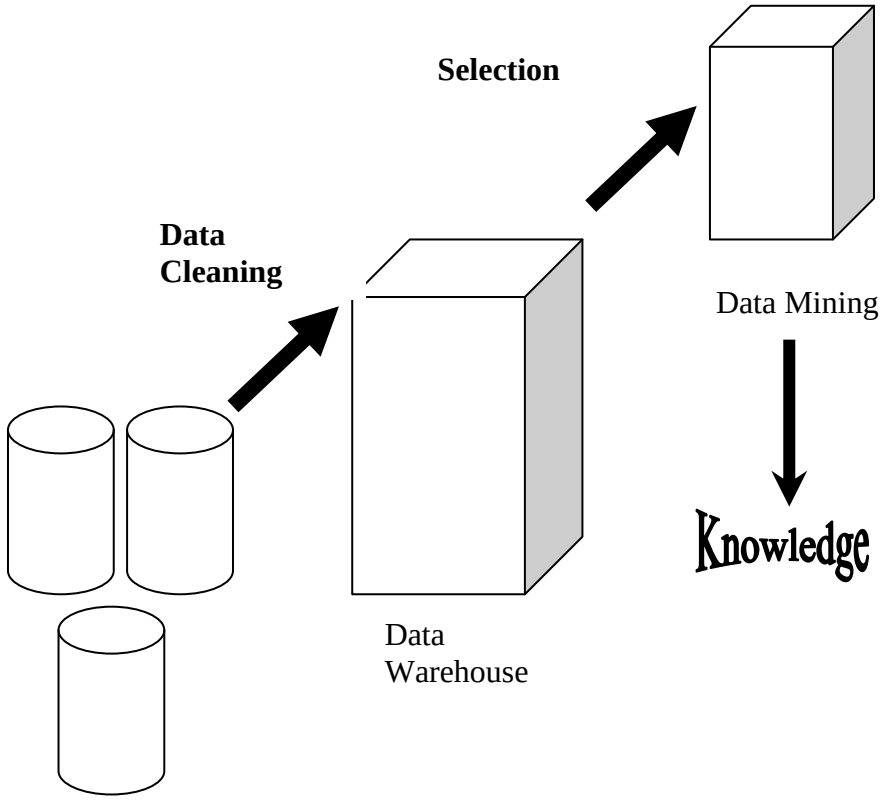

Figure 1.3 KDD process

Learning the application domain

»Relevant prior knowledge and goals of application

- Creating a target data set: data selection
- Data cleaning and preprocessing
- Data reduction and transformation

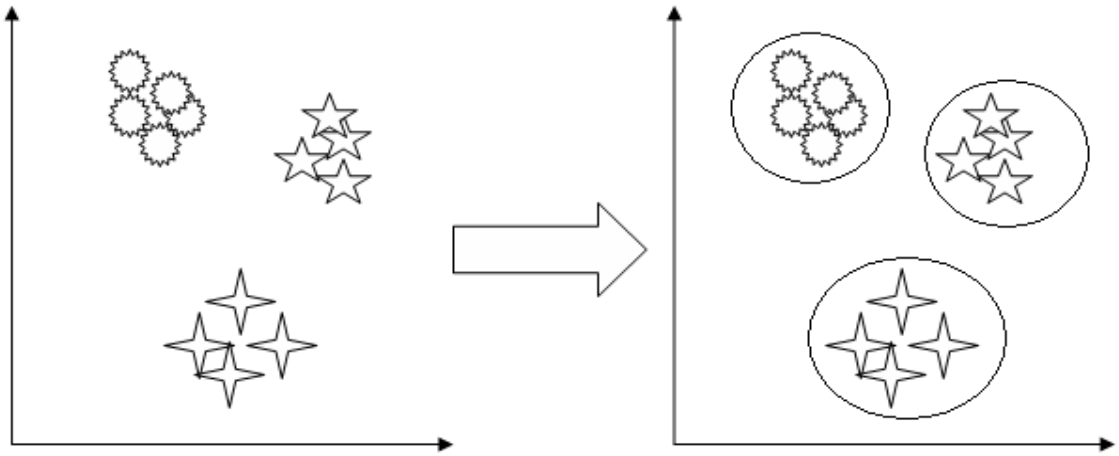

Figure 1.4 Grouping a set of data objects into clusters

1.2 Clustering

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines.

Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics.

A definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. We can show this with a simple graphical example:

1.2.1 Clustering Techniques

Clustering algorithms may be classified as listed below:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering
- Partitioning Algorithms
- Grid-Based Algorithms
- Algorithms Based on Co-Occurrence of Categorial Data
- Constraint-Based Clustering
- Evolutionary Algorithms
- Scalable Clustering Algorithms
- Algorithms for High Dimensional Data

Centered - Based Partitioning Clustering Techniques:

- K-Means Clustering Algorithm
- K-Medoids Clustering Algorithm
- CLARANS Algorithm

Hierarchical Clustering Techniques :

- Agglomerative Clustering Techniques
- Divisive Clustering Techniques
- BIRCH Algorithm
- CURE Algorithm
- ROCK Algorithm
- CHAMELEON Algorithm

Density Based Clustering Techniques:

- DBSCAN: Martin Ester, Hans-Peter Kriegel, Jörg Sander & Xiaowei Xu (KDD'96)

- OPTICS: Ankerst, Breunig, Kriegel, & Sander (SIGMOD'99).
- DENCLUE: Hinneburg & D. Keim (KDD'98)
- CLIQUE: Agrawal, Gehrke, Gunopulos, & Raghavan (SIGMOD'98)

Grid-Based Clustering Techniques :

- STING Algorithm
- WaveCluster Algorithm

1.2.2 Where to Use Clustering?

- Data mining
- Information retrieval
- Text mining
- Web analysis
- Marketing
- Medical diagnostic

1.2.3 Possible Applications

Clustering algorithms can be applied in many fields, for instance:

- Marketing: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- Biology: classification of plants and animals given their features;
- Libraries: book ordering;
- Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- City-planning: identifying groups of houses according to their house type, value and geographical location;
- Earthquake studies: clustering observed earthquake epicenters to identify dangerous zones;
- WWW: document classification; clustering weblog data to discover groups of similar access patterns.

1.2.4 Distance Functions

- Numeric data: euclidean, manhattan distances
- Categorical data: 0/1 to indicate presence/absence followed by

✓ Hamming distance (dissimilarity)

✓ Jaccard coefficients

✓ data dependent measures: similarity of A and B depends on co-occurance with C.

- Combined numeric and categorical data:

✓ weighted normalized distance

1.2.5 Distance Measurements

- Euclidean Distance
- Minkowski Distance
- Mahalanobis D2 Distance
- Hotelling T2 Distance
- Canberra Distance
- Manhattan City-Block Distance

1.3 Classification

Definition;

- Given a collection of records (training set)

–Each record contains a set of attributes, one of the attributes is the class.

- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as correctly as possible.

– A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to confirm it.

Basic classification techniques used in data mining are decision trees, bayesian classification, neural network approach and genetic algorithms.

1.3.1 Classification Techniques

Several major kinds of classification method including;

- Decision Tree based Methods
- Rule-based Methods
- Memory based reasoning
- Neural Networks
- Naive Bayes and Bayesian Belief Networks
- Support Vector Machines

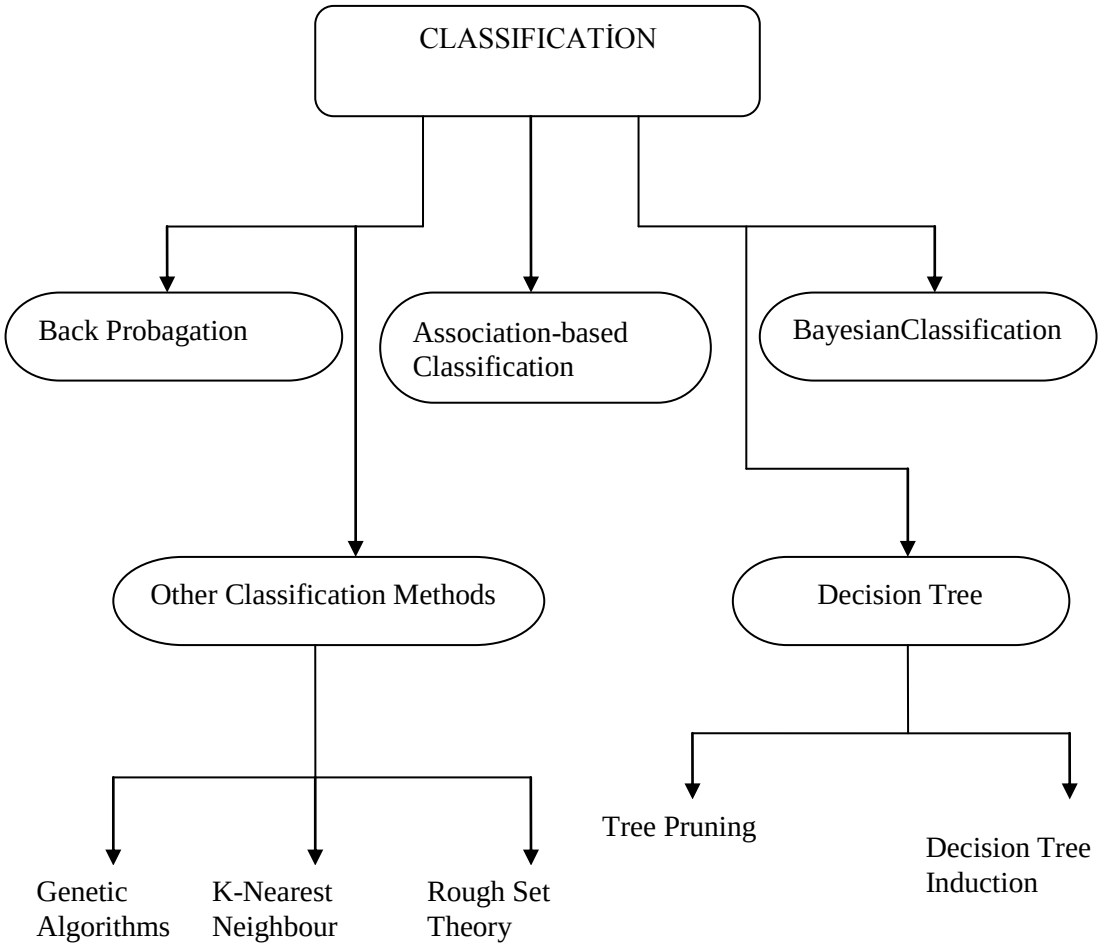

Figure 1.5 Classification techniques

Decision tree induction is a classification technique for partitioning a training file into a set of rules. A decision tree comes out of nodes and edges. Each node is labeled with attribute names and each edge is labeled with possible values for this attribute. The starting node is called the root node. Depending upon the results of a test, the training files are partitioned into two or more sub-sets. The end result is a set of rules covering all the possibilities. They are labeled with different classes. To construct a decision tree, we could start with the first attribute, find a certain threshold, go on to the next one, find a certain threshold, and repeat this process until we have made a correct classification for our customers, thus creating a decision tree for our database. The tree induction algorithms scale up very well for large data sets (Who, 2001).

Tree pruning, when a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of over fitting the data. Such methods typically use statistical measures to remove the least reliable branches, generally resulting in faster classification and an improvement in the ability of the tree to correctly classify independent test data.

Backpropagation is a neural network learning algorithm. The field of neural networks was originally kindled by psychologists and neurobiologists who sought to develop and test computational analogues of neurons. Roughly speaking, a neural network is a set of connected input/output units where each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input samples. Neural network learning is also referred to as connectionist learning due to the connections between units. (Who, 2001).

Neural networks involve long training times, and are therefore more suitable for applications where this is feasible. They require a number of parameters which are typically best determined empirically, such as the network topology or "structure". Neural networks have been criticized for their poor interpretability, since it is difficult for humans to interpret the symbolic meaning behind the learned weights. These features initially made neural networks less desirable for data mining.

Advantages of neural networks, however, include their high tolerance to noisy data as well as their ability to classify patterns on which they have not been trained. In addition, several algorithms have recently been developed for the extraction of rules from trained neural networks. These factors contribute towards the usefulness of neural networks for classification in data mining.

Other classification methods

* K - Nearest Neighbor Method is based on learning by analogy. The training samples are described by n-dimensional numeric attributes. Each sample represents a

point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space.

* Case-based reasoning (CBR) classifiers are instanced-based. Unlike nearest neighbor classifiers, which store training samples as points in Euclidean space, the samples or "cases" stored by CBR are complex symbolic descriptions.

* Genetic algorithms attempt to incorporate ideas of natural evolution. In general, genetic learning starts as follows. An initial population is created consisting of randomly generated rules. Each rule can be represented by a string of bits.

*Rough set theory can be used for classification to discover structural relationships within imprecise or noisy data. It applies to discrete-valued attributes.
Continuous-valued attributes must therefore be discretized prior to its use. Rough set theory is based on the establishment of equivalence classes within the given training data. All of the data samples forming an equivalence class are indiscernible, that is, the samples are identical with respect to the attributes describing the data (Who, 2001).

1.3.2 Where to Use Classification?

- Computer vision
- Medical ismaging and medical image analysis
- Optical character recognition
- Drug discovery and development
- Handwriting recognition
- Natural language processing
- Document classification

CHAPTER TWO
DBSCAN ALGORITHM

DBSCAN is a density based clustering method that converts the high-density objects regions into clusters with arbitrary shapes and sizes. It was first introduced by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996, and depend on a density-based notion of clusters. Clusters are recognized by looking at the density of points. This algorithm is especially agreeable to implement with large datasets, with noise, and is able to define clusters with different sizes and shapes.

Dbscan is based on two main concepts: *density reachability* and *density connectability*. These both concepts depend on two input parameters of the dbscan clustering: *the size of epsilon neighborhood* ε and *the minimum points in a cluster*. The number of points parameter impacts detection of outliers. Points are declared to be outliers if there are few othe points in the ε -Eucledean neighborhood. ε parameter controlls the size of the neighborhood, as well as the size of the clusters. If the ε is big enough, the would be one big cluster.

A cluster, which is a subset of the points of the database, satisfies two properties:

1. All points within the cluster are mutually density-connected.
2. If a point is density-connected to any point of the cluster, it is part of the cluster as well.

2.1 Related Work

In the literature, the DBSCAN algorithm has been used in many studies. Many researchers have found that the DBSCAN algorithm accomplishes very good performance in their experiments on different data sets.

The first studies with DBSCAN algorithm developed by Ester et al. (1996) and Sander et al. (1998). They introduced the approach of DBSCAN to address the detection of clusters in a spatial database according to a difference in density.

Performance analysis of DBSCAN algorithm has been identified with a variety of algorithms. For example, *Density-based clustering algorithms – DBSCAN and SNN*, the use of these two density-based clustering algotihms have been discussed. These algorithms were implemented within the context of the LOCAL project as part of a task that aims to create models of the geographic space (Space Models) to be used in context-aware mobile systems. There, the role of the clustering algorithms is to identify clusters of Points of Interest (POIs) and then use the clusters to automatically characterize geographic regions (Moreira, Santos, & Carneiro, 2005).

In contrast to the existing density-based clustering algorithms, a new density-based clustering algorithm, *ST-DBSCAN: An algorithm for clustering spatial–temporal data* which is based on DBSCAN were presented. They proposed three marginal extensions to DBSCAN related with the identification of core objects, noise objects, and adjacent clusters. In contrast to the existing density-based clustering algorithms, their algorithm had the ability of discovering clusters according to non-spatial, spatial and temporal values of the objects. In their project, they also presented a spatial-temporal data warehouse system designed for storing and clustering a wide range of spatial-temporal data. They showed an implementation of their algorithm by using this data warehouse and presented the data mining results (Birant & Kut, 2007).

In just a few years, gene expression microarrays have rapidly become a standard experimental tool in the biological and medical research. The estimation of the number of clusters in datasets is one of the main problems of clustering microarrays**.** The DBSCAN and other existing methods were compared using the microarray data from two datasets used for diagnosis of leukemia and lung cancer. This work reported the application of DBSCAN algorithm that proved to be useful in determining the number of clusters in a microarray experiments (Raczynski, Wozniak, Rubel, & Zaremba, 2010).

Researchers are putting their best work to reach the fast and well-organized algorithm on medical datasets. Different data mining techniques for effective

implementation of clinical data were developed (Saranya & Hemalatha, 2011). The main aim of their work was to discover various data mining techniques such as DBSCAN, CURE, CLARANS, BIRCH, and STING on clinical and spatial datasets.

Most of the recent work on spatial data has used various clustering techniques due to the nature of the data. Spatial Data Mining is the process of discovering interesting and previously unknown but potentially useful patterns from large spatial datasets. For example, there was given a detailed survey of the existing density based algorithms namely *DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN* and *DBCLASD* based on the essential parameters needed for a good clustering algorithm in spatial data. (Parimala, Lopez, & Senthilkumar, 2011).

2.2 Algorithm

DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (MinPts). It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ε-environment of a different point and hence be made part of a cluster.

If a point is found to be part of a cluster, its ε-neighborhood is also part of that cluster. Hence, all points that are found within the ε-neighborhood are added, as is their own ε-neighborhood. This process continues until the cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

Algorithm DBSCAN (*D, Eps, MinPts*)
// Precondition: All objects in *D* are unclassified.
FORALL objects *o* in *D* DO:
IF *o* is unclassified
call function *expand_cluster* to construct a cluster wrt.
Eps and MinPts containing *o*.
FUNCTION *expand_cluster* (*o, D, Eps, MinPts*):
retrieve the *Eps*-neighborhood *NEps*(*o*) of *o*;
IF | *NEps*(*o*) | < *MinPts // i.e. o is not a core object*
mark o as noise and RETURN;
ELSE // i.e. o is a core object
select a new cluster-id and mark all objects in NEps(*o*)
with this current cluster-id;
push all objects from NEps(*o*)*\{o} onto the stack seeds;*
WHILE NOT seeds.empty() DO
currentObject := seeds.top();
retrieve the *Eps*-neighborhood *NEps*(*currentObject*)
of *currentObject*;
IF | *NEps*(*currentObject*) | ³ *MinPts*
select all objects in NEps(*currentObject*) *not yet*
classified or are marked as noise,
push the unclassified objects onto seeds
and mark all of these objects with current
cluster-id;
seeds.pop();
RETURN

Figure 2.1 Algorithm DBSCAN (Ester, Kriegel, Sander, Wimmer, & Xu, 96)

2.3 Pseudocode

```
DBSCAN(D, eps, MinPts)
  C = 0
  for each unvisited point P in dataset D
```

```
    mark P as visited
    N = getNeighbors (P, eps)
    if sizeof(N) < MinPts
      mark P as NOISE
    else
      C = next cluster
      expandCluster(P, N, C, eps, MinPts)

expandCluster(P, N, C, eps, MinPts)
  add P to cluster C
  for each point P' in N
    if P' is not visited
      mark P' as visited
      N' = getNeighbors(P', eps)
      if sizeof(N') >= MinPts
        N = N joined with N'
    if P' is not yet member of any cluster
      add P' to cluster C
```

Density based spatial clustering of applications with noise, DBSCAN; rely on a density-based notion of clusters, which is designed to discover clusters of arbitrary shape and also have ability to handle noise. The main task of this algorithm is class identification, i.e. the grouping of the objects into meaningful subclasses.

Two global parameters for DBSCAN algorithms are:

• Eps: Maximum radius of the neighborhood

• MinPts: Minimum number of points in an Eps-neighborhood of that point

Core Object: Object with at least MinPts objects within a radius ‘Eps-neighborhood’

Border Object: Object that on the border of a cluster

NEps(p): {q belongs to D | dist(p,q) <= Eps}

Directly Density-Reachable: A point p is directly density-reachable from a point q w.r.t Eps, MinPts if

p belongs to NEps(q)

|*NEps (q)*| >= *MinPts*

Density-Reachable: A point *p* is density-reachable from a point *q* w.r.t *Eps*, *MinPts* if there is a chain of
points *p1*, …, *pn*, *p1 = q*, *pn = p* such that *pi+1* is directly density-reachable from *pi*
Density-Connected: A point *p* is density-connected to a point *q* w.r.t *Eps*, *MinPts* if there is a point *o* such
that both, *p* and *q* are density-reachable from *o* w.r.t *Eps* and *MinPts*.

Algorithm:
• Arbitrary select a point *p*
• Retrieve all points density-reachable from *p* w.r.t *Eps* and *MinPts*.
• If *p* is a core point, a cluster is formed.
• If *p* is a border point, no points are density-reachable from *p* and DBSCAN visits the next point of the database.
• Continue the process until all of the points have been processed.

2.4 Simularity Measures

Common distance functions:

- The Euclidean distance, A review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance
- The Manhattan distance
- The maximum norm
- The Mahalanobis distance corrects data for different scales and correlations in the variables
- The angle between two vectors can be used as a distance measure when clustering high dimensional data.
- The Hamming distance measures the minimum number of substitutions required to change one member into another.

Since similarity is fundamental to the definition of a cluster, a measure of the similarity between two patterns drawn from the same feature space is essential to most clustering procedures. Because of the variety of feature types and scales, the distance measure (or measures) must be chosen carefully. It is most common to calculate the dissimilarity
between two patterns using a distance measure defined on the feature space. We will focus on the well-known distance measures used for patterns whose features are all continuous. The most popular metric for continuous features is the Euclidean distance.

$$d_2\left(x_i, x_j\right) = \left(\sum_{k=1}^{d}\left(x_{i,k} - x_{j,k}\right)^2\right)^{1/2}$$

$$= \| x_i - x_j \|_2,$$

which is a special case (p52) of the Minkowski metric

$$d_p\left(x_i, x_j\right) = \left(\sum_{k=1}^{d} \mathrm{I} x_{i,k} - x_{j,k} \mathrm{I}^p\right)^{1/p}$$

$$= \| x_i - x_j \|_p$$

The Euclidean distance has an intuitive appeal as it is commonly used to evaluate the proximity of objects in two or three-dimensional space. It works well when a data set has “compact” or “isolated” clusters (Mao & Jain, 1996).

The drawback to direct use of the Minkowski metrics is the tendency of the largest-scaled feature to dominate the others. Solutions to this problem include normalization of the continuous features (to a common range or variance) or other weighting schemes. Linear
correlation among features can also distort distance measures; this distortion can be alleviated by applying a whitening transformation to the data or by using the squared Mahalanobis distance

$$d_{M(Xi,X_j)} = (x_i - x_j)\sum^{-1}(x_i - x_j)^T ,$$

where the patterns **x***i* and **x***j* are assumed to be row vectors, and S is the sample covariance matrix of the patterns or the known covariance matrix of the pattern generation process; *dM*~z , z! assigns different weights to different features based on their variances and pairwise linear correlations. Here, it is implicitly assumed that class conditional densities are unimodal and characterized by multidimensional spread, i.e., that the densities are multivariate Gaussian. The regularized Mahalanobis distance was used in Mao and Jain (1996) to extract hyperellipsoidal clusters. Recently, several researchers have used the Hausdorff distance in a point set matching context (Dubuisson & Jain, 1994).

2.5 Advantages- Disadvantages

Advantages;

- DBSCAN does not require you to know the number of clusters in the data a priori, as opposed to k-means.
- DBSCAN can find arbitrarily shaped clusters. It can even find clusters completely surrounded by (but not connected to) a different cluster. Due to the MinPts parameter, the so-called single-link effect (different clusters being connected by a thin line of points) is reduced.
- DBSCAN has a notion of noise.
- DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database. (Only points sitting on the edge of two different clusters might swap cluster membership if the ordering of the points is changed, and the cluster assignment is unique only up to isomorphism.)

Disadvantages;

- DBSCAN needs to materialize the distance matrix for finding the neighbords. It has a complexity of $O((n^2-n)/2)$ since only an upper matrix is needed. Within the distance matrix the nearest neighbors can be detected by selecting a tuple with minimums functions over the rows and columns. Databases solve the neighborhood problem with indexes specifically designed for this type of application. For large scale applications, you cannot afford to materialize the distance matrix.
- Finding neighbords is an operation based on distance (generally the Euclidean distance) and the algorithm may find the curse of dimensionality problem
- DBSCAN cannot cluster data sets well with large differences in densities, since the minPts-epsilon combination cannot be chosen appropriately for all clusters then.

2.6 Complexity

DBSCAN visits each point of the database, possibly multiple times (e.g., as candidates to different clusters). For practical considerations, however, the time complexity is mostly governed by the number of getNeighbors queries. DBSCAN executes exactly one such query for each point, and if an indexing structure is used that executes such a neighborhood query in $O(\log n)$, an overall runtime complexity of $O(n.\log n)$, is obtained. Without the use of an accelerating index structure, the run time complexity is $O(n^2)$.

CHAPTER THREE
K-NN(K-NEAREST NEIGHBOUR) ALGORITHM

An instance based learning method called the K-Nearest Neighbor or K-NN algorithm has been used in many applications in areas such as data mining, statistical pattern recognition, image processing. Successful applications include recognition of handwriting, satellite image and EKG pattern.

The k Nearest Neighbour (k-NN) method is a widely used technique which has found several applications in clustering and classification.

In pattern recognition, the k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning.

The training examples are mapped into multidimensional feature space. The space is partitioned into regions by class labels of the training samples. A point in the space is assigned to the class *c* if it is the most frequent class label among the *k* nearest training samples. Usually Euclidean distance is used.

The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the actual classification phase, the same features as before are computed for the test sample (whose class is not known). Distances from the new vector to all stored vectors are computed and *k* closest samples are selected. The new point is predicted to belong to the most numerous class within the set.

The best choice of *k* depends upon the data; generally, larger values of *k* reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good *k* can be selected by parameter optimization using, for example, cross-validation. The special case where the class is predicted to be the class of the closest training sample (i.e. when $k = 1$) is called the nearest neighbour algorithm.

The accuracy of the k-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the features scales are not consistent with their relevance. Much research effort has been placed into selecting or scaling features to improve classification. A particularly popular approach is the use of evolutionary algorithms to optimize feature scaling. Another popular approach is to scale features by the mutual information of the training data with the training classes.

The algorithm is easy to implement, but it is quantitively intensive, especially when the size of the training set grows. Many optimizations have been proposed over the years; these generally seek to reduce the number of distances actually computed. Some optimizations involve partitioning the feature space, and only computing distances within specific nearby volumes. Several different types of nearest neighbour finding algorithms include:

- Linear scan
- Kd-trees
- Balltrees
- Metric trees
- Locality-sensitive hashing (LSH)

3.1 Related Work

The k-nearest neighbor method has been frequently used for the classification of biological and medical data. While searching, we attached mostly on the methods they used. For this purpose we investigated some thesis and article mentioned below.

Application of K-nearest neighbors algorithm on breast cancer diagnosis problem, there was a work which addressed the Breast Cancer diagnosis problem as a pattern classification problem. Specifically, the problem was studied using the Wisconsin-Madison Breast Cancer data set. The K-nearest neighbors algorithm was employed as the classifier. The K-NN algorithm assigned the class label of the new datum based on the class label that most of the K-closest training data process. The K-NN

algorithm yielded the best classification performance that was obtained so far on this problem. (Sarkar & Leong, 2000).

The automatic diagnosis of breast cancer is an important, real-world medical problem. There was a project, Fuzzy Logic system designed for diagnosing and analyzing the breast cancer and the learning procedure of this system was described. For their purpose they dealt with Wisconsin Breast Cancer Database (WBCD). The system extracted classification rules from trained network based on Fuzzy Logic. Analyzing both malignant and benign cell features, he could also generated the rules for classification depending on the cell features using Fuzzy Inference System (FIS) editor using MATLAB. In the project, they described the accuracy of the trained networks and compare the result with the outputs of the classifiers constructed by using both k-nearest neighbor (KNN) and Bayes classifier. (Kırtulukoğlu, 2009).

In another study, classification methods for biomedical analysis were used in a project. Their sample database had breast cancer data which was one of the most cause of cancer, because detection of this cancer was very important. Database had 9 attributes which was used for classification. These attributes were numerical numbers. After having numerical numbers via FNA procedure, class labels could be given both by classical examinations and by classification algorithms. Their database included a class column which real diagnosis exists. The study aimed to consider classification algorithms results carefully. Different methods and algorithms had been used; classification accuracies had been given depending on real values. Results were compared and some ideas could arise for using software programs to classify sickness instances. Computer supported diagnosis could be used more commonly. (Güneşer, 2009).

Another study, a system aimed at classifying the medical data by using artifical immune system and k-nn classification algorithm, is proposed. In the proposed classification structure, with the artifical immune system, the features which characterize the data are selected, and with k-nn, these features which are reduced

from data, are classified. For his purpose, experiments are done about the data of protein settlement place of E.coli bacterium which are taken from UCI data base. (Kaymaz, 2007)

Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. There was a research intended to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that were in use in today's medical research particularly in heart disease prediction. This research aimed to analyze the different predictive/ descriptive data mining techniques such as KNN, Neural Networks, Naive Bayes, proposed in recent years for the diagnosis of heart disease. (Soni, 2011)

3.2 Algorithm

The k-nearest neighbor algorithm is within the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor.

1 NN;

• Estimate the same value/class as the nearest instance in the training set

k-NN;

• Find the k closest training points

• Predicted class: majority vote

• Predicted value: average weighted by inverse distance

k-NN Classification;

- Calculate distances of all training vectors to test vector
- Pick k closest vectors
- Calculate average/majority

Neighborhood size;

• Choice of k

• Smaller k higher variance (less stable)

• Larger k higher bias (less precise)

• Proper choice of k dependends on the data

• Adaptive methods, heuristics

• Cross-validation

Distance metric;

• Distance used: Euclidean, Manhattan, etc.

• Issue: scaling of different dimensions

• Selecting/scaling features: common problem for all methods

• but affects k NN even more

→ use mutual information between feature and output

3.3 Parameter Selection

The best option of *k* depends to the data; usually, larger values of *k* reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good *k* can be selected by various heuristic techniques, for example, cross-validation. The special case where the class is predicted to be the class of the closest training sample (i.e. when $k = 1$) is called the nearest neighbor algorithm.

The accuracy of the *k*-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. Much research effort has been put into selecting or scaling features to improve classification. A particularly popular approach is the use of evolutionary algorithms to optimize feature scaling. Another popular approach is to scale features by the mutual information of the training data with the training classes. In binary (two class) classification problems, it is helpful to choose *k* to be an odd number as this avoids tied votes.

3.4 Distance Function

We have seen above how, for a new record, the *k*-nearest neighbor algorithm assigns the classification of the most similar record or records. But just how do we define similar?

Data analysts define distance metrics to measure similarity. A distance metric or distance function is a real-valued function *d*, such that for any coordinates *x*, *y*, and *z*:

1. $d(x,y) \geq 0$, and $d(x,y) = 0$ if and only if $x = y$
2. $d(x,y) = d(y,x)$
3. $d(x,z) \leq d(x,y) + d(y,z)$

Property 1 assures us that distance is always nonnegative, and the only way for distance to be zero is for the coordinates (e.g., in the scatter plot) to be the same. Property 2 indicates commutativity, so that, for example, the distance from New York to Los Angeles is the same as the distance from Los Angeles to New York. Finally, property 3 is the *triangle inequality*, which states that introducing a third point can never shorten the distance between two other points. The most common distance function is Euclidean distance, which represents the usual manner in which humans think of distance in the real world:

$$d\text{Euclidean}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

where $\mathbf{x} = x1, x2, \ldots, xm$, and $\mathbf{y} = y1, y2, \ldots, ym$ represent the *m* attribute values of two records. For example, suppose that patient A is $x1 = 12$ years old and has a Na/K ratio of $x2 = 8$, while patient B is $y1 = 20$ years old and has a Na/K ratio of $y2 = 4$.

Then the Euclidean distance between these points, as shown in Figure 3.1 is

$$d_{Euclidean}(x,y) = \sqrt{\sum_i (x_i - y_i)^2} = \sqrt{(12-20)^2 + (8-4)^2}$$

$$= \sqrt{64+16} = 8.94$$

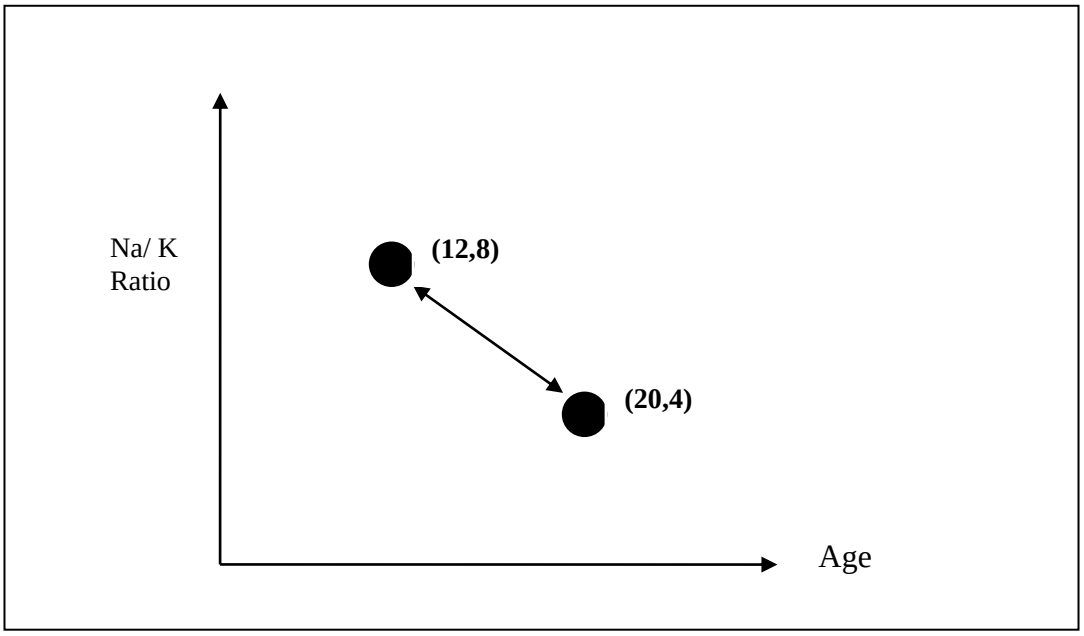

Figure 3.1 Euclidean distance

3.5 Advantages- Disadvantages

k-NN is very simple to understand and easy to implement. So it should be considered
in seeking a solution to any classification problem. Some advantages of k-NN are as follows :

– Because the process is transparent, it is easy to implement and debug.

– In situations where an explanation of the output of the classifier is useful, k-NN can be very effective if an analysis of the neighbours is useful as explanation.

– There are some noise reduction techniques that work only for k-NN that can be effective in improving the accuracy of the classifier (Cunningham & Delany, 2007).

On the other hand, some significant disadvantages are as follows:

– Because all the work is done at run-time, k-NN can have poor run-time performance if the training set is large.

– k-NN is very sensitive to irrelevant or redundant features because all features contribute to the similarity and thus to the classification. This can be ameliorated by careful feature selection or feature weighting (Cunningham & Delany, 2007).

CHAPTER FOUR
IMPLEMENTATION

In the scope of this study, a program was developed to analyse mammographic images and to obtain necessary results. This application describes the implementation of two algorithms: DBSCAN (Ester, 1996) and K-NN (Hastie & Tibshirani, 1996). These algorithms were implemented within the MIAS Database (The Mammographic Image Analysis Society Digital Mammogram Database). This program was developed by using Visual Studio. NET technology and Visual C# programming language.

Main form of the program can be seen on Figure 4.1. Some important functionalities such as DbScan clustering and K-NN classification, can be reached from main form. In the groupbox of "DBSCAN CLUSTERING", there are three button and two textbox. User can apply dbscan algorithm on the dataset by using Eps and MinPts parameters. After this application, the datas which are clustered can be written on a text file by using Output.txt button. In the groupbox of "K-NN CLASSIFICATION", there are one button and one textbox. User can apply K-NN algorithm on the noise point of the data by using K parameter.

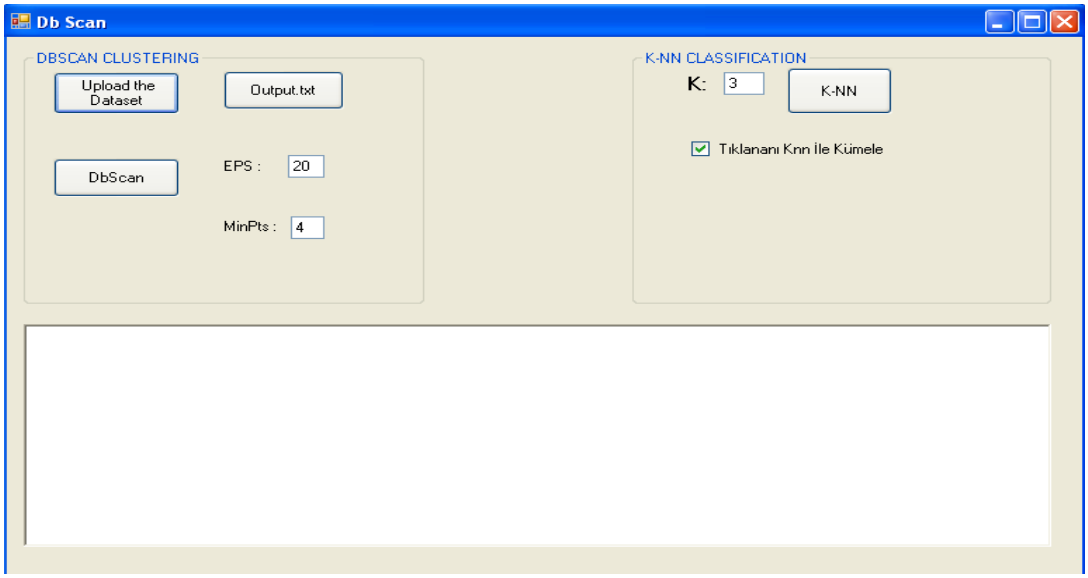

Figure 4.1 Interface of the program

By using page given in Figure 4.2, user can see detailed information of MIAS database (Suckling et al., 1994). At first, the dataset will be uploading on the screen when user click on the button which name is “Upload the Dataset”. When user click on the button “Upload the Dataset”, the red points will be appeared on the panel. The red points show that x,y image-coordinates of centre of abnormality. User can see detailed information about the MIAS database from the richtextbox on the panel.

There are 330 data in the MIAS database. 121 data of them have x,y image coordinates which can be shown as 5th, 6th columns of the database information. 209 datas of them don’t have x,y image coordinates. And they are seen as normal (NORM). Thatswhy we didn’t use these data.

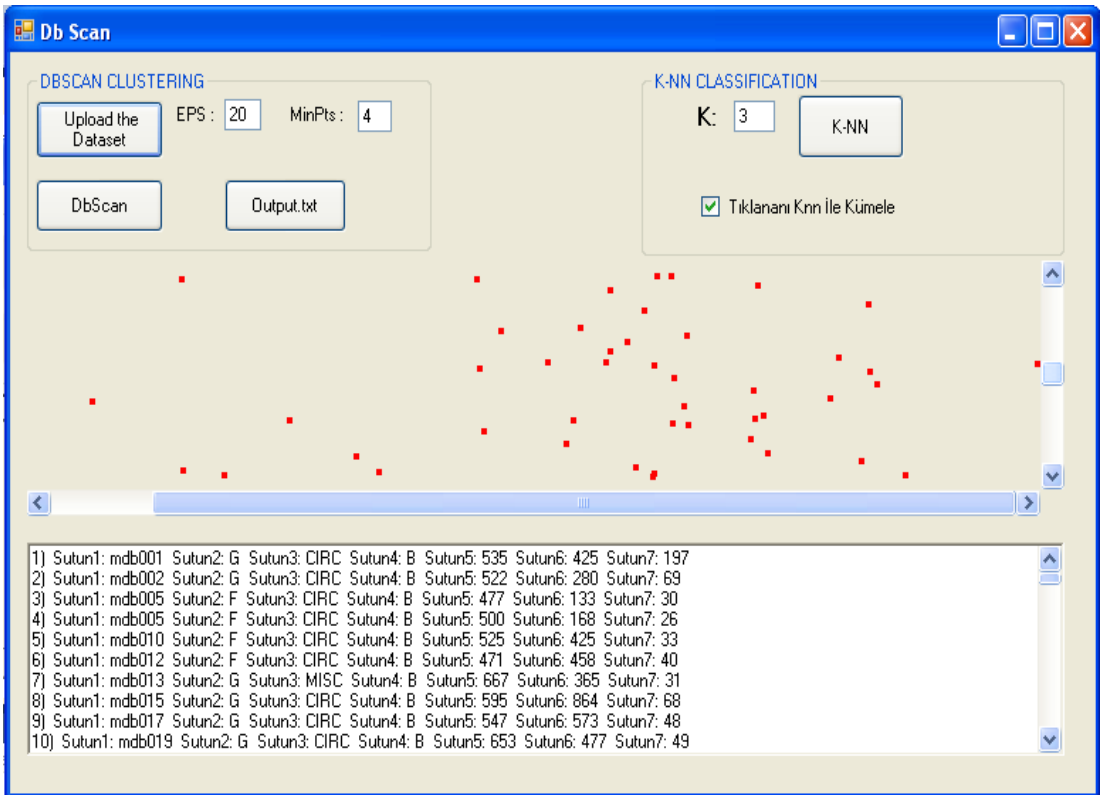

Figure 4.2 The page of uploading the dataset

In figure 4.3, user can learn that the red point clicked on belongs to which mammographic image. When user click on a red point, a message box appears on the screen and it gives information (MIAS database reference number) about which image is it.

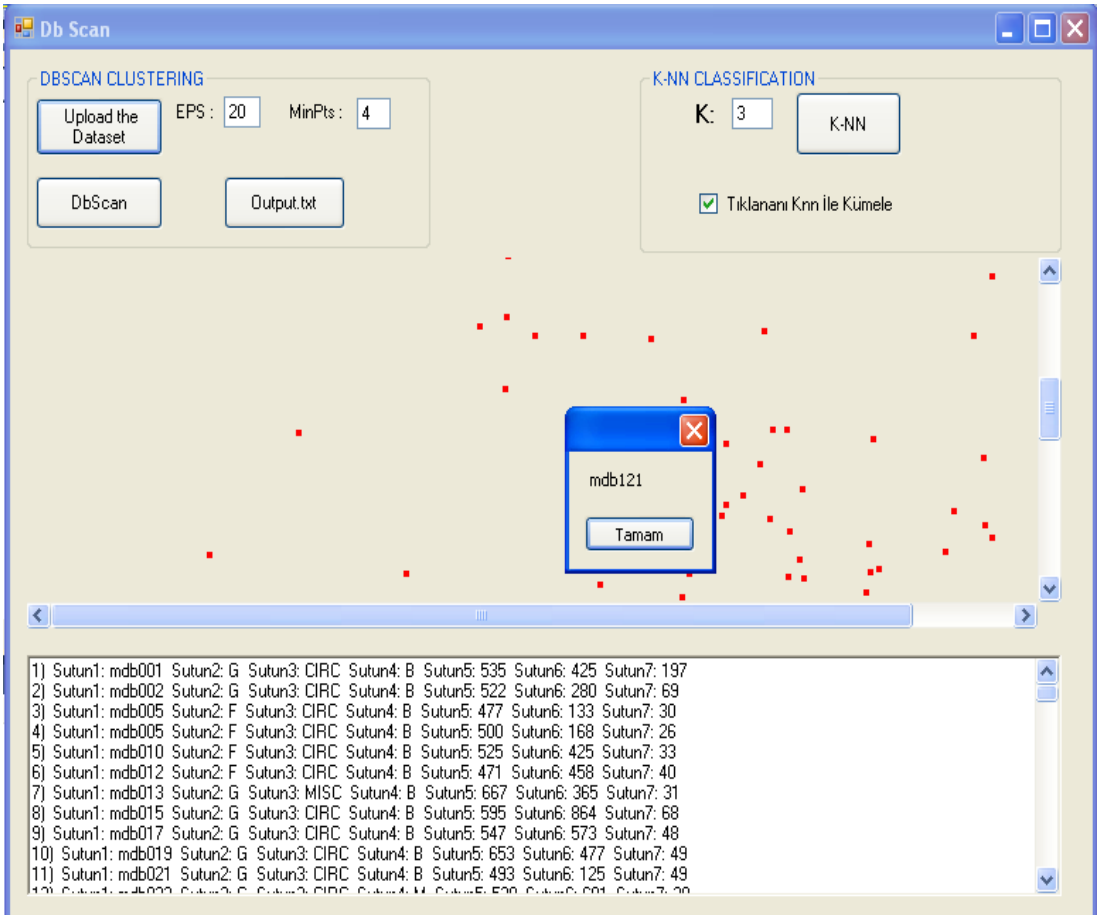

Figure 4.3 The page of uploading the dataset

Dataset information;

This file lists the films in the MIAS(The Mammographic Image Analysis Society Digital Mammogram Database) database and provides appropriate details as follows:

1st column: MIAS database reference number.

2nd column: Character of background tissue:

F - Fatty

G - Fatty-glandular

D - Dense-glandular

3rd column: Class of abnormality present:

CALC - Calcification

CIRC - Well-defined/circumscribed masses

SPIC - Spiculated masses

MISC - Other, ill-defined masses

ARCH - Architectural distortion

ASYM - Asymmetry

NORM – Normal

4th column: Severity of abnormality;

B - Benign

M - Malignant

5th,6th columns: x,y image-coordinates of centre of abnormality.

7th column: Approximate radius (in pixels) of a circle enclosing the abnormality.

Note: The size of ALL the images is 1024 pixels x 1024 pixels. The images have been centered in the matrix.

Mammography remains the key screening tool for breast abnormalities detection, because it allows identification of tumour before being palpable. Cancer begins with uncontrolled division of one cell, which results in a visible mass named tumor. Tumor can be benign or malignant. Malignant tumor grows rapidly and invades its surrounding tissues causing their damage

The Small Part of Dataset;

mdb001 G CIRC B 535 425 197

mdb002 G CIRC B 522 280 69

mdb003 D NORM

mdb004 D NORM

mdb005 F CIRC B 477 133 30

mdb005 F CIRC B 500 168 26

mdb006 F NORM

mdb007 G NORM

mdb008 G NORM
mdb009 F NORM
mdb010 F CIRC B 525 425 33
mdb011 F NORM
mdb012 F CIRC B 471 458 40
mdb013 G MISC B 667 365 31
mdb014 G NORM
mdb015 G CIRC B 595 864 68
mdb016 G NORM
mdb017 G CIRC B 547 573 48
mdb018 G NORM
mdb019 G CIRC B 653 477 49
mdb020 G NORM
mdb021 G CIRC B 493 125 49

-
-
-
-
-

mdb321 D NORM
mdb322 D NORM

Mamographic image displays of the MIAS database;

On figure 4.4 and 4.5, you can see that two examples of mamographic images in the MIAS database. First one is a normal mamographic image display (Figure 4.4) and second one is a display of a mammographic image which has an asymmetrical abnormality (Figure 4.5). In figure 4.6, you can see examples of different types of breast tissue in the MIAS database. (a) fatty, (b) glandular, and (c) dense.

In the MIAS database, there are 54 negative samples (malignant), 67 positive samples (benign) and 209 normal samples.

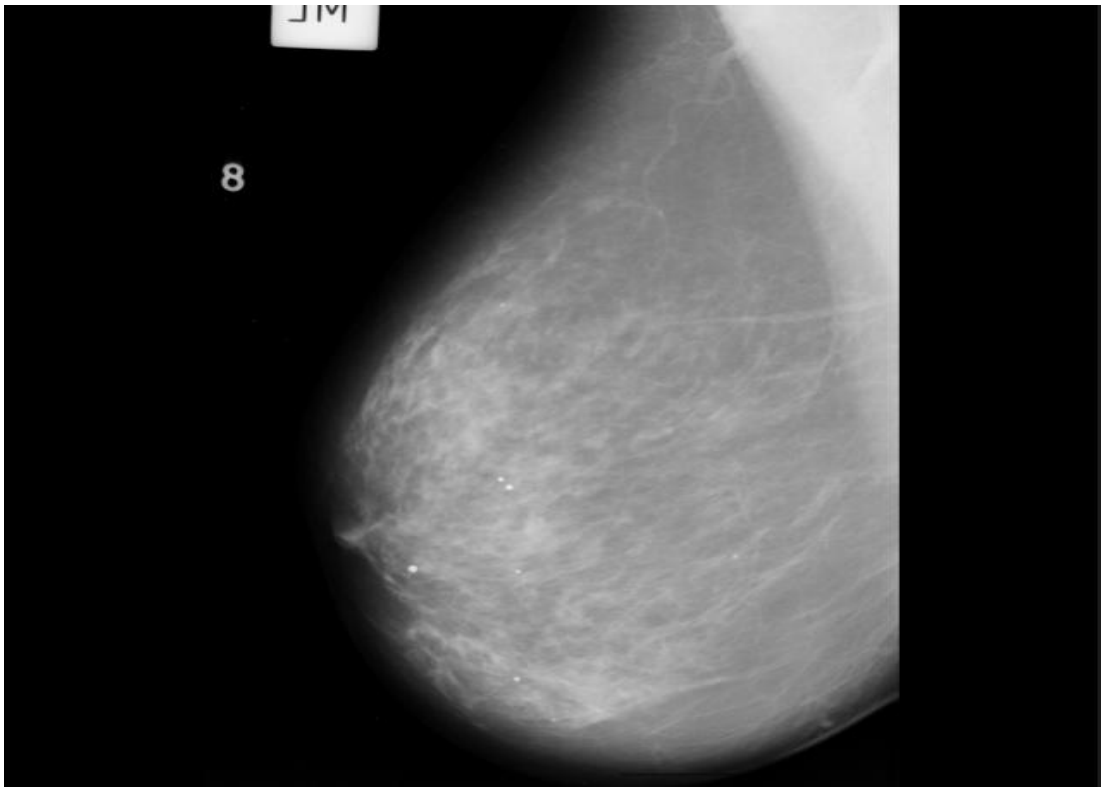

Figure 4.4 Display of a normal mammographic image (mdb057 D NORM)

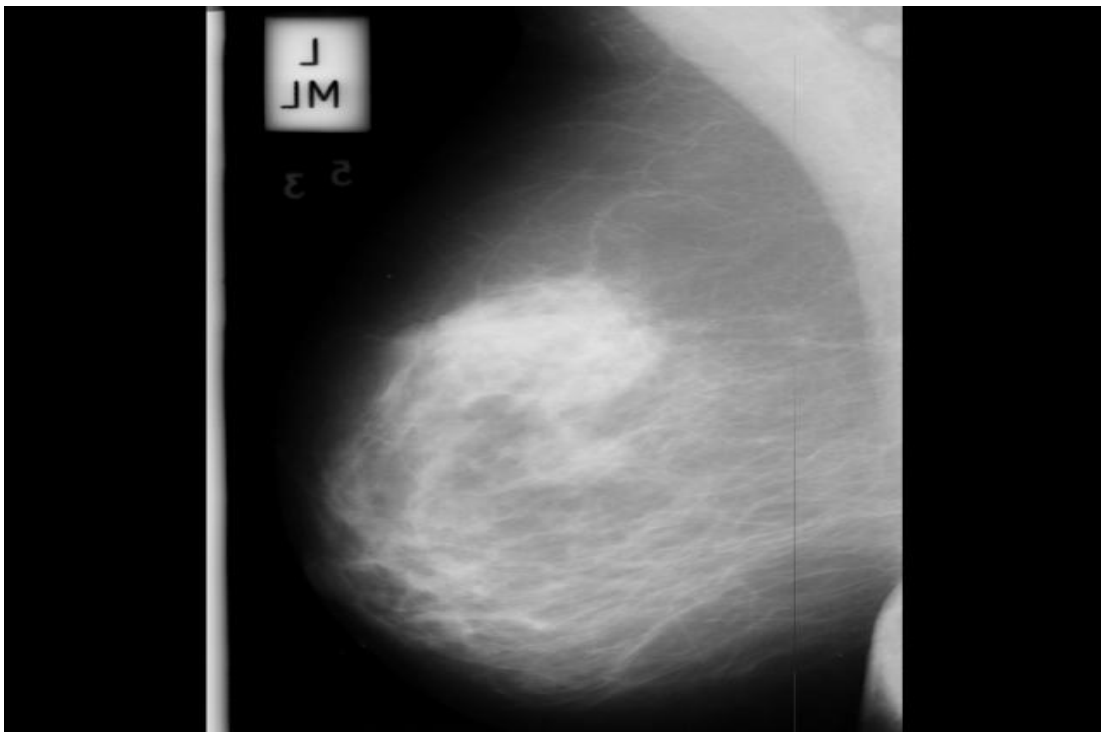

Figure 4.5 Display of a mammographic image which has an asymmetrical abnormality (mdb111 D ASYM M 505 575 107).

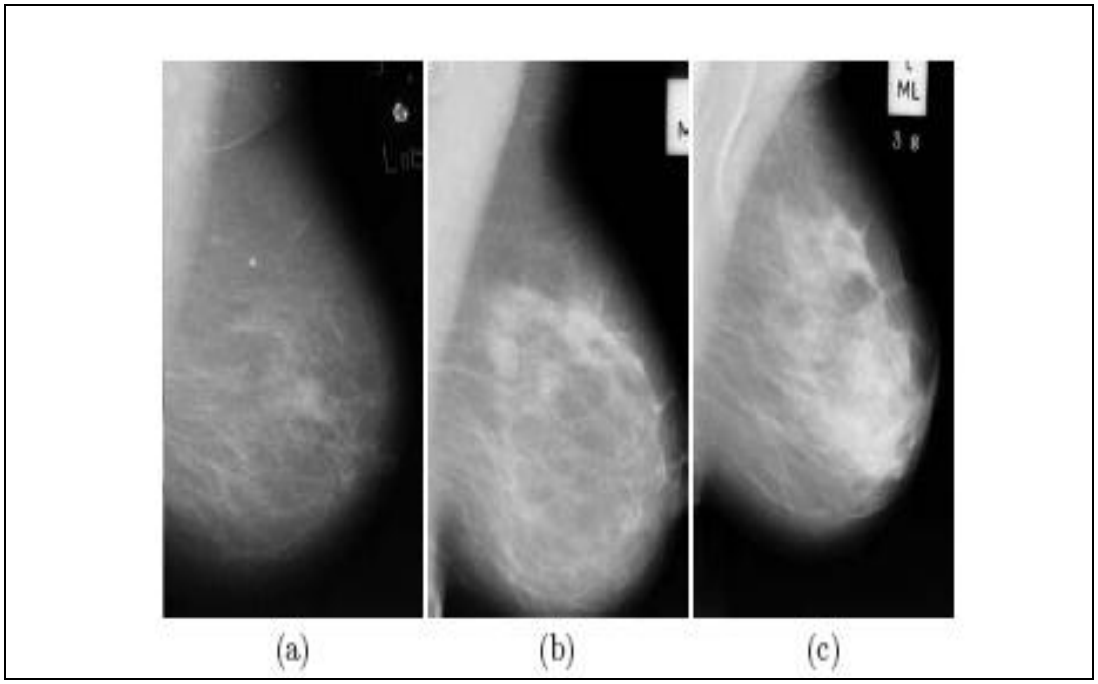

Figure 4.6 Examples of different types of breast tissue in the MIAS database . (a) fatty, (b) glandular, and (c) dense.

On figure 4.7, user can apply DbScan algorithm on the dataset by using DbScan buton. There are two parameters which name are Eps and MinPts. For example when user choose Eps=40 and Minpts=8 parameter values, five circle will be occured on the panel. This means that the dataset were divided into 5 clusters for these parameter values. When user clicked on the Output.txt buton, datas which are clustered can be written on a text file under the Visual Studio2010->Projects-DbScan-bin-Debug-output.txt and there will be detailed information about clusters which are occured. On the other hand, user can see these information by using the richtextbox at below on the panel.

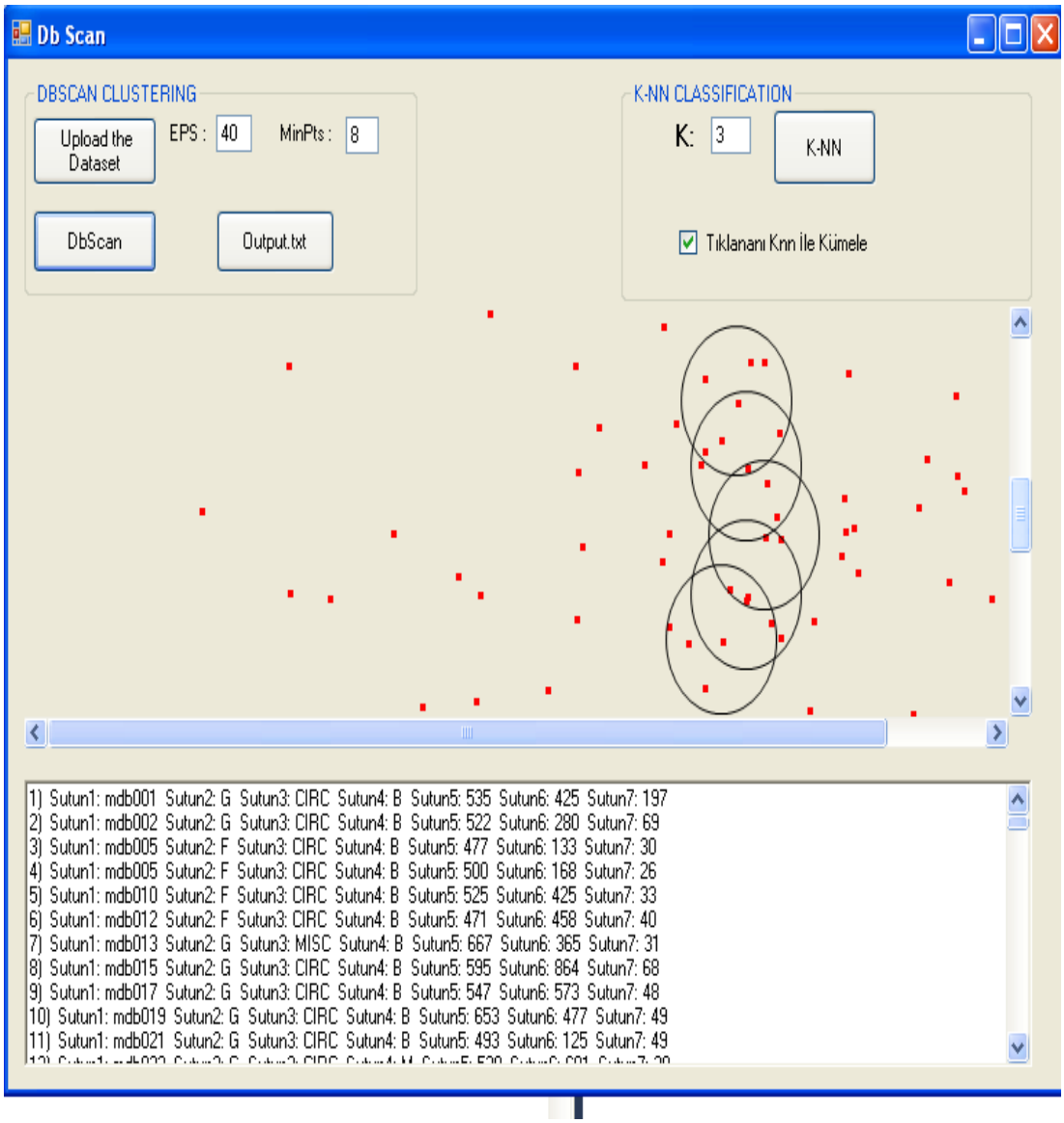

Figure 4.7 The page of applying DbScan algorithm

Output.txt File for Eps:40 and Minpts:8 Values;

Five cluster occured after applying DbScan algorithm for Eps:40 and Minpts:8 values. There are 8 point in every clusters.

• Eps: Maximum radius of the neighborhood

• MinPts: Minimum number of points in an Eps-neighborhood of that point

Table 4.1 Content of the output.txt file for Eps:40 and Minpts:8 values

1.Grup: 36_mdb111 İçindekiler(8) ->

Sutun1: 25_mdb090 Sutun2: G Sutun3: ASYM Sutun4: M Sutun5: 510 Sutun6: 547 Sutun7: 49

Sutun1: 36_mdb111 Sutun2: D Sutun3: ASYM Sutun4: M Sutun5: 505 Sutun6: 575 Sutun7: 107

Sutun1: 38_mdb117 Sutun2: G Sutun3: ARCH Sutun4: M Sutun5: 480 Sutun6: 576 Sutun7: 84

Sutun1: 44_mdb127 Sutun2: G Sutun3: ARCH Sutun4: B Sutun5: 523 Sutun6: 551 Sutun7: 48

Sutun1: 58_mdb158 Sutun2: F Sutun3: ARCH Sutun4: M Sutun5: 540 Sutun6: 565 Sutun7: 88

Sutun1: 66_mdb178 Sutun2: G Sutun3: SPIC Sutun4: M Sutun5: 492 Sutun6: 600 Sutun7: 70

Sutun1: 96_mdb238 Sutun2: F Sutun3: CALC Sutun4: M Sutun5: 522 Sutun6: 553 Sutun7: 17

Sutun1: 101_mdb244 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 466 Sutun6: 567 Sutun7: 52

Kesiştiği Gruplar : 44_mdb127 İle Kesişiyor

2.Grup:44_mdb127 İçindekiler(8)->

Sutun1: 9_mdb017 Sutun2: G Sutun3: CIRC Sutun4: B Sutun5: 547 Sutun6: 573 Sutun7: 48

Sutun1: 25_mdb090 Sutun2: G Sutun3: ASYM Sutun4: M Sutun5: 510 Sutun6: 547 Sutun7: 49

Sutun1: 36_mdb111 Sutun2: D Sutun3: ASYM Sutun4: M Sutun5: 505 Sutun6: 575 Sutun7: 107

Sutun1: 44_mdb127 Sutun2: G Sutun3: ARCH Sutun4: B Sutun5: 523 Sutun6: 551 Sutun7: 48

Sutun1: 58_mdb158 Sutun2: F Sutun3: ARCH Sutun4: M Sutun5: 540 Sutun6: 565 Sutun7: 88

Sutun1: 59_mdb160 Sutun2: F Sutun3: ARCH Sutun4: B Sutun5: 536 Sutun6:

Table 4.1 (continue)

519 Sutun7: 61

Sutun1: 84_mdb213 Sutun2: G Sutun3: CALC Sutun4: M Sutun5: 547 Sutun6: 520 Sutun7: 45

Sutun1: 96_mdb238 Sutun2: F Sutun3: CALC Sutun4: M Sutun5: 522 Sutun6: 553 Sutun7: 17

Kesiştiği Gruplar : 36_mdb111 İle Kesişiyor 59_mdb160 İle Kesişiyor

---

3.Grup:59_mdb160 İçindekiler(8)->

Sutun1: 25_mdb090 Sutun2: G Sutun3: ASYM Sutun4: M Sutun5: 510 Sutun6: 547 Sutun7: 49

Sutun1: 44_mdb127 Sutun2: G Sutun3: ARCH Sutun4: B Sutun5: 523 Sutun6: 551 Sutun7: 48

Sutun1: 59_mdb160 Sutun2: F Sutun3: ARCH Sutun4: B Sutun5: 536 Sutun6: 519 Sutun7: 61

Sutun1: 61_mdb165 Sutun2: D Sutun3: ARCH Sutun4: B Sutun5: 537 Sutun6: 490 Sutun7: 42

Sutun1: 84_mdb213 Sutun2: G Sutun3: CALC Sutun4: M Sutun5: 547 Sutun6: 520 Sutun7: 45

Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6: 482 Sutun7: 29

Sutun1: 96_mdb238 Sutun2: F Sutun3: CALC Sutun4: M Sutun5: 522 Sutun6: 553 Sutun7: 17

Sutun1: 103_mdb249 Sutun2: D Sutun3: CALC Sutun4: M Sutun5: 544 Sutun6: 508 Sutun7: 48

Kesiştiği Gruplar : 44_mdb127 İle Kesişiyor 88_mdb223 İle Kesişiyor .

4.Grup:88_mdb223 İçindekiler(9)->

Sutun1: 18_mdb063 Sutun2: D Sutun3: MISC Sutun4: B Sutun5: 546 Sutun6: 463 Sutun7: 33

Sutun1: 23_mdb081 Sutun2: G Sutun3: ASYM Sutun4: B Sutun5: 492 Sutun6: 473 Sutun7: 131

Table 4.1 (continue)

Sutun1: 59_mdb160 Sutun2: F Sutun3: ARCH Sutun4: B Sutun5: 536 Sutun6: 519 Sutun7: 61

Sutun1: 61_mdb165 Sutun2: D Sutun3: ARCH Sutun4: B Sutun5: 537 Sutun6: 490 Sutun7: 42

Sutun1: 63_mdb170 Sutun2: D Sutun3: ARCH Sutun4: M Sutun5: 489 Sutun6: 480 Sutun7: 82

Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6: 482 Sutun7: 29

Sutun1: 93_mdb227 Sutun2: G Sutun3: CALC Sutun4: B Sutun5: 504 Sutun6: 467 Sutun7: 9

Sutun1: 103_mdb249 Sutun2: D Sutun3: CALC Sutun4: M Sutun5: 544 Sutun6: 508 Sutun7: 48

Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93

Kesiştiği Gruplar : 59_mdb160 İle Kesişiyor 117_mdb315 İle Kesişiyor

---

5.Grup:117_mdb315 İçindekiler(8)->

Sutun1: 1_mdb001 Sutun2: G Sutun3: CIRC Sutun4: B Sutun5: 535 Sutun6: 425 Sutun7: 197

Sutun1: 5_mdb010 Sutun2: F Sutun3: CIRC Sutun4: B Sutun5: 525 Sutun6: 425 Sutun7: 33

Sutun1: 18_mdb063 Sutun2: D Sutun3: MISC Sutun4: B Sutun5: 546 Sutun6: 463 Sutun7: 33

Sutun1: 23_mdb081 Sutun2: G Sutun3: ASYM Sutun4: B Sutun5: 492 Sutun6: 473 Sutun7: 131

Sutun1: 40_mdb121 Sutun2: G Sutun3: ARCH Sutun4: B Sutun5: 492 Sutun6: 434 Sutun7: 87

Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6: 482 Sutun7: 29

Sutun1: 93_mdb227 Sutun2: G Sutun3: CALC Sutun4: B Sutun5: 504 Sutun6: 467 Sutun7: 9

Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93

Kesiştiği Gruplar : 88_mdb223 İle Kesişiyor

To learn that how the number of clusters will be changed while the value of Eps is increasing and Minpts is constant (Minpts=4), we run the program for different Eps values and the results showed that number of clusters were increased. On table 4.2, we can see the results for different Eps values.

Table 4.2 Number of clusters for Parameter values that MinPts= 4 and Eps is increasing.

EPS	MINPTS	CLUSTER
4	4	0
8	4	0
10	4	0
16	4	0
19	4	3
20	4	4
25	4	10
32	4	27
50	4	58
100	4	104

To learn that how the number of clusters will be changed while the value of Minpts is increasing and Eps is constant (Eps=40), we run the program for different Minpts values and the results showed that the number of clusters were dicreased. On table 4.3, we can see the results for different Minpts values.

Table 4.3 Number of clusters for Parameter values that Eps= 40 and Minpts is increasing.

EPS	MINPTS	CLUSTER
40	4	42
40	5	27
40	6	14
40	7	9
40	8	5
40	9	1
40	10	0
40	15	0
40	20	0
40	40	0

On table 4.4, we can see the results for Minpts= Eps values. When MinPts is equal to Eps, the number of clusters are zero.

Table 4.4 Cluster parameter values when Eps=MinPts

EPS	MINPTS	CLUSTER
10	10	0
15	15	0
20	20	0
26	26	0
34	34	0
42	42	0
50	50	0
52	52	0
64	64	0
100	100	0

On figure 4.8, user can apply K-NN Algorithm on the dataset by using K-NN buton. There is a parameter which name is K. For example when user choose K=3 parameter value and click on the noise point which MIAS database reference number

is mdb226, this data will be included in clusters which is accordance with its characteristics with the help of K-NN classification algorithm.

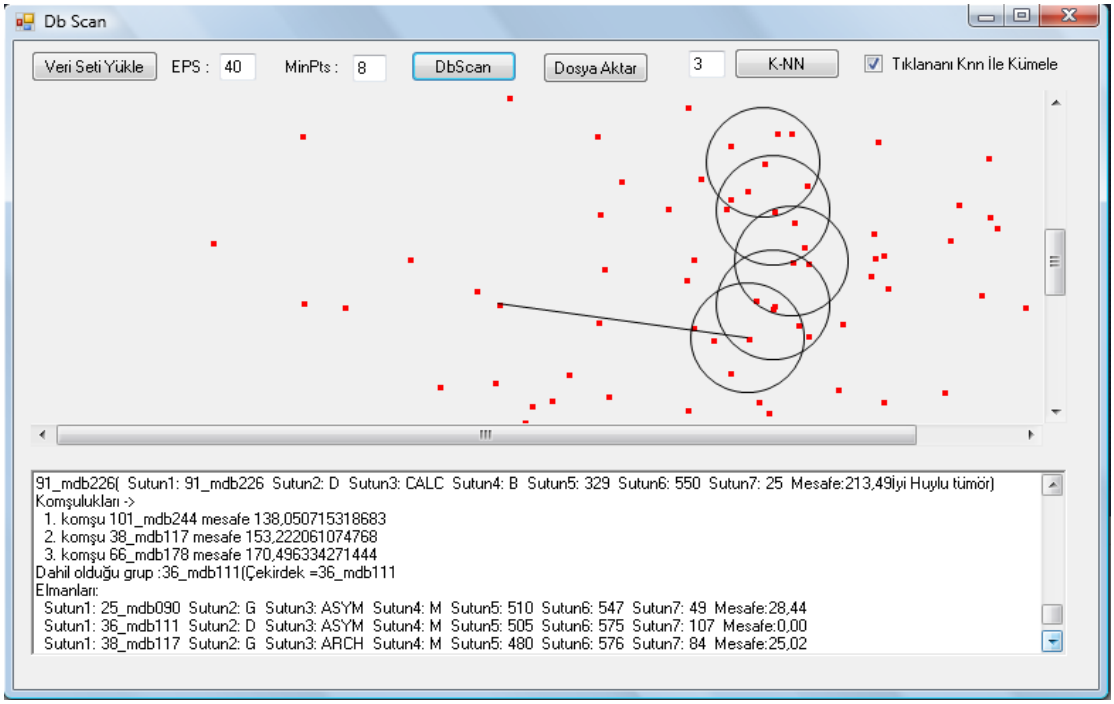

Figure 4.8 The page of applying K-NN algorithm for K=3

--------For K=3 and Test Data = mdb226------

Table 4.5 Testing result for K=3 and test data= mdb226

91_**mdb226**(Sutun1: 91_mdb226 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 329 Sutun6: 550 Sutun7: 25 Mesafe:213,49 İyi Huylu tümör)

Komşulukları ->
1. komşu 101_mdb244 mesafe 138,050715318683
2. komşu 38_mdb117 mesafe 153,222061074768
3. komşu 66_mdb178 mesafe 170,496334271444

Dahil olduğu grup :36_mdb111**(Çekirdek =36_mdb111)**

Elmanları:

Sutun1: 25_mdb090 Sutun2: G Sutun3: ASYM Sutun4: M Sutun5: 510 Sutun6: 547 Sutun7: 49 Mesafe:28,44

Sutun1: 36_mdb111 Sutun2: D Sutun3: ASYM Sutun4: M Sutun5: 505 Sutun6: 575 Sutun7: 107 Mesafe:0,00

Table 4.5 (continue)

Sutun1: 38_mdb117 Sutun2: G Sutun3: ARCH Sutun4: M Sutun5: 480 Sutun6: 576 Sutun7: 84 Mesafe:25,02

Sutun1: 44_mdb127 Sutun2: G Sutun3: ARCH Sutun4: B Sutun5: 523 Sutun6: 551 Sutun7: 48 Mesafe:30,00

Sutun1: 58_mdb158 Sutun2: F Sutun3: ARCH Sutun4: M Sutun5: 540 Sutun6: 565 Sutun7: 88 Mesafe:36,40

Sutun1: 66_mdb178 Sutun2: G Sutun3: SPIC Sutun4: M Sutun5: 492 Sutun6: 600 Sutun7: 70 Mesafe:28,18

Sutun1: 96_mdb238 Sutun2: F Sutun3: CALC Sutun4: M Sutun5: 522 Sutun6: 553 Sutun7: 17 Mesafe:27,80

Sutun1: 101_mdb244 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 466 Sutun6: 567 Sutun7: 52 Mesafe:39,81
Grupta 2 adet iyi huylu tümör var, 6 adet kotu huylu tümör var

On figure 4.9, when user choose K=5 parameter value and click on the noise point which MIAS database reference number is mdb226, this data will be included in clusters which is accordance with its characteristics with the help of K-NN classification algorithm.

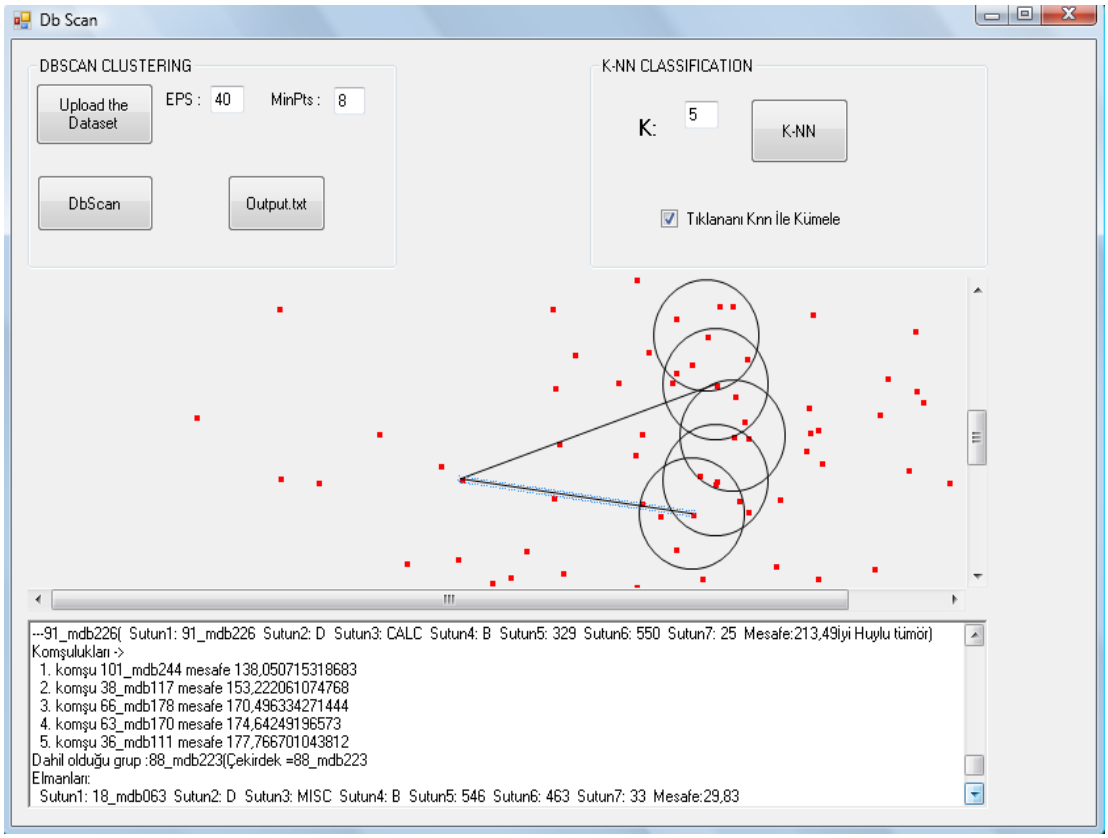

Figure 4.9 The page of applying K-NN algorithm for K=5

-----For K=5 and Test Data = mdb226-----

Table 4.6 Testing result for K=5 and test data= mdb226

---91_**mdb226**(Sutun1: 91_mdb226 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 329 Sutun6: 550 Sutun7: 25 Mesafe:213,49İyi Huylu tümör)

Komşulukları ->
1. komşu 101_mdb244 mesafe 138,050715318683
2. komşu 38_mdb117 mesafe 153,222061074768
3. komşu 66_mdb178 mesafe 170,496334271444
4. komşu 63_mdb170 mesafe 174,64249196573
5. komşu 36_mdb111 mesafe 177,766701043812

Dahil olduğu grup :88_mdb223(**Çekirdek =88_mdb223**)

Elmanları:

Sutun1: 18_mdb063 Sutun2: D Sutun3: MISC Sutun4: B Sutun5: 546 Sutun6: 463 Sutun7: 33 Mesafe:29,83

Sutun1: 23_mdb081 Sutun2: G Sutun3: ASYM Sutun4: B Sutun5: 492 Sutun6: 473 Sutun7: 131 Mesafe:32,28

Table 4.6 (continue)

Sutun1: 59_mdb160 Sutun2: F Sutun3: ARCH Sutun4: B Sutun5: 536 Sutun6: 519 Sutun7: 61 Mesafe:39,22

Sutun1: 61_mdb165 Sutun2: D Sutun3: ARCH Sutun4: B Sutun5: 537 Sutun6: 490 Sutun7: 42 Mesafe:16,12

Sutun1: 63_**mdb170** Sutun2: D Sutun3: ARCH Sutun4: M Sutun5: 489 Sutun6: 480 Sutun7: 82 Mesafe:34,06

Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6: 482 Sutun7: 29 Mesafe:0,00

Sutun1: 93_mdb227 Sutun2: G Sutun3: CALC Sutun4: B Sutun5: 504 Sutun6: 467 Sutun7: 9 Mesafe:24,21

Sutun1: 103_mdb249 Sutun2: D Sutun3: CALC Sutun4: M Sutun5: 544 Sutun6: 508 Sutun7: 48 Mesafe:33,42

Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93 Mesafe:35,69

Grupta 7 adet iyi huylu tümör var, 2 adet kotu huylu tümör var

On figure 4.10, when user choose K=3 parameter value and click on the K-NN buton, all datas in our database will be included in clusters which is accordance with their characteristics with the help of K-NN classification algorithm.

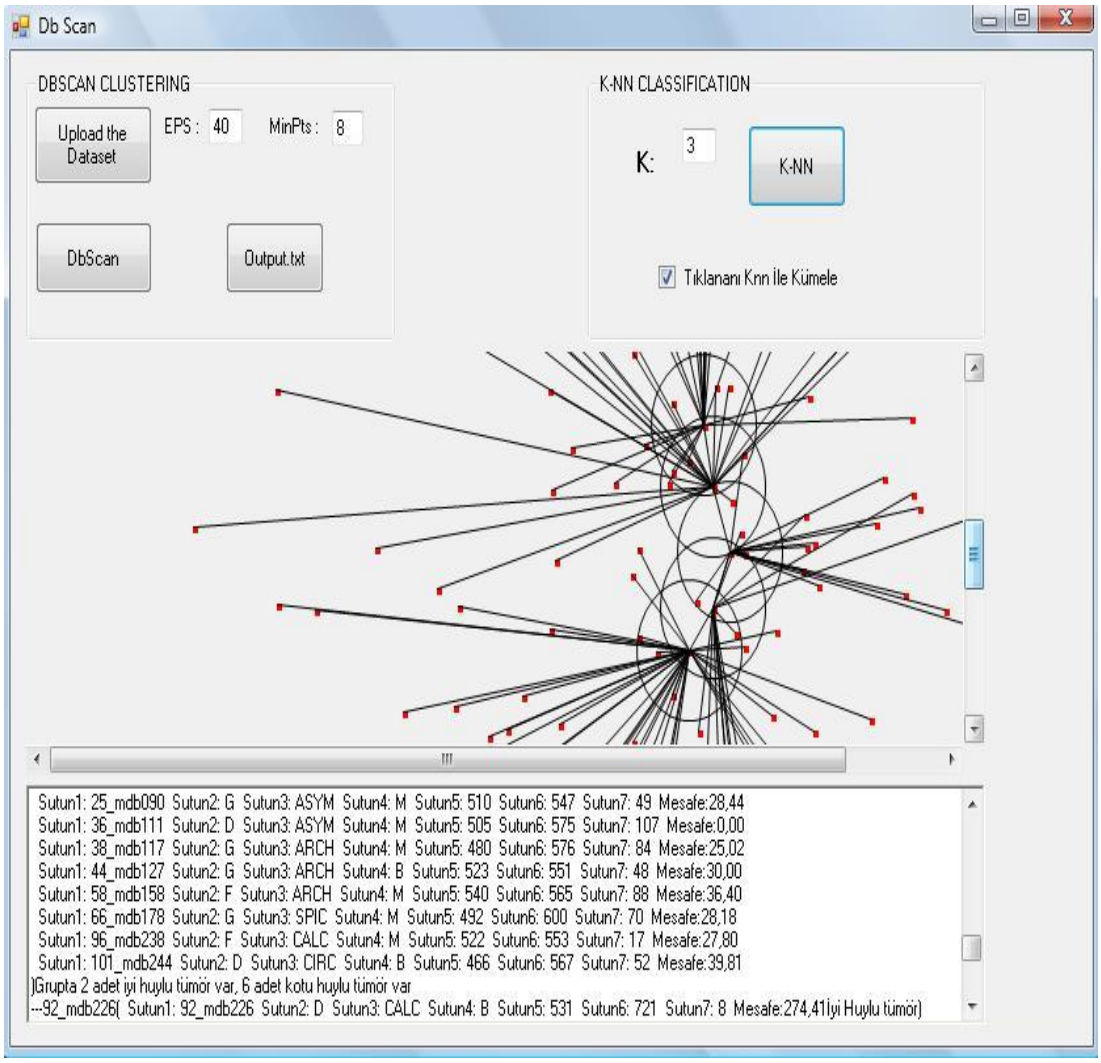

Figure 4.10 The page of applying K-NN algorithm for K=3

On table 4.7, you can see that testing results for all datas in our database which are included in clusters in accordance with their characteristics with the help of K-NN classification algorithm. But there were a lot of record, thatswhy we couldn’tshow all of them on the table.

-----For K=3 and Test Data= All Noise Datas-----

Table 4.7 Testing result for K=3 and test data= all noise datas

1_**mdb001**(Sutun1: 1_mdb001 Sutun2: G Sutun3: CIRC Sutun4: B Sutun5: 535 Sutun6: 425 Sutun7: 197 Mesafe:29,07İyi Huylu tümör) Komşulukları ->

1. komşu 1_mdb001 mesafe 0
2. komşu 5_mdb010 mesafe 10
3. komşu 117_mdb315 mesafe 29,0688837074973

Dahil olduğu grup :88_**mdb223** Çekirdek =88_mdb223

Elmanları:

Sutun1: 18_mdb063 Sutun2: D Sutun3: MISC Sutun4: B Sutun5: 546 Sutun6: 463 Sutun7: 33 Mesafe:29,83

Sutun1: 23_mdb081 Sutun2: G Sutun3: ASYM Sutun4: B Sutun5: 492 Sutun6: 473 Sutun7: 131 Mesafe:32,28

Sutun1: 59_mdb160 Sutun2: F Sutun3: ARCH Sutun4: B Sutun5: 536 Sutun6: 519 Sutun7: 61 Mesafe:39,22

Sutun1: 61_mdb165 Sutun2: D Sutun3: ARCH Sutun4: B Sutun5: 537 Sutun6: 490 Sutun7: 42 Mesafe:16,12

Sutun1: 63_mdb170 Sutun2: D Sutun3: ARCH Sutun4: M Sutun5: 489 Sutun6: 480 Sutun7: 82 Mesafe:34,06

Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6: 482 Sutun7: 29 Mesafe:0,00

Sutun1: 93_mdb227 Sutun2: G Sutun3: CALC Sutun4: B Sutun5: 504 Sutun6: 467 Sutun7: 9 Mesafe:24,21

Sutun1: 103_mdb249 Sutun2: D Sutun3: CALC Sutun4: M Sutun5: 544 Sutun6: 508 Sutun7: 48 Mesafe:33,42

Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93 Mesafe:35,69

Grupta 7 adet iyi huylu tümör var, 2 adet kotu huylu tümör var

Table 4.7 (continue)

---2_**mdb002**(Sutun1: 2_mdb002 Sutun2: G Sutun3: CIRC Sutun4: B Sutun5: 522 Sutun6: 280 Sutun7: 69 Mesafe:167,11İyi Huylu tümör) Komşulukları ->

1. komşu 5_mdb010 mesafe 145,0310311623
2. komşu 1_mdb001 mesafe 145,581592242976
3. komşu 40_mdb121 mesafe 156,894869259641

Dahil olduğu grup :117_**mdb315**(Çekirdek =117_mdb315)

Elmanları:

Sutun1: 1_mdb001 Sutun2: G Sutun3: CIRC Sutun4: B Sutun5: 535 Sutun6: 425 Sutun7: 197 Mesafe:29,07

Sutun1: 5_mdb010 Sutun2: F Sutun3: CIRC Sutun4: B Sutun5: 525 Sutun6: 425 Sutun7: 33 Mesafe:23,77

Sutun1: 18_mdb063 Sutun2: D Sutun3: MISC Sutun4: B Sutun5: 546 Sutun6: 463 Sutun7: 33 Mesafe:34,00

Sutun1: 23_mdb081 Sutun2: G Sutun3: ASYM Sutun4: B Sutun5: 492 Sutun6: 473 Sutun7: 131 Mesafe:35,38

Sutun1: 40_mdb121 Sutun2: G Sutun3: ARCH Sutun4: B Sutun5: 492 Sutun6: 434 Sutun7: 87 Mesafe:27,29

Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6: 482 Sutun7: 29 Mesafe:35,69

Sutun1: 93_mdb227 Sutun2: G Sutun3: CALC Sutun4: B Sutun5: 504 Sutun6: 467 Sutun7: 9 Mesafe:23,32

Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93 Mesafe:0,00

Grupta 8 adet iyi huylu tümör var, 0 adet kotu huylu tümör var

---3_**mdb005**(Sutun1: 3_mdb005 Sutun2: F Sutun3: CIRC Sutun4: B Sutun5: 477 Sutun6: 133 Sutun7: 30 Mesafe:316,41İyi Huylu tümör) Komşulukları ->

1. komşu 5_mdb010 mesafe 295,918907810907
2. komşu 1_mdb001 mesafe 297,704551527181
3. komşu 40_mdb121 mesafe 301,373522393723

Table 4.7 (continue)

Dahil olduğu grup :117_**mdb315**(Çekirdek =117_mdb315
Elmanları:
Sutun1: 1_mdb001 Sutun2: G Sutun3: CIRC Sutun4: B Sutun5: 535 Sutun6: 425 Sutun7: 197 Mesafe:29,07
Sutun1: 5_mdb010 Sutun2: F Sutun3: CIRC Sutun4: B Sutun5: 525 Sutun6: 425 Sutun7: 33 Mesafe:23,77
Sutun1: 18_mdb063 Sutun2: D Sutun3: MISC Sutun4: B Sutun5: 546 Sutun6: 463 Sutun7: 33 Mesafe:34,00
Sutun1: 23_mdb081 Sutun2: G Sutun3: ASYM Sutun4: B Sutun5: 492 Sutun6: 473 Sutun7: 131 Mesafe:35,38
Sutun1: 40_mdb121 Sutun2: G Sutun3: ARCH Sutun4: B Sutun5: 492 Sutun6: 434 Sutun7: 87 Mesafe:27,29
Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6: 482 Sutun7: 29 Mesafe:35,69
Sutun1: 93_mdb227 Sutun2: G Sutun3: CALC Sutun4: B Sutun5: 504 Sutun6: 467 Sutun7: 9 Mesafe:23,32
Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93 Mesafe:0,00
Grupta 8 adet iyi huylu tümör var, 0 adet kotu huylu tümör var

---4_**mdb005**(Sutun1: 4_mdb005 Sutun2: F Sutun3: CIRC Sutun4: B Sutun5: 500 Sutun6: 168 Sutun7: 26 Mesafe:279,46İyi Huylu tümör) Komşulukları ->
1. komşu 5_mdb010 mesafe 258,213090295593
2. komşu 1_mdb001 mesafe 259,372319263255
3. komşu 40_mdb121 mesafe 266,120273560659

Dahil olduğu grup :117_**mdb315** Çekirdek =117_mdb315
Elmanları:
Sutun1: 1_mdb001 Sutun2: G Sutun3: CIRC Sutun4: B Sutun5: 535 Sutun6: 425 Sutun7: 197 Mesafe:29,07
Sutun1: 5_mdb010 Sutun2: F Sutun3: CIRC Sutun4: B Sutun5: 525 Sutun6: 425

Table 4.7 (continue)

Sutun7: 33 Mesafe:23,77

Sutun1: 18_mdb063 Sutun2: D Sutun3: MISC Sutun4: B Sutun5: 546 Sutun6: 463 Sutun7: 33 Mesafe:34,00

Sutun1: 23_mdb081 Sutun2: G Sutun3: ASYM Sutun4: B Sutun5: 492 Sutun6: 473 Sutun7: 131 Mesafe:35,38

Sutun1: 40_mdb121 Sutun2: G Sutun3: ARCH Sutun4: B Sutun5: 492 Sutun6: 434 Sutun7: 87 Mesafe:27,29

Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6: 482 Sutun7: 29 Mesafe:35,69

Sutun1: 93_mdb227 Sutun2: G Sutun3: CALC Sutun4: B Sutun5: 504 Sutun6: 467 Sutun7: 9 Mesafe:23,32

Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93 Mesafe:0,00

Grupta 8 adet iyi huylu tümör var, 0 adet kotu huylu tümör var

---5_**mdb010**(Sutun1: 5_mdb010 Sutun2: F Sutun3: CIRC Sutun4: B Sutun5: 525 Sutun6: 425 Sutun7: 33 Mesafe:23,77İyi Huylu tümör) Komşulukları ->

1. komşu 5_mdb010 mesafe 0
2. komşu 1_mdb001 mesafe 10
3. komşu 117_mdb315 mesafe 23,7697286480094

Dahil olduğu grup :88_**mdb223** Çekirdek =88_mdb223

Elmanları:

Sutun1: 18_mdb063 Sutun2: D Sutun3: MISC Sutun4: B Sutun5: 546 Sutun6: 463 Sutun7: 33 Mesafe:29,83

Sutun1: 23_mdb081 Sutun2: G Sutun3: ASYM Sutun4: B Sutun5: 492 Sutun6: 473 Sutun7: 131 Mesafe:32,28

Sutun1: 59_mdb160 Sutun2: F Sutun3: ARCH Sutun4: B Sutun5: 536 Sutun6: 519 Sutun7: 61 Mesafe:39,22

Sutun1: 61_mdb165 Sutun2: D Sutun3: ARCH Sutun4: B Sutun5: 537 Sutun6: 490 Sutun7: 42 Mesafe:16,12

Sutun1: 63_mdb170 Sutun2: D Sutun3: ARCH Sutun4: M Sutun5: 489 Sutun6:

Table 4.7 (continue)

480 Sutun7: 82 Mesafe:34,06
Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6: 482 Sutun7: 29 Mesafe:0,00
Sutun1: 93_mdb227 Sutun2: G Sutun3: CALC Sutun4: B Sutun5: 504 Sutun6: 467 Sutun7: 9 Mesafe:24,21
Sutun1: 103_mdb249 Sutun2: D Sutun3: CALC Sutun4: M Sutun5: 544 Sutun6: 508 Sutun7: 48 Mesafe:33,42
Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93 Mesafe:35,69
Grupta 7 adet iyi huylu tümör var, 2 adet kotu huylu tümör var

---6_**mdb012**(Sutun1: 6_mdb012 Sutun2: F Sutun3: CIRC Sutun4: B Sutun5: 471 Sutun6: 458 Sutun7: 40 Mesafe:46,32İyi Huylu tümör) Komşulukları ->
1. komşu 23_mdb081 mesafe 25,8069758011279
2. komşu 63_mdb170 mesafe 28,4253408071038
3. komşu 40_mdb121 mesafe 31,890437438204

Dahil olduğu grup :117_**mdb315** Çekirdek =117_mdb315
Elmanları:
Sutun1: 1_mdb001 Sutun2: G Sutun3: CIRC Sutun4: B Sutun5: 535 Sutun6: 425 Sutun7: 197 Mesafe:29,07
Sutun1: 5_mdb010 Sutun2: F Sutun3: CIRC Sutun4: B Sutun5: 525 Sutun6: 425 Sutun7: 33 Mesafe:23,77
Sutun1: 18_mdb063 Sutun2: D Sutun3: MISC Sutun4: B Sutun5: 546 Sutun6: 463 Sutun7: 33 Mesafe:34,00
Sutun1: 23_mdb081 Sutun2: G Sutun3: ASYM Sutun4: B Sutun5: 492 Sutun6: 473 Sutun7: 131 Mesafe:35,38
Sutun1: 40_mdb121 Sutun2: G Sutun3: ARCH Sutun4: B Sutun5: 492 Sutun6: 434 Sutun7: 87 Mesafe:27,29
Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6: 482 Sutun7: 29 Mesafe:35,69
Sutun1: 93_mdb227 Sutun2: G Sutun3: CALC Sutun4: B Sutun5: 504 Sutun6:

Table 4.7 (continue)

467 Sutun7: 9 Mesafe:23,32

Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93 Mesafe:0,00

Grupta 8 adet iyi huylu tümör var, 0 adet kotu huylu tümör var

.
.
.
.
.
.
.
.
.
.
.
.

---116_**mdb314**(Sutun1: 116_mdb314 Sutun2: F Sutun3: MISC Sutun4: B Sutun5: 518 Sutun6: 191 Sutun7: 39 Mesafe:256,01İyi Huylu tümör) Komşulukları ->

1. komşu 5_mdb010 mesafe 234,104677441524
2. komşu 1_mdb001 mesafe 234,616708697399
3. komşu 40_mdb121 mesafe 244,386988196999

Dahil olduğu grup :117_**mdb315** Çekirdek =117_mdb315

Elmanları:

Sutun1: 1_mdb001 Sutun2: G Sutun3: CIRC Sutun4: B Sutun5: 535 Sutun6: 425 Sutun7: 197 Mesafe:29,07

Sutun1: 5_mdb010 Sutun2: F Sutun3: CIRC Sutun4: B Sutun5: 525 Sutun6: 425 Sutun7: 33 Mesafe:23,77

Sutun1: 18_mdb063 Sutun2: D Sutun3: MISC Sutun4: B Sutun5: 546 Sutun6: 463 Sutun7: 33 Mesafe:34,00

Table 4.7 (continue)

Sutun1: 23_mdb081 Sutun2: G Sutun3: ASYM Sutun4: B Sutun5: 492 Sutun6: 473 Sutun7: 131 Mesafe:35,38

Sutun1: 40_mdb121 Sutun2: G Sutun3: ARCH Sutun4: B Sutun5: 492 Sutun6: 434 Sutun7: 87 Mesafe:27,29

Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6: 482 Sutun7: 29 Mesafe:35,69

Sutun1: 93_mdb227 Sutun2: G Sutun3: CALC Sutun4: B Sutun5: 504 Sutun6: 467 Sutun7: 9 Mesafe:23,32

Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93 Mesafe:0,00

Grupta 8 adet iyi huylu tümör var, 0 adet kotu huylu tümör var

---117_**mdb315**(Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93 Mesafe:0,00İyi Huylu tümör) Komşulukları ->

1. komşu 117_mdb315 mesafe 0
2. komşu 93_mdb227 mesafe 23,3238075793812
3. komşu 5_mdb010 mesafe 23,7697286480094

Dahil olduğu grup :117_**mdb315**(Çekirdek =117_mdb315

Elmanları:

Sutun1: 1_mdb001 Sutun2: G Sutun3: CIRC Sutun4: B Sutun5: 535 Sutun6: 425 Sutun7: 197 Mesafe:29,07

Sutun1: 5_mdb010 Sutun2: F Sutun3: CIRC Sutun4: B Sutun5: 525 Sutun6: 425 Sutun7: 33 Mesafe:23,77

Sutun1: 18_mdb063 Sutun2: D Sutun3: MISC Sutun4: B Sutun5: 546 Sutun6: 463 Sutun7: 33 Mesafe:34,00

Sutun1: 23_mdb081 Sutun2: G Sutun3: ASYM Sutun4: B Sutun5: 492 Sutun6: 473 Sutun7: 131 Mesafe:35,38

Sutun1: 40_mdb121 Sutun2: G Sutun3: ARCH Sutun4: B Sutun5: 492 Sutun6: 434 Sutun7: 87 Mesafe:27,29

Sutun1: 88_mdb223 Sutun2: D Sutun3: CALC Sutun4: B Sutun5: 523 Sutun6:

Table 4.7 (continue)

482 Sutun7: 29 Mesafe:35,69 Sutun1: 93_mdb227 Sutun2: G Sutun3: CALC Sutun4: B Sutun5: 504 Sutun6: 467 Sutun7: 9 Mesafe:23,32 Sutun1: 117_mdb315 Sutun2: D Sutun3: CIRC Sutun4: B Sutun5: 516 Sutun6: 447 Sutun7: 93 Mesafe:0,00)Grupta 8 adet iyi huylu tümör var, 0 adet kotu huylu tümör var

CHAPTER FIVE
CONCLUSION

Clustering and classification methods are widely used in data mining. There are many clustering and classification algorithms in the literature. It is a great importance to choose the appropriate clustering or classificaion technique for an application.

In the scope of this thesis, a system aimed at clustering the medical data by using a density based clustering algorithm and classifying by using a classification algorithm, is proposed. Our application described the implementation of these two algorithms: DBSCAN (Ester, 1996) and K-NN (Hastie & Tibshirani, 1996). So, clustering and classification algorithms were used together in the project, DBSCAN algorithm for clustering and K-NN algorithm for classification were implemented within the MIAS database.

After analyzing mammography dataset, we saw that x,y image coordinates of centre of abnormality were numerical and we decided to use this data knowledge. We need to use an algorithm that separates the data in the appropriate number of clusters by itself. We choosed a density-based algorithm, DBSCAN. On the other hand, to classification the noise points which can not be included in any cluster with DBSCAN algorithm, the most suitable classification method is K-NN classifier for our dataset.

Firstly, the data, x-y image coordinates of abnormality centre, divided into clusters. And then, noise points (which can not be included in any cluster with DBSCAN algorithm) were included in clusters with K-NN classification algorithm.

Intensity of a point depends on the value of Eps. So, if the radius (Eps) is kept large, density of each data set will be equal to the number of points. Similarly, if the radius is kept very small, intensity of each point will be itself (1). For low-dimensional data, there are techniques for finding the appropriate radius.

The results of the DBSCAN algorithm for different parameter values were presented on the table 4.1, table 4.2, table 4.3 and table 4.4. The results of the K-NN classifier for K=3 and K=5 parameter values and selecting test data were presented on the table 4.5, table 4.6 and table 4.7.

When parameter values were examined, the results showed that DBSCAN algorithm produced the best result for optimal parameter values, Eps=40 and MinPts=8 on MIAS database. On the other hand, K-NN algorithm produced the best result for the value of k = 5. If we choose small k nearest neighbor parameter value, we may get incorrect results. The accuracy of the k-NN algorithm can be severely degraded by the presence of noisy or irrelevant features.

While the value of Eps is increasing and Minpts is constant (Minpts=4), we run the program for different Eps values and the results showed that number of clusters were increased. On table 4.2, we can see the results for different Eps values. while the value of Minpts is increasing and Eps is constant (Eps=40), we run the program for different Minpts values and the results showed that the number of clusters were dicreased. On table 4.3, we can see the results for different Minpts values. On table 4.4, we can see the results for Minpts= Eps values. When MinPts is equal to Eps, the number of clusters are zero.

REFERENCES

A tutorial on clustering algorithm. (n.d). Retrieved August 8, 2011, from http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/

Birant, D., & Kut, A., (2007). *ST-DBSCAN: An Algorithm for Clustering Spatial-temporal data.* Data and Knowledge Engineering, pp. 208-221.

Birant, D. (2004). *Modeling and analyzing marine data using data mining techniques.* İzmir.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: From concept to implementation*. Upper Saddle River, NJ: Prentice Hall.

Cunningham, P. & Delany, S. J. (2007). K-nearest neighbour classifiers. *Technical Report UCD-CSI-2007-4.*

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*, in: Proceedings of Second International Conference on Knowledge Discovery and Data Mining, Portland, pp. 226-231.

Güneser, C. (2009). *Classification of wisconsin breast cancer database*, İzmir.

Han, J., & Kamber, M. (2000). *Data mining: Concepts and techniques,* Simon Fraser University.

Hand, D., Manila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge: MIT Press.

Han, J., & Kambe, M. (2006). *Data mining concepts and techniques* (2nd Edition).

Inmon, W. H. (1995). What is a Data Warehouse. *Prism Tech Topic*, *1*.

Kaymaz, E. D., (2007). *Yapay bağışıklık sistemi tabanlı k-nn sınıflandırma algoritması ile protein örüntülerinin hücredeki yerleşim yerlerinin belirlenmesi*. Fırat üniversitesi yüksek lisans tezi, Elazığ.

Kumar, K. A., & Rangan, C. P. (2007). *Privacy preserving DBSCAN algorithm for clustering***.**

Department of Computer Science and Engineering**,** Indian Institute of Technology – Madras**,** Chennai - 600036, India.

Kozma, L. (February 2, 2008). *K nearest neighbors algorithm (kNN).* Retrieved March 17, 2011, from http://www.lkozma.net/knn2.pdf.

Kırtulukoğlu, S. (2009). *Neuro – fuzzy classification of wisconsin breast cancer database*, İzmir.

Leung, K. M. (November 13, 2007*). K-nearest neighbor algorithm for classification* Retrieved July 12, 2011, from http://cis.poly.edu/~mleung/FRE7851/f07/k-NearestNeighbor.pdf.

Moreira, A., Santos, M. Y., & Carneiro, S. (2005). *Density-based clustering algorithms – DBSCAN and SNN*, University of Minho – Portugal.

Parimala, M., Lopez, D. & Senthilkumar, N.C. (2011). A survey on density based clustering algorithms for mining large spatial databases. *International Journal of Advanced Science and Technology*, *31*, 59.

Raczynski, L., Wozniak, K., Rubel, T., & Zaremba, K. (2010). Application of density based clustering to microarray data analysis. *Intl Journal of Electronics and Telecommunications*, *56*, pp. 281-286.

Rehman, M., & Mehdi, S. A. (n.d.). *Comparıson of densıty-based clusterıng algorithms*. Lahore College for Women University & University of Management and Technology Lahore, Pakistan.

Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., Ricketts, I., Stamatakis, E., Cerneaz, N., Kok, S., Taylor, P., Betal, D., & Savage, J. (1994). *The mammographic image analysis society digital mammogram database.* In: International Workshop on Digital Mammography.

Sarkar, M., & Leong, T. Y. (2000). *Application of K-nearest neighbors algorithm on breast cancer diagnosis problem,* Medical Computing Laboratory, Department of Computer Science, School of Computing, The National University of Singapore.

Saranya, N. N., & Hemalatha, M. (2011). Potential research into spatial cancer database by using data clustering techniques. *International Journal of Computer Science and Information Security (IJCSIS), 9.*

Soni, J. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications (0975 – 8887), 17.*

The mini-MIAS database of mammograms. (n.d.). Retrieved March 5, 2011, from http://peipa.essex.ac.uk/info/mias.html.

The Gartner Group. Retrieved March 12, 2011, from http://www.gartner.com.

Voulgaris, Z., & Magoulas G. D. (n.d.). *Extensions of the k nearest neighbour methods for classification problems*. Retrieved July 12, 2011, from http://www.dcs.bbk.ac.uk/~gmagoulas/IASTED-AIA08.pdf

Who, X. Q., (2001). *Data mining a promising research area.* Retrieved August 02, 2011, from http://digital.cs.usu.edu/~xqi/DataMining.html.

ABBREVIATIONS

DBSCAN	Density Based Spatial Clustering of Applications with Noise
K-NN	K- Nearest Neighbour
MIAS	Mammographic Image Analysis Society Digital Mammogram Database
KDD	Knowledge Discovery in Databases
OPTICS	Ordering Points To Identify the Clustering Structure
DENCLUE	Density-based Clustering
CLIQUE	Clustering In Quest
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CLARANS	Clustering Large Applicationss
CURE	Clustering Using Representatives
ROCK	Robust Clustering Using links
STING	Statistical Information Grid-Based method
SVM	Support Vector Machine
EM	Expectation Maximization
CRM	Customer Relationship Management
EKG	Electrocardiography