© 2018 IJSRSET | Volume 4 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

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, Chennaii, 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, un-supervised 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 bio-informatics, 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 un-supervised 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 1

DISEASEID	DISEASENAME	ADDDESCRIPTION	DISEASEID1	SYMPTOMNAME
1001	GUSTROINTESTINAL BLEEDING	GUSTROINTESTINAL BLEEDING	1001	BLEEDING
1001	GUSTROINTESTINAL BLEEDING	GUSTROINTESTINAL BLEEDING	1001	RED COLORED VOMIT
1001	GUSTROINTESTINAL BLEEDING	GUSTROINTESTINAL BLEEDING	1001	COFFEE GROUNDS COLORED VOMIT
1002	FOOD	FOOD	1002	VOMITIN

	POISONING	POISONING		G
1002	FOOD POISONING	FOOD POISONING	1002	DIARRHE A
1002	FOOD POISONING	FOOD POISONING	1002	PAIN
1003	GASTROEN TERITIS	STOMACH FLUE	1003	VOMITIN G
1003	GASTROEN TERITIS	STOMACH FLUE	1003	PAIN
1003	GASTROEN TERITIS	STOMACH FLUE	1003	BLOATIN G
1003	GASTROEN TERITIS	STOMACH FLUE	1003	DECREAS ED APPETITE
1003	GASTROEN TERITIS	STOMACH FLUE	1003	DIARRHE A
1003	GASTROEN TERITIS	STOMACH FLUE	1003	RED COLORE D VOMIT
1004	GENERALIZ ED ANXIETY DISORDER	GENERALIZ ED ANXIETY DISORDER	1004	CHILLS
1004	GENERALIZ ED ANXIETY DISORDER	GENERALIZ ED ANXIETY DISORDER	1004	PAIN
1004	GENERALIZ ED ANXIETY DISORDER	GENERALIZ ED ANXIETY DISORDER	1004	ANXIETY
1004	GENERALIZ ED ANXIETY DISORDER	GENERALIZ ED ANXIETY DISORDER	1004	DIZZINES S
1004	GENERALIZ ED ANXIETY DISORDER	GENERALIZ ED ANXIETY DISORDER	1004	VOMITIN G
1004	GENERALIZ ED	GENERALIZ ED	1004	AGITATI ON

	ANXIETY DISORDER	ANXIETY DISORDER		
1005	INTESTINA L LIEUS	INTESTINA L LIEUS	1005	PAIN
1005	INTESTINA L LIEUS	INTESTINA L LIEUS	1005	DECREAS ED APPETITE
1005	INTESTINA L LIEUS	INTESTINA L LIEUS	1005	CONSTIP ATION
1005	INTESTINA L LIEUS	INTESTINA L LIEUS	1005	VOMITIN G
1005	INTESTINA L LIEUS	INTESTINA L LIEUS	1005	STOMAC H CRAMPS
1006	IRRITABLE BOWEL SYNDROME	IRRITABLE BOWEL SYNDROME	1006	BLOATIN G
1006	IRRITABLE BOWEL SYNDROME	IRRITABLE BOWEL SYNDROME	1006	DIARRHE A
1006	IRRITABLE BOWEL SYNDROME	IRRITABLE BOWEL SYNDROME	1006	FREQUEN T URGE TO HAVE BOWEL MOVEME NT
1006	IRRITABLE BOWEL SYNDROME	IRRITABLE BOWEL SYNDROME	1006	INCREAS ED PASSING GAS
1006	IRRITABLE BOWEL SYNDROME	IRRITABLE BOWEL SYNDROME	1006	PAIN
1006	IRRITABLE BOWEL SYNDROME	IRRITABLE BOWEL SYNDROME	1006	CONSTIP ATION
1006	IRRITABLE BOWEL SYNDROME	IRRITABLE BOWEL SYNDROME	1006	FREQUEN T BOWEL MOVEME NT
1007	NARCOTIC	OPIATE	1007	PAIN

	ABUSE	ADDICTION		
1007	NARCOTIC ABUSE	OPIATE ADDICTION	1007	CONFUSI ON
1007	NARCOTIC ABUSE	OPIATE ADDICTION	1007	CONSTIP ATION
1007	NARCOTIC ABUSE	OPIATE ADDICTION	1007	VOMITIN G
1007	NARCOTIC ABUSE	OPIATE ADDICTION	1007	GIDDINE SS
1007	NARCOTIC ABUSE	OPIATE ADDICTION	1007	ITCHING AND BURNING
1008	PANIC ATTACKS	PANIC DISORDER	1008	ANXIETY
1008	PANIC ATTACKS	PANIC DISORDER	1008	DIZZINES S
1008	PANIC ATTACKS	PANIC DISORDER	1008	VOMITIN G
1008	PANIC ATTACKS	PANIC DISORDER	1008	GIDDINE SS
1008	PANIC ATTACKS	PANIC DISORDER	1008	IRREGUL AR HEART BEAT
1008	PANIC ATTACKS	PANIC DISORDER	1008	PAIN
1009	PEPTIC ULCER	PEPTIC ULCER	1009	RED COLORE D VOMIT
1009	PEPTIC ULCER	PEPTIC ULCER	1009	BLACK COLORE D STOOLS
1009	PEPTIC ULCER	PEPTIC ULCER	1009	WEIGHT LOSS
1009	PEPTIC ULCER	PEPTIC ULCER	1009	VOMITIN G
1009	PEPTIC ULCER	PEPTIC ULCER	1009	RED COLORE D STOOLS
1009	PEPTIC	PEPTIC	1009	PAIN

	ULCER	ULCER		
1010	IRON POISONING	IRON POISONING	1010	RED COLORE D VOMIT
1010	IRON POISONING	IRON POISONING	1010	PAIN
1010	IRON POISONING	IRON POISONING	1010	DIARRHE A
1010	IRON POISONING	IRON POISONING	1010	BLACK COLORE D STOOLS
1010	IRON POISONING	IRON POISONING	1010	RED COLORE D STOOLS

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 types of classification is performed in dynamic manner and creates feature extractions and patters automatically [2][38][39].

List of Un- Supervised Algorithm

1. Partioned Clustering
2. Hierarical 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=countofsysmptoms 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=countofsysmptoms 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

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Information Gain for Attribute A

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

Information Gain

$$Gain(A) = Info(D) - Info_A(D)$$

Classification Example :

DTC Classification

DiseaseID	DiseaseName	Symptom 1	Symptom 2	Symptom 3	Symptom 4	Symptom 5	Symptom 6	Symptom 7	Symptom 8	Condition
1001	GUSTROINTESTINAL BLEEDING	BLEEDING	COFFEE GROUNDS COLORED VOMIT	RED COLORED VOMIT	NIL	NIL	NIL	NIL	NIL	YES
1002	FOOD POISONING	DIARRHEA	PAIN	VOMITING	NIL	NIL	NIL	NIL	NIL	YES
1003	GASTROENTERITIS	BLOATING	DECREASED APPETITE	DIARRHEA	PAIN	RED COLORED VOMIT	VOMITING	NIL	NIL	YES
1004	GENERALIZED ANXIETY DISORDER	AGITATION	ANXIETY	CHILLS	DIZZINESS	PAIN	VOMITING	NIL	NIL	YES
1005	INTESTINAL LIEUS	CONSTIPATION	DECREASED APPETITE	PAIN	STOMACH CRAMPS	VOMITING	NIL	NIL	NIL	YES
1007	NARCOTIC ABUSE	CONFUSION	CONSTIPATION	GIDDINESS	ITCHING AND BURNING	PAIN	VOMITING	NIL	NIL	YES
1006	IRRITABLE BOWEL SYNDROME	BLOATING	CONSTIPATION	DIARRHEA	FREQUENT BOWEL MOVEMENT	FREQUENT URGE TO HAVE BOWEL MOVEMENT	INCREASED PASSING GAS	PAIN	NIL	YES
1008	PANIC ATTACKS	ANXIETY	DIZZINESS	GIDDINESS	IRREGULAR HEART BEAT	PAIN	VOMITING	NIL	NIL	YES

Figure 1. Decision Tree

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 %

6.3 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]

Bays theorem as follows

$$P(H/\mathbf{X}) = \frac{P(\mathbf{X}/H)\,P(H)}{P(\mathbf{X})}$$

Disease Name :GUSTROINTESTINAL BLEEDING

Attribute Name	Attribute Value	postivefreq	negativefreq
Symptom1	BLEEDING	2	0
Symptom1	BLACK COLORED STOOLS	0	1
Symptom1	DIARRHEA	0	1
Attribute Name	Attribute Value	postivefreq	negativefreq
Symptom2	PAIN	0	1
Symptom2	RED COLORED VOMIT	1	0
Symptom2	DIARRHEA	0	1
Symptom2	COFFEE GROUNDS COLORED VOMIT	1	0
Attribute Name	Attribute Value	postivefreq	negativefreq
Symptom3	RED COLORED VOMIT	1	0
Symptom3	PAIN	0	1
Symptom3	VOMITING	0	1
Symptom3	COFFEE GROUNDS COLORED VOMIT	1	0
Attribute Name	Attribute Value	postivefreq	negativefreq
Symptom4	RED COLORED STOOLS	0	1
Attribute Name	Attribute Value	postivefreq	negativefreq
Symptom5	RED COLORED VOMIT	0	1

Table 2. Bayesian Classification Example

TEST DATA OUTPUT

Disease Identified : GUSTROINTESTINAL 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

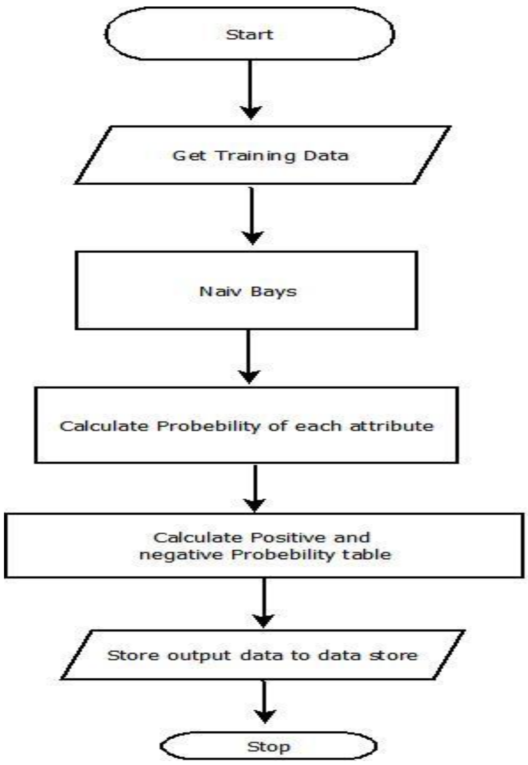
Disease Identified : INTESTINAL LIEUS

Positive Probability:40 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

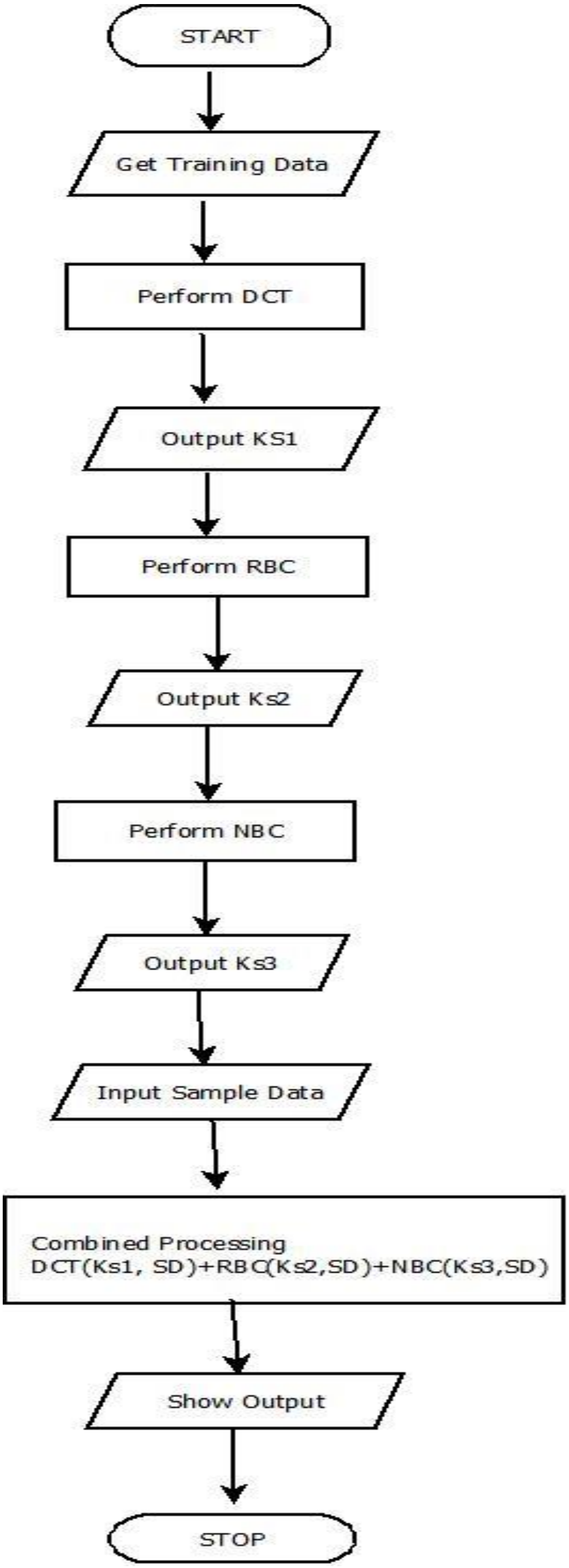


6.4 ADAPTIVE CLASSIFIER

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 :GUSTROINTESTINAL BLEEDING Identified

DTC Based :GUSTROINTESTINAL BLEEDING Identified

NBC Based :GUSTROINTESTINAL 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 LIEUS Identified

DTC Based :INTESTINAL LIEUS Identified

NBC Based :INTESTINAL LIEUS 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].

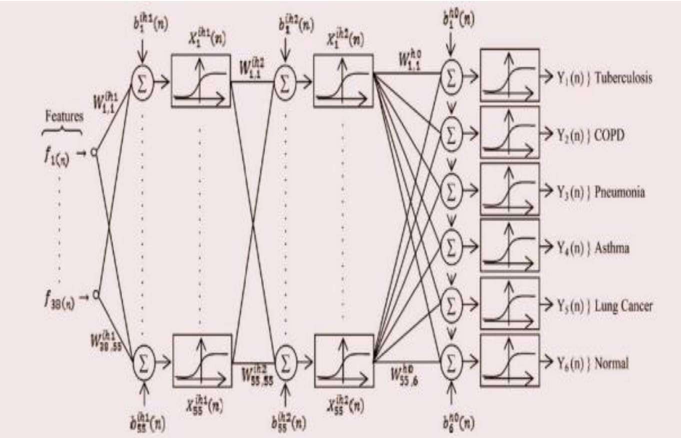


Figure 3. Classification of Chest Diseases using Multilayer Neural Network

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].

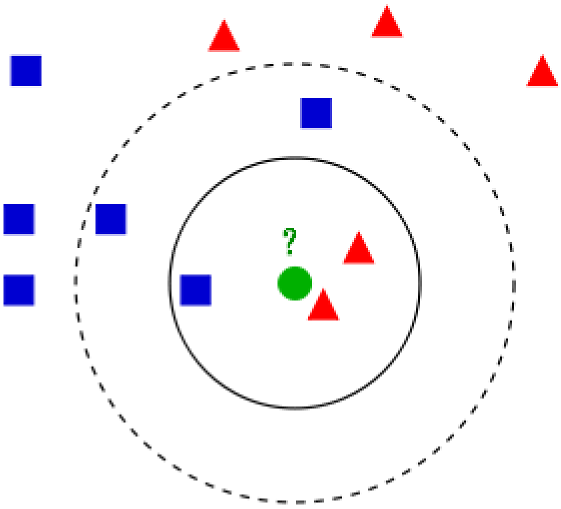


Figure 4. Example of k-NN classification

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 : 11341.2866113153

Distance with :PANIC ATTACKS : 13888.741087658
Min Distance Obtained:GUSTROINTESTINAL BLEEDING
Expected Disease Name: FOOD POISONING
Symptoms :DIARRHEA,PAIN,VOMITING,,,,,
Distance with :GUSTROINTESTINAL BLEEDING : 14.0356688476182
Distance with :INTESTINAL LIEUS : 11344.1185642605
Distance with :PEPTIC ULCER : 13893.9370230327
Distance with :GASTROENTERITIS : 13891.0571591942
Distance with :NARCOTIC ABUSE : 13889.3221576865
Distance with :FOOD POISONING : 0
Distance with :GENERALIZED ANXIETY DISORDER : 13885.8779700817
Distance with :IRRITABLE BOWEL SYNDROME : 16029.0093892293
Distance with :IRON POISONING : 11341.2920339792
Distance with :PANIC ATTACKS : 13888.7457317067
Min Distance Obtained: FOOD POISONING

2222226.7 SVM

SVM is invented by vapnik etal based on statistical learning. SVM method initially support binary classification, further it could be extended for multi class problems. SVM creates hyper plane for single dimension problem and multiple hyper plan for multiple problems. These make SVM most familiar, and produces hyper plane 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 hyper plane, hyper plane constructed with the help of support vectors. SVM support both linear and non-linear data separation[55][56][67].

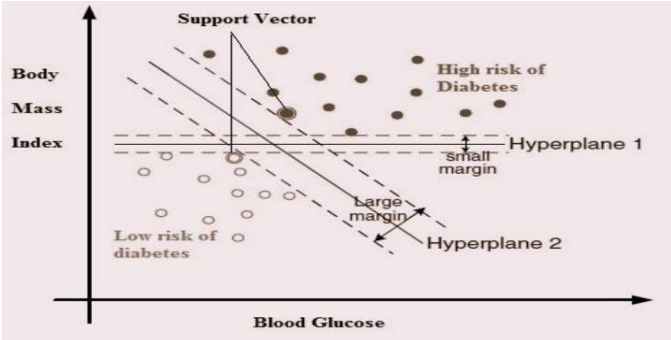


Figure 4. Classification of Diabetic Patients using Support Vector Machine

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 PARTIONED 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.

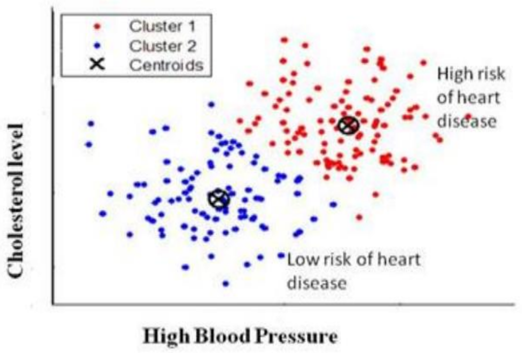
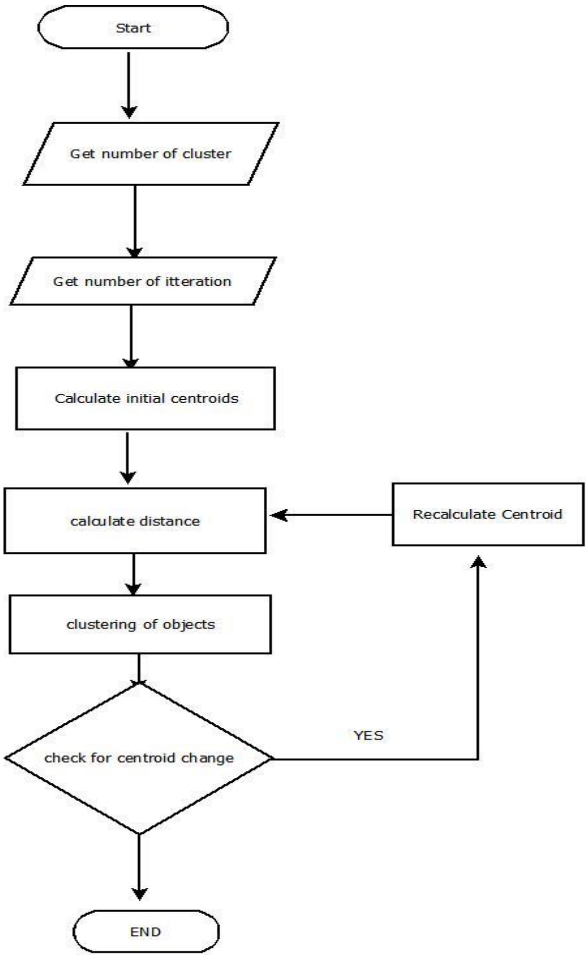


Figure 5. Example for K-Means Algorithm for Indentifying High Blood Pressure and cholesterol .



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. Hierarchal 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.

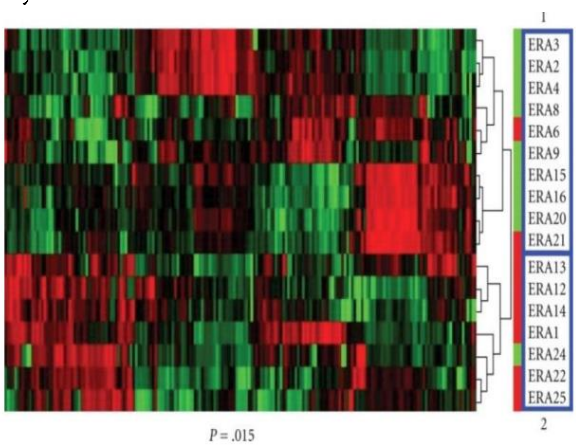


Figure 6. Hierarchical Clustering for Grouping the Patients into Two Cluster using 192-gene Expression Profile [78]

6.10.DENSITY BASED CLUSTERING

Density based clustering contains many advantages than hierarcial and portioned based clustering hierarchal 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.

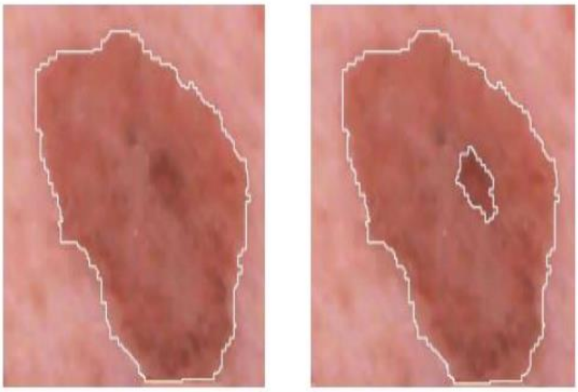


Figure 7. Clustering of Skin Wound Image using DBSCAN

VII. COMPARATIVE TABLE (SUPERVISED)

Table 4. Classification Comparative Table

Methods	Advantage	Disadvantage
K-NN	1. It is easy to implement. 2. Training is done in faster manner.	1. It requires large storage space. 2. Sensitive to noise. 3. Testing is slow.
Decision Tree	1. There are no requirements of domain knowledge in the construction of decision tree. 2. It minimizes the ambiguity of complicated decisions and assigns 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.	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 dataset. 4. If the type of dataset is numeric than it generates a complex decision tree
Support Vector Machine	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.	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.
Neural	1. Easily	1. Local minima.

Network	identify complex relationships between dependent and independent variables. 2. Able to handle noisy data.	2. Over-fitting. 3. The processing of ANN network is difficult to interpret and require high processing time if there are large neural networks.
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.

VIII. COMPARATIVE TABLE

(UN-SUPERVISED)

Methods	Advantage	Disadvantage
K-means Clustering	1. Simple clustering approach. 2. Efficient. 3. Less complex method.	1. Requires number of cluster in advance. 2. Problem with handling categorical attributes. 3. Not discover the cluster with non-convex shape. 4. Result varies in the presence of outlier.
Hierarchical Clustering	1. Easy to implement. 2. Having good visualization capability. 3. There is no need to specify the number of clusters in advance.	1. Have cubic time complexity in many cases so it is slower. 2. Decision regarding selection of merge or split point. Once a decision is made it cannot be undone. 3. Not work well in the presence of noise and outlier. 4. Not scalable.
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.

IX. CONCLUSION

Thus, this paper provides complete survey for medical data mining techniques, Classification methods, clustering and pattern recognition. In this paper we have studied disadvantages and advantages of all methods and algorithm involved in the field of medical data mining. Classification methods mostly categorized into supervised and semi-supervised learning classification, labeled data classification also called statistical data classification. Cluster analysis used for unlabeled data, cluster analysis also called as un-supervised learning. Cluster analysis contains different types of methods, to handle Spherical and Arbitrary type data problems.

Pattern recognition is the process of extracting useful features and generates knowledge using classification and cluster analysis. The Second main use of pattern recognition provide reverse process of predicting given data point using specific classification or cluster analysis. This paper shows all about machine learning approach to medical data, machine learning is an 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).

[2]. Hernan Dar ıo Vargas Cardona, AlvaroAngel Orozco and Mauricio A.Alvarez , "Unsupervised Learning applied in MER and ECG Signals through Gaussians Mixtures with the Expectation-Maximization Algorithm and Variational Bayesian Inference", 35th Annual International Conference of the IEEE EMBS Osaka, Japan, 3 - 7 July, 2013.

[3]. Devendra Naga, Dr. Swati Sharma, "Simultaneous 12-Lead QRS Detection by K-means Clustering Algorithm", IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), May 09-11, 2014, Jaipur India.

[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.

[5]. Su Liu, Nuri F. Ince, Senior Member IEEE, Akin Sabanci, Aydin Aydoseli, Yavuz Aras, Altay Sencer,Nerses Bebek, Zhiyi Sha and Candan Gurses, "Detection of High Frequency Oscillations in Epilepsy with K-means Clustering Method", 7th Annual International IEEE EMBS Conference on Neural Engineering Montpellier, France, 22 - 24 April, 2015.

[6]. U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery in databases", Commun. ACM, vol. 39, no. 11, (1996), pp. 24-26.

[7]. C. McGregor, C. Christina and J. Andrew, "A process mining driven framework for clinical guideline improvement in critical care", Learning from Medical Data Streams 13th Conference on Artificial Intelligence in Medicine (LEMEDS). http://ceur-ws. org, vol. 765, (2012).

[8]. M. Silver, T. Sakara, H. C. Su, C. Herman, S. B. Dolins and M. J. O'shea, "Case study: how to apply data mining techniques in a healthcare data warehouse", Healthc. Inf. Manage, vol. 15, no. 2, (2001), pp. 155-164.

[9]. P. R. Harper, "A review and comparison of classification algorithms for medical decision making", Health Policy, vol. 71, (2005), pp. 315-331.

[10]. V. S. Stel, S. M. Pluijm, D. J. Deeg, J. H. Smit, L. M. Bouter and P. Lips, "A classification tree for predicting recurrent falling in community-dwelling older persons", J. Am. Geriatr. Soc., vol. 51, (2003), pp. 1356-1364.

[11]. R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines", Int. J. Med. Inform., vol. 77, (2008), pp. 81-97.

[12]. R. D. Canlas Jr., "Data Mining in Healthcare:Current Applications and Issues", (2009).

[13]. F. Hosseinkhah, H. Ashktorab, R. Veen, M. M. Owrang O., "Challenges in Data Mining on Medical Databases", IGI Global, (2009), pp. 502-511.

[14]. M. Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST ISSN: 2229- 4333, vol. 2, no. 2, (2011) June.

[15]. J. Soni, U. Ansari, D. Sharma and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", (2011).

[16]. C. S. Dangare and S. S. Apte, "Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques", (2012).

[17]. K. Srinivas, B. Kavihta Rani and Dr. A.Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering, vol. 02, no. 02, (2010), pp. 250-255.

[18]. A. A. Aljumah, M. G.Ahamad and M. K. Siddiqui, "Predictive Analysis on Hypertension Treatment Usinging Approach in Saudi Arabia", Intelligent Information Management, vol. 3, (2011), pp. 252-261.

[19]. D. Delen, "Analysis of cancer data: a data mining approach", (2009).

[20]. A. O. Osofisan, O. O. Adeyemo, B. A. Sawyerr and O. Eweje, "Prediction of Kidney Failure Using Artificial Neural Networks", (2011).

[21]. S. Floyd, "Data Mining Techniques for Prognosis in Pancreatic Cancer", (2007).

[22]. M.-J. Huang, M.-Y. Chen and S.-C. Lee, "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis", Expert Systems with Applications, vol. 32, (2007), pp. 856-867.

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

[24]. K. S. Kavitha, K. V. Ramakrishnan and M. K. Singh, "Modeling and design of evolutionary neural network for heart disease detection", IJCSI International Journal of Computer Science Issues, ISSN (Online): 1694- 0814, vol. 7, no. 5, (2010) September, pp. 272-283.

[25]. S. H. Ha and S. H. Joo, "A Hybrid Data Mining Method for the Medical Classification of Chest Pain", International Journal of Computer and Information Engineering, vol. 4, no. 1, (2010), pp. 33-38.

[26]. R. Parvathi and S. Palaniammalì, "An Improved Medical Diagnosing Technique Using Spatial Association Rules", European Journal of Scientific Research ISSN 1450-216X, vol. 61, no. 1, (2011), pp. 49-59.

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

[28]. A. Habrard, M. Bernard and F. Jacquenet, "Multi-Relational Data Mining in Medical Databases", SpringerVerlag, (2003).

[29]. S. B. Patil and Y. S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research ISSN 1450-216X, EuroJournals Publishing, Inc., vol. 31, no. 4, (2009), pp. 642-656.

[30]. A.Shukla, R. Tiwari, P. Kaur, Knowledge Based Approach for Diagnosis of Breast Cancer, IEEE International Advance Computing Conference,IACC 2009.

[31]. L. Duan, W. N. Street & E. Xu, Healthcare information systems: data mining methods in the creation of a clinical recommender system, Enterprise Information Systems, 5:2, pp169-181 , 2011.

[32]. D. S. Kumar, G. Sathyadevi and S. Sivanesh, "Decision Support System for Medical Diagnosis Using Data Mining", (2011).

[33]. S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", (2008).

[34]. H. Hu, J. Li, A. Plank, H. Wang and G. Daggard, "A Comparative Study of Classification Methods For Microarray Data Analysis", Proc. Fifth Australasian Data Mining Conference (AusDM2006), Sydney, Australia. CRPIT, ACS, vol. 61, (2006), pp. 33-37.

[35]. C. Hattice and K. Metin, "A Diagnostic Software tool for Skin Diseases with Basic and Weighted K-NN", Innovations in Intelligent Systems and Applications (INISTA), (2012).

[36]. R. Potter, "Comparison of classification algorithms applied to breast cancer diagnosis and prognosis", advances in data mining, 7th Industrial Conference, ICDM 2007, Leipzig, Germany, (2007) July, pp. 40-49.

[37]. G. Beller, "The rising cost of health care in the United States: is it making the United States globally noncompetitive?", J. Nucl. Cardiol., vol. 15, no. 4, (2008), pp. 481-482.

[38]. D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala and G. Wang, "Algorithmic prediction of health-care costs", Oper. Res., vol. 56, no. 6, (2008), pp. 1382-1392.

[39]. C. H. Jena, C. C. Wang, B. C. Jiangc, Y. H. Chub and M. S. Chen, "Application of classification techniques on development an early-warning systemfor chronic illnesses", Expert Systems with Applications, vol. 39, (2012), pp. 8852-8858.

[40]. M. Shouman, T. Turner and R. Stocker, "Applying K-Nearest Neighbour in Diagnosing Heart Disease Patients", International Conference on Knowledge Discovery (ICKD-2012), (2012).

[41]. D. Y. Liu, H. L. Chen, B. Yang, X. E. Lv, N. L. Li and J. Liu, "Design of an Enhanced Fuzzy k-nearest Neighbor Classifier Based Computer Aided Diagnostic System for Thyroid Disease", Journal of Medical System, Springer, (2012).

[42]. W. L. Zuoa, Z. Y. Wanga, T. Liua and H. L. Chenc, "Effective detection of Parkinson's disease using an

adaptive fuzzy k-nearest neighbor approach", Biomedical Signal Processing and Control, Elsevier, (2013), pp. 364-373.

[43]. Goharian & Grossman, Data Mining Classification, Illinois Institute of Technology, http://ir.iit.edu/~nazli/cs422/CS422-Slides/DM-Classification.pdf, (2003).

[44]. Apte & S.M. Weiss, Data Mining with Decision Trees and Decision Rules, T.J. Watson Research Center, http://www.research.ibm.com/dar/papers/pdf/fgcsaptewe issue_with_cover.pdf, (1997).

[45]. M. U. Khan, J. P. Choi, H. Shin and M. Kim, "Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare", 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, (2008) August 20-24.

[46]. C. Chien and G. J. Pottie, "A Universal Hybrid Decision Tree Classifier Design for Human Activity Classification", 34th Annual International Conference of the IEEE EMBS San Diego, California USA, (2012) August 28-September 1.

[47]. S. S. Moon, S. Y. Kang, W. Jitpitaklert and S. B. Kim, "Decision tree models for characterizing smoking patterns of older adults", Expert Systems with Applications, Elsevier, vol. 39, (2012), pp. 445-451.

[48]. C. L. Chang and C. H. Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis", Expert Systems with Applications, Elsevier, vol. 36, (2009), pp. 4035-4041.

[49]. V. Vapnik, "Statistical Learning Theory", Wiley, (1998).

[50]. V. Vapnik, "The support vector method of function estimation", (1998).

[51]. N. Chistianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines, and other kernel-based learning methods", Cambridge University Press, (2000).

[52]. N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines", Cambridge University Press, (2000).

[53]. T. H. A. Soliman, A. A. Sewissy and H. A. Latif, "A Gene Selection Approach for Classifying Diseases Based on Microarray Datasets", 2nd International Conference on Computer Technology and Development (lCCTD 2010), (2010).

[54]. S. W. Fei, "Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine", Expert Systems with Applications, Elsevier, vol. 37, (2010), pp. 6748-6752.

[55]. C. L. Huang, H. C. Liao and M. C. Chen, "Prediction model building and feature selection with support vector machines in breast cancer diagnosis", Expert Systems with Applications, vol. 34, (2008), pp. 578-587.

[56]. E. Avci, "A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier", Expert Systems with Applications, Elsevier, vol. 36, (2009), pp. 10618-10626.

[57]. M. J. Abdi and D. Giveki, "Automatic detection of erythemato-squamous diseases using PSO–SVM based on association rules", Engineering Applications of Artificial Intelligence, vol. 26, (2013), pp. 603-608.

[58]. M. H. Dunham, "Data mining introductory and advanced topics", Upper Saddle River, NJ: Pearson Education, Inc., (2003).

[59]. O. Er, N. Yumusakc and F. Temurtas, "Chest diseases diagnosis using artificial neural networks", Expert Systems with Applications, vol. 37, (2010), pp. 7648-7655.

[60]. R. Das, I. Turkoglub and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles", Expert Systems with Applications, vol. 36, (2009), pp. 7675-7680.

[61]. S. Gunasundari and S. Baskar, "Application of Artificial Neural Network in identification of Lung Diseases", Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on. IEEE, (2009).

[62]. K. F. R. Liu and C. F. Lu, "BBN-Based Decision Support for Health Risk Analysis", Fifth International Joint Conference on INC, IMS and IDC, (2009).

[63]. D. I. Curiac, G. Vasile, O. Banias, C. Volosencu and A. Albu, "Bayesian Network Model for Diagnosis of Psychiatric Diseases", Proceedings of the ITI 2009 31st Int. Conf. on Information Technology Interfaces, Cavtat, Croatia, (2009) June 22-25.

[64]. J. Fox, "Applied Regression Analysis, Linear Models, and Related Methods", (1997).

[65]. P. A. Gutiérrez, C. Hervás-Martínez and F. J. Martínez-Estudillo, "Logistic Regression by Means of Evolutionary Radial Basis Function Neural

Networks", IEEE Transactions on Neural Networks, vol. 22, no. 2, (2011), pp. 246-263.

[66]. C. Gennings, R. Ellis and J. K. Ritter, "Linking empirical estimates of body burden of environmental chemicals and wellness using NHANES data", http://dx.doi.org/10.1016/j.envint.2011.09.002,2011.

[67]. Divya and S. Agarwal, "Weighted Support Vector Regression approach for Remote Healthcare monitoring", IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011, 978-1-4577- 0590-8/11/$26.00 © 2011 IEEE MIT, Anna University, Chennai, (2011) June 3-5.

[68]. J. J. Tapia, E. Morett and E. E. Vallejo, "A Clustering Genetic Algorithm for Genomic Data Mining", Foundations of Computational Intelligence, vol. 4 Studies in Computational Intelligence, vol. 204, (2009), pp. 249-275.

[69]. A. K. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: a review", ACM Compute, Surveys, vol. 31, (1996).

[70]. G. Hamerly and C. Elkan, "Learning the K in K-means", Proceedings of the 17th Annual Conference on Neural Information Processing Systems, British Columbia, Canada, (2003).

[71]. L. Lenert, A. Lin, R. Olshen and C. Sugar, "Clustering in the Service of the Public's Health", http://wwwstat.stanford.edu/~olshen/manuscripts/helsinki.PDF.

[72]. S. Belciug, F. Gorunescu, A. Salem and M. Gorunescu, "Clustering-based approach for detecting breast cancer recurrence", 10th International Conference on Intelligent Systems Design and Applications, (2010).

[73]. T. Balasubramanian and R. Umarani, "An Analysis on the Impact of Fluoride in Human Health (Dental) using Clustering Data mining Technique", Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, (2012) March 21-23.

[74]. J. Escudero, J. P. Zajicek and E. Ifeachor, "Early Detection and Characterization of Alzheimer's Disease in Clinical Scenarios Using Bioprofile Concepts and K-Means", 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, (2011) August 30-September 3.

[75]. H. Chipman and R. Tibshirani, "Hybrid hierarchical clustering with applications to microarray data", Biostatistics, vol. 7, no. 2, (2009), pp. 286-301.

[76]. T. S. Chen, T. H. Tsai, Y. T. Chen, C. C. Lin, R. C. Chen, S. Y. Li and H. Y. Chen, "A Combined K-Means and Hierarchical Clustering Method for improving the Clustering Efficiency of Microarray", Proceedings of 2005 International Symposium on Intelligent Signal Processing and Communication Systems, (2005).

[77]. S. Belciug, "Patients length of stay grouping using the hierarchical clustering algorithm", Annals of University of Craiova, Math. Comp. Sci. Ser., ISSN: 1223-6934, vol. 36, no. 2, (2009), pp. 79-84.

[78]. Z. Liu, T. Sokka, K. Maas, N. J. Olsen and T. M. Aune, "Prediction of Disease Severity in Patients with Early Rheumatoid Arthritis by Gene Expression Profiling", Human Genomics and Proteomics, (2009).

[79]. M. E. Celebi, Y. A. Aslandogan and R. P. Bergstresser, "Mining Biomedical Images with Density-based Clustering", Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05), (2005).

[80]. R. Agrawal, T. Imielinski and A. N. Swami, "Mining Association Rules between Sets of Items in Large Databases. SIGMOD", vol. 22, no. 2, (1993) June, pp. 207-16.

[81]. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", VLDB, Chile, ISBN 1-55860- 153-8, (1994) September 12-15, pp. 487-99.

[82]. J. Yanqing, H. Ying, J. Tran, P. Dews, A. Mansour and R. Michael Massanari, "Mining Infrequent Causal Associations in Electronic Health Databases", 11th IEEE International Conference on Data Mining Workshops, (2011).