A Comprehensive and Experimental Survey on Medical Data Classification and Pattern Recognition

R. Subathra Devi
Research Scholar, PG and Research Department of Computer Science, Presidency College, Chennai, Tamil Nadu, India

ABSTRACT

This paper is proposed to compare and analyze various type of medical data classification and pattern recognition methods. Medical data classification methods majorly divided into three categories such as supervised, classification and also semi-supervised classification. Pattern recognition and data classifications are both overlapped domain for useful knowledge generation and prediction from training data. The field of medical diagnosis (or) clinical support system needs in intelligent data classification and pattern recognition algorithms for more accuracy in clinical decision making. Supervised classification contains many methods such as rule based classification, decision tree based classification, Bayesian classification, KNN probabilistic neural network, SVM and more, combination of supervised classification called as ensemble algorithm. These types of mixed algorithm provide more accuracy. Unsupervised classification called lazy learner (or) clustering for example automatic classification of unlabeled data. Unsupervised classification also contains some types such as K-means, deep learning methods, hierarchical clustering and more. In this paper we have to analyze various types of classification algorithms using sample medical record of upper abdomen diseases database. In this paper we have to analyze maximum of algorithms in experimental using same training data, this will used for various performance and accuracy analysis.

Keywords: Medical support system, clinical support system, medical data classification, supervised classification, unsupervised classification, rule-based classification, DCT, Bayesian classification, PNN artificial neural network adaptive classifier, K-NN, K-means, machine learning, svm, abdominal diseases.

I. INTRODUCTION

Medical data contains large volume of information in an unstructured format, data mining discovers insightful, important and good patterns which are descriptive, understandable and predictive from large amount of data[1].

Data mining includes important techniques such as association, clustering, classification and prediction[6][7]. Classification is also one of the most important techniques in mining process[10][11]. The challenge in knowledge discovery is constructing fast and accurate classifier for large data set [2]. Medical data mining is an trending technology in medical field that solves the most traditional problems, such as congestion long wait time and delayed patient use. Clinical support systems help doctors to make accurate diagnosis of most diseases. Most medical data sets are widely distributed and unclassified medical data also look like heterogeneous and huge size/volume[8][9]. These data need to be organized in a form which is classified and understandable. Advantages of using data mining techniques in medical domain is to
improve the accuracy of the output with large amount of data. Medical data mining has great potential for exploring useful patterns among medical data set. Knowledge generation and retrieval are performed using classification algorithms, data mining and artificial neural networks. In this paper we have to analyze various types of data classification algorithm in chapter 5, Introduction to pattern recognition in chapter 2, and comparison and analysis on chapter 7 and 8.

II. PATTERN RECOGNITION

Recognizing different types of objects in real world environment is a complex task for humans. To overcome this problem we have to implement artificial (or) computerizing methods[15]. Because in computers pattern recognition and machine learning not developed well. Pattern recognition methods provide solution for various problems such as bioinformatics, document analysis, industrial automation, image analysis, remote sensing, handwritten text analysis, medical diagnosis, speech recognition, IS and many more. Pattern recognition mostly involved in three steps one is extracting features from given information and second one is classifying extracted patterns using specific methods[17][18].

Third one is data acquisition. Data acquisition is the process of converting information from one form to system readable-digital form, for example computer systems can handle different types of data such as audio-speech, text-character, picture- image. Data acquisition is performed by different types of sensors, such as mic-audio (or)speech, cosensor, LDR-light sensor, scanner-image and more pattern recognition training performed by train data and system performance tested by test data.

III. CLASSIFICATION

It’s a process for updating (or) adding unlabeled data point to labeled classified data group. Unlabeled data recognized and organized to labeled classified data using various classification methods(9). Devising a procedure for classification in which exact classes are known in advanced termed as pattern recognition(or) supervised learning[23][25].

In un-supervised learning classification classes not known in advance. There are three standard classification techniques available, such as machine learning based classification, statistical based classification and neural network based classification. These types of classification algorithms further subdivided (or) contains both supervised and unsupervised methods[32][29].

IV. SAMPLE DATA

In this paper we have used upper abdomen diseases data set, this training and testing data set contains sample of 10 disease for training and 10 disease for testing (noisy (or) un classified data).

4.1 SAMPLE DATA- TRAINING SET

<table>
<thead>
<tr>
<th>DISEASEID</th>
<th>DISEASENAME</th>
<th>ADDDESCRIPTION</th>
<th>DISEASEID1</th>
<th>SYMPTOMNAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>GASTROINTESTINAL BLEEDING</td>
<td>GASTROINTESTINAL BLEEDING</td>
<td>1001</td>
<td>BLEEDING</td>
</tr>
<tr>
<td>1001</td>
<td>GASTROINTESTINAL BLEEDING</td>
<td>GASTROINTESTINAL BLEEDING</td>
<td>1001</td>
<td>RED COLORED VOMIT</td>
</tr>
<tr>
<td>1001</td>
<td>GASTROINTESTINAL BLEEDING</td>
<td>GASTROINTESTINAL BLEEDING</td>
<td>1001</td>
<td>COFFEE GROUNDS COLORED VOMIT</td>
</tr>
<tr>
<td>1002</td>
<td>FOOD</td>
<td>FOOD</td>
<td>1002</td>
<td>VOMITIN</td>
</tr>
<tr>
<td>Code</td>
<td>Term</td>
<td>Code</td>
<td>Term</td>
<td>Code</td>
</tr>
<tr>
<td>------</td>
<td>-----------------------------</td>
<td>------</td>
<td>-----------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1002</td>
<td>POISONING</td>
<td>1002</td>
<td>ANXIETY DISORDER</td>
<td>1005</td>
</tr>
<tr>
<td></td>
<td>FOOD POISONING</td>
<td></td>
<td>INTESTINAL LIEUS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1005</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1003</td>
<td>GASTROENTERITIS</td>
<td>1003</td>
<td>VOMITING</td>
<td>1005</td>
</tr>
<tr>
<td></td>
<td>STOMACH FLUE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1003</td>
<td>GASTROENTERITIS</td>
<td>1003</td>
<td>PAIN</td>
<td>1005</td>
</tr>
<tr>
<td></td>
<td>STOMACH FLUE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1003</td>
<td>GASTROENTERITIS</td>
<td>1003</td>
<td>BLOATING</td>
<td>1005</td>
</tr>
<tr>
<td></td>
<td>STOMACH FLUE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1003</td>
<td>GASTROENTERITIS</td>
<td>1003</td>
<td>DECREASED APPETITE</td>
<td>1005</td>
</tr>
<tr>
<td></td>
<td>STOMACH FLUE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1003</td>
<td>GASTROENTERITIS</td>
<td>1003</td>
<td>DIARRHEA</td>
<td>1006</td>
</tr>
<tr>
<td></td>
<td>STOMACH FLUE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1003</td>
<td>GASTROENTERITIS</td>
<td>1003</td>
<td>RED COLORED VOMIT</td>
<td>1006</td>
</tr>
<tr>
<td></td>
<td>STOMACH FLUE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1004</td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td>1004</td>
<td>CHILLS</td>
<td>1006</td>
</tr>
<tr>
<td></td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1004</td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td>1004</td>
<td>PAIN</td>
<td>1006</td>
</tr>
<tr>
<td></td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1004</td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td>1004</td>
<td>ANXIETY</td>
<td>1006</td>
</tr>
<tr>
<td></td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1004</td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td>1004</td>
<td>DIZZINESS</td>
<td>1006</td>
</tr>
<tr>
<td></td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1004</td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td>1004</td>
<td>VOMITING</td>
<td>1006</td>
</tr>
<tr>
<td></td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1004</td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td>1004</td>
<td>AGITATION</td>
<td>1006</td>
</tr>
<tr>
<td></td>
<td>GENERALIZED ANXIETY DISORDER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1007</td>
<td>NARCOTIC OPIATE</td>
<td></td>
<td>PAIN</td>
<td>1006</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
V. TYPES OF DATA CLASSIFICATION ALGORITHM

The data classification algorithms are broadly classified into two categories:

1) Supervised Classification

Supervised learning of data classification method works based on pre-defined static rules and validated using test data. Supervised classification is most suitable for simple problems[41]. Supervised classification algorithm is further sub divided into semi-supervised. This type of classification algorithm is most suitable for simple and moderate problem[36][35].

List of Supervised and Semi-supervised Classification Algorithm
1. Rule Based Classification
2. Decision Tree Classification
3. Bayesian Classification
4. Adaptive Classifier
5. Neural Network
6. K-NN
7. SVM
2) **Un-Supervised(or) Automatic Classification**  
Un-supervised classification is performed without the interaction or control of user. Un-supervised classification is widely used for most complex data analysis. This type of classification is performed in dynamic manner and creates feature extractions and patterns automatically [2][38][39].

**List of Un-Supervised Algorithm**

1. Partitioned Clustering  
2. Hierarchical Clustering  
3. Density Based Clustering.

**VI. EXISTING METHODS**

6.1 **RULE BASED CLASSIFICATION**  
Rule Based Classification performs based on simple if-else-end conditional model. Every traditional programming language provide conditional statement such as if-else. Syntax for rule based classification model is follows.[1]  
If (Condition) Then Statement 1 to n End  
This method is used to perform simple problems, easy to program, easy to understand by others and also it renders decent performance [15][1][14].

**Classification Example**  
If symptom0=BLEEDING Then  
weight=weight+1  
End If  
If symptom1=COFFEE GROUNDS COLORED VOMIT Then  
weight=weight+1  
End If  
If symptom2=RED COLORED VOMIT Then  
weight=weight+1  
End If  
If weight=countofsymptoms Then  
DiseaseName=GUSTROINTESTINAL BLEEDING  
End If  

***************END OF RULE***************

If symptom0=DIARRHEA Then  
weight=weight+1  
End If  
If symptom1=PAIN Then  
weight=weight+1  
End If  
If symptom2=VOMITING Then  
weight=weight+1  
End If  
If weight=countofsymptoms Then  
DiseaseName=FOOD POISONING  
End If  

***************END OF RULE***************

**TEST DATA OUTPUT**  
Disease Identified : GUSTROINTESTINAL BLEEDING  
Disease Identified : FOOD POISONING  
Disease Identified : GENERALIZED ANXIETY DISORDER  
Disease Identified : INTESTINAL LIEUS  
Disease Identified : IRRITABLE BOWEL SYNDROME  
Disease Identified : PEPTIC ULCER  
Disease Identified : IRON POISONING  
Total Input Records :10  
Total Success Count :7  
Success Percentage :70 %

6.2 **DECISION-TREE CLASSIFICATION**  
Decision Tree Classification works based on decision tree induction. It may similar to rule based classification, combined with tree, but have many advantages then rule based classification. Decision tree represented in graph structure and contains secondary leafs and nodes, the top most node is called root node.[1][14][47][46][44]

Information Gain for Class D  
\[ Info(D) = - \sum_{i=1}^{n} p_i \log_2(p_i) \]
Information Gain for Attribute A

\[ \text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j) \]

Information Gain

\[ \text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \]

### Classification Example:

**Decision Tree**

**TEST DATA OUTPUT**

Disease Identified: GASTROINTESTINAL BLEEDING
Disease Identified: FOOD POISONING
Disease Identified: GENERALIZED ANXIETY DISORDER
Disease Identified: INTESTINAL LIEMS
Disease Identified: IRRITABLE BOWEL SYNDROME
Disease Identified: PEPTIC ULCER
Disease Identified: IRON POISONING
Total Input Records: 10
Total Success Count: 7
Success Percentage: 70%

### 6.3 BAYESIAN CLASSIFICATION

Bayesian classification works based on Bayes theorem, probability and class frequency. It is better than DT Classification and Rule based classification. [1][2][63][86][87]

Bayes theorem as follows

\[ P(H/X) = \frac{P(X/H)P(H)}{P(X)} \]

**Table 2. Bayesian Classification Example**

**TEST DATA OUTPUT**

Disease Identified: GASTROINTESTINAL BLEEDING
Positive Probability: 104 Negative Probability: 39
Disease Identified: FOOD POISONING
Positive Probability: 40 Negative Probability: 0
Disease Identified: GASTROENTERITIS
Positive Probability: 48 Negative Probability: 0
Disease Identified: GENERALIZED ANXIETY DISORDER
Positive Probability: 48 Negative Probability: 0
Disease Identified: IRRITABLE BOWEL SYNDROME
Positive Probability: 56 Negative Probability: 0

Disease Identified: PANIC ATTACKS
Positive Probability: 15 Negative Probability: 0

Disease Identified: PEPTIC ULCER
Positive Probability: 12 Negative Probability: 0

Disease Identified: NARCOTIC ABUSE
Positive Probability: 12 Negative Probability: 0

Disease Identified: IRON POISONING
Positive Probability: 20 Negative Probability: 0

Total Input Records: 10
Total Success Count: 10
Success Percentage: 100%

ALGORITHM FOR NAVI BAYES CLASSIFICATION

Step 1: START
Step 2: Get disease (or) patient record as table -> ITB
Step 3: Calculate probability table for (ITB) ->Attribute table (Frequency table)
Step 4: Calculate negative and positive probability table -> FPT
Step 5: Test with test data (FPT, Test Data) -> Output
Step 6: Check for positive and negative probability value which is higher
Step 7: If (Positive > Negative) then
  Show output class as positive
else
  Show output class as negative
Step 8: END

6.4 ADAPTIVE CLASSIFIER

Adaptive classifier contains combined future of rule-based classification, decision tree classification and Bayesian classification. This algorithm is derived by combining the above three algorithms. Adaptive classifier perform well in medical data classification.[1][86][63].

Algorithm for Adaptive Classifier

Step 1: Start
Step 2: Get Training data -> TD
Step 3: Perform DCT(TD) -> KS1
Step 4: Perform RBC(TD) -> KS2
Step 5: Perform NB(TD) -> KS3
Step 6: Get Sample Test Data -> STD
Step 7: Perform DCT(STD, TD) -> OT1
Step 8: Perform RBC(STD, TD) -> OT2
Step 9: Perform NB(STD, TD) -> OT3
Step 10: Combine(OT1, OT2, OT3) -> FT
Step 12: END
**TEST DATA OUTPUT**

Rule Based: GASTROINTESTINAL BLEEDING Identified
DTC Based: GASTROINTESTINAL BLEEDING Identified
NBC Based: GASTROINTESTINAL BLEEDING Identified

Rule Based: FOOD POISONING Identified
DTC Based: FOOD POISONING Identified
NBC Based: FOOD POISONING Identified
NBC Based: GASTROENTERITIS Identified

Rule Based: GENERALIZED ANXIETY DISORDER Identified
DTC Based: GENERALIZED ANXIETY DISORDER Identified
NBC Based: GENERALIZED ANXIETY DISORDER Identified

Rule Based: INTESTINAL LIUS Identified
DTC Based: INTESTINAL LIUS Identified
NBC Based: INTESTINAL LIUS Identified

Rule Based: IRRITABLE BOWEL SYNDROME Identified
DTC Based: IRRITABLE BOWEL SYNDROME Identified
NBC Based: IRRITABLE BOWEL SYNDROME Identified

NBC Based: PANIC ATTACKS Identified
Rule Based: PEPTIC ULCER Identified
DTC Based: PEPTIC ULCER Identified
NBC Based: PEPTIC ULCER Identified

NBC Based: NARCOTIC ABUSE Identified
Rule Based: IRON POISONING Identified
DTC Based: IRON POISONING Identified
NBC Based: IRON POISONING Identified

Total Input Records: 10
Total Success Count: 7
Success Percentage: 70%
Total Positive Count: 24
Total Negative Count: 6
Total True Positive Count: 7
Total True Negative Count: 3
6.5 NEURAL NETWORK

Artificial Neural Network works based on biological nervous system. Artificial neural network classification algorithm uses gradient decent method, Neural network contains multiple neurons for combined processing, neural network has two phases training and testing, formerly neural networks used for classification and pattern recognition.[5][24][29]

Neural network natively dynamic because its dynamically changes its structure and weights between neurons. Weight adjustment is performed for minimizing errors. Weight adjustment performed based on input and output of current training phase. In ANN multiclass problems are solved by multilayer feed forward network, to identify chest disease by implementing probabilistic neural networks[59] and show diagnostic of various multi layer neural network[59][48][61].

Figure 3. Classification of Chest Diseases using Multilayer Neural Network

4.6 K-NN

This method is mostly used in pattern recognition. K –nearest neighborhood method is used for both classification and regression analysis [4][40][41][42].

Figure 4. Example of k-NN classification

The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If k = 3 (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If k = 5 (dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).

TEST DATA OUTPUT

Expected Disease Name: GUSTROINTESTINAL BLEEDING

Symptoms : BLEEDING, COFFEE GROUNDS COLORED VOMIT, RED COLORED VOMIT,...,

Distance with : GUSTROINTESTINAL BLEEDING : 0
Distance with : INTESTINAL LIEUS : 11344.1157874909
Distance with : PEPTIC ULCER : 13893.9396500777
Distance with : GASTROENTERITIS : 13891.0517240416
Distance with : NARCOTIC ABUSE : 13889.3179458172
Distance with : FOOD POISONING : 14.0356688476182
Distance with : GENERALIZED ANXIETY DISORDER : 13885.8628828028
Distance with : IRRITABLE BOWEL SYNDROME : 16029.0039303757
Distance with : IRON POISONING : 11341.286613153
Distance with :PANIC ATTACKS : 13888.741087658
Min Distance Obtained:GASTROINTESTINAL BLEEDING
Expected Disease Name: FOOD POISONING
Symptoms :DIARRHEA,PAIN,VOMITING,,,,
Distance with :GASTROINTESTINAL BLEEDING : 14.035688476182
Distance with :INTESTINAL LIEUS : 1134.1185642605
Distance with :PEPTIC ULCER : 13893.9370230327
Distance with :GASTROENTERITIS : 13891.057191942
Distance with :NARCOTIC ABUSE : 13889.3221576865
Distance with :FOOD POISONING : 0
Distance with :GENERALIZED ANXIETY DISORDER : 13885.877907817
Distance with :IRRITABLE BOWEL SYNDROME : 16029.0093892293
Distance with :IRON POISONING : 11341.2920339792
Distance with :PANIC ATTACKS : 13888.7457317067
Min Distance Obtained: FOOD POISONING

SVM is invented by vapnik et al. based on statistical learning. SVM method initially support binary classification, further it could be extended for multi class problems. SVM creates hyperplane for single dimension problem and multiple hyperplane for multiple problems. These make SVM most familiar, and produces hyperplane from given input space and separate data points as different classes. Data separation done via original finite dimensional space into new higher dimension space[50][51][52].

Kernel function are used for non-linear mapping of training sample to high dimensional space. Different types of kernel functions available such as Gaussian, polynomial, sigmoid, etc. In short SVM separate input data points into hyperplane, hyperplane constructed with the help of support vectors. SVM support both linear and non-linear data separation[55][56][67].

**Figure 4. Classification of Diabetic Patients using Support Vector Machine**

**Introduction to Clustering**
Clustering comes under the category of unsupervised learning method, clustering different from classification because classification most relevant to supervised and clusters based on data point similarity between given data points. There are many clustering algorithms available for various problems. For example gene expression data clustering using hieratical and genetic algorithm approach[68][72].

**6.8 PARTITIONED CLUSTERING**
In this clustering unknown or unlabeled data set, n data points portioned into K Clusters each clusters must contains one data points and each data point must hooked to one cluster. In this method we need to define k-number of cluster in initial stage partition of clustering, further divided into two major methods such as k-means and k-mediods[71][74]. k-means method is an widely adopted and enhanced periodically. In k-means n data points into k-cluster using Euclidian distance between data points and cluster. Distance of data point and different clusters may vary short distance with cluster center taken as friend for categorization[5][68].

**Pseudo code for K-means algorithm:**
Step 1: Start
Step 2: Get number of clusters --> NC
Step 3: Get number iterations --> NI
Step 4: Calculate initial centroids(NC,NI) --> IC
Step 5: Calculate Euclidian distance(data items IC)
Step 6: Cluster data items(IC,Data items)
Step 7: Check for centroid relocation(IC, Data items) > RC
Step 8: End

For Example discovered the causes of risk related with fluoride content in drinking water using k-means algorithm[73]. Figure k means shows high blood pressure and cholesterol using k-means clustering.

**Figure 5.** Example for K-Means Algorithm for Identifying High Blood Pressure and cholesterol.

**6.9 HIERARCHICAL CLUSTERING.**

In Hierarchical clustering we don’t need to input n-number of clusters in advance. Hierarchical clustering portioning done via Hierarchical way[75]. There is two way available one is top-up approach and another one is bottom up approach. Hierachal clustering further sub-divided into two categories agglomerative and another one is divisive method. Agglomerative check the input data point to any relevant subcategory or cluster and hook to that cluster. Agglomerative frequently check if the data point only attached to one subcategory and need termination condition for each data point[5]. Divisive method opposite to agglomerative, in divisive all data points initially subdivided into two large sub categories, then further sub divided into two recursively, to complete divisive hierarchal clustering need termination condition[76][77].

Mixed clustering methods provide more performance, for example combine k-means and hierarchical approach to cluster micro – array data gives better performance than expected.[76] Another example in fig two cluster of 192-gene expression data[78] to identify disease.

**Figure 6.** Hierarchical Clustering for Grouping the Patients into Two Cluster using 192-gene Expression Profile [78]

**6.10 DENSITY BASED CLUSTERING**

Density based clustering contains many advantages than hierarcial and portioned based clustering hierachal clustering only handle spherical type data
problems. But not to handle outlier and arbitrary shaped data. Density based clustering handle both arbitrary and outlier problems. There is an two most familiar methods available one is DBSCAN and another one is OPTICS to density cluster data.[79][72] DENCLUE is an another method for density based data clustering[5]. Figure [79] Provide un healthy skin clustering of wounded skin using DBSCAN Algorithm.

![Image](image1)

**Figure 7.** Clustering of Skin Wound Image using DBSCAN

## VII. COMPARATIVE TABLE (SUPERVISED)

**Table 4.** Classification Comparative Table

<table>
<thead>
<tr>
<th>Methods</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>K-NN</strong></td>
<td>1. It is easy to implement. 2. Training is done in faster manner.</td>
<td>1. It requires large storage space. 2. Sensitive to noise. 3. Testing is slow.</td>
</tr>
<tr>
<td><strong>Decision Tree</strong></td>
<td>1. There are no requirements of domain knowledge in the construction of decision tree. 2. It minimizes the ambiguity of complicated decisions and assigns</td>
<td>1. It is restricted to one output attribute. 2. It generates categorical output. 3. It is an unstable classifier i.e. performance of classifier is depend upon the type of exact values to outcomes of various actions. 3. It can easily process the data with high dimension. 4. It is easy to interpret. 5. Decision tree also handles both numerical and categorical data.</td>
</tr>
<tr>
<td><strong>Support Vector Machine</strong></td>
<td>1. Better Accuracy as compare to other classifier. 2. Easily handle complex nonlinear data points. 3. Over fitting problem is not as much as other methods.</td>
<td>1. Computationally expensive. 2. The main problem is the selection of right kernel function. For every dataset different kernel function shows different results. 3. As compare to other methods training process take more time. 4. SVM was designed to solve the problem of binary class. It solves the problem of multi class by breaking it into pair of two classes such as oneagainst-one and one-againstall.</td>
</tr>
<tr>
<td><strong>Neural</strong></td>
<td>1. Easily</td>
<td>1. Local minima.</td>
</tr>
</tbody>
</table>
Network |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>identify complex relationships between dependent and independent variables.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Able to handle noisy data.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Over-fitting.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. The processing of ANN network is difficult to interpret and require high processing time if there are large neural networks.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bayesian Belief Network |
| 1. It makes computations process easier. |
| 2. Have better speed and accuracy for huge datasets. |
| 1. It does not give accurate results in some cases where there exists dependency among variables. |

Density Based Clustering |
| 1. No need to specify number of cluster in advance. |
| 2. Easily handle cluster with arbitrary shape. |
| 3. Worked well in the presence of noise. |
| 1. Not handle the data points with varying densities. |
| 2. Results depend on the distance measure. |

VIII. COMPARATIVE TABLE

(UN-SUPERVISED)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical Clustering</td>
<td>1. Easy to implement. 2. Having good visualization capability. 3. There is no need to specify</td>
<td>1. Have cubic time complexity in many cases so it is slower. 2. Decision regarding</td>
</tr>
</tbody>
</table>
analysis. This paper shows all about machine learning approach to medical data, machine learning is a broad domain which is include data classification, cluster analysis and pattern recognition. This paper also state that which type of algorithm suitable for specific data type such as statistical, spherical and arbitrary.

X. REFERENCES

[1]. Sneha Chandra and Maneet Kaur, "Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the Field of Medical Data Mining, 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).


[4]. Mohammadreza Balouchestani, Member, IEEE and Sridhar Krishnan, Senior Member, IEEE, "Fast Clustering Algorithm for Large ECG Data Sets Based on CS theory in Combination with PCA and K-NN Methods", 978-1-4244-7929-0/14/$26.00 ©2014 IEEE.


S. Gupta, D. Kumar and A. Sharma, "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis", (2011).


S. Chao and F. Wong, "An Incremental Decision Tree Learning Methodology Regarding Attributes in Medical Data Mining", (2009).


W. L. Zuoan, Z. Y. Wanga, T. Liua and H. L. Chenc, "Effective detection of Parkinson’s disease using an


