International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015
ISSN 2091-2730

Survey on Decision Support System for Medical Diagnosis Using Data Mining

Huzaifa Shabbir Dhorajiwala[1], Er. Asadullah Shaikh[2]

[1]Department of Computer Engineering, M.H. Saboo Siddik College of Engineering, Mumbai, Maharashtra, India
dhorajiwalahuzaifa@gmail.com Mob: 9820047237

Abstract— Industries in healthcare gathers a large amount of data that has not been appropriately mined and which is not feasible. Research of these covered patterns and relationships among them have not been put into use. The main motto of our project based on Medical diagnosis is to prepare knowledge based aspects with the help of collecting data regarding various kinds of diseases such as heart disease and diabetes. Another aspect of our research is to develop decision support systems in medical field to ease the workload of the physicians. In our proposed system, we make use of different algorithms such as ID3 algorithm, CART algorithm, Genetic algorithm and LS-SVM algorithm which classifies the above mentioned diseases to compare the usefulness and how effective it is.

Keywords— Healthcare, Medical Diagnosis, ID3 algorithm, CART, Genetic and LS-VSM, Decision Support.

INTRODUCTION

The Healthcare Industry gathers data in bulk that is not well mined and not put into use which is feasible. Analysis of the hidden patterns and the relationships among them often goes undetected. We can overcome these shortcomings by using advanced data mining techniques. With the help of these data mining techniques and decision support techniques we can improve the management of heart disease.

Providing quality services at minimal costs is the major challenge faced by the Healthcare Industries.

Quality services means treating patients correctly, taking care of the financial conditions of each and every individual. Substandard decisions can lead to worse results. Even the highly rated hospitals and clinics in India do not have software that checks and predicts a disease through data mining techniques.

There is a large amount of data which goes undetected and this data can be turned into information which could prove useful in various Healthcare centres. Medical diagnosis is said to be instinctive. It is totally dependent on the physician who examines the patients. It becomes very difficult for the physician to make good decisions if the data is too large. The data often becomes impossible to manage.

In such cases, the doctor or the physician uses machine learning techniques to derive past rules, patients who were treated successfully and this helps the physician to make the examination process more trustworthy.

The concept of Decision Support System [DSS] is very large because of many different approaches and a vast range of domains which prove helpful in making decisions. DSS phraseology implies to a class of computer information systems which includes knowledge based systems that help us in making decisions

LITERATURE SURVEY

Algorithms such as ID3 and C4.5 were proposed by Quinlan for promoting classification models which was extracted from data known as decision tress. Set of information are known to us. Per capita of information has the same format consisting of number of element pairs. An element represents the group of the information. The dilemma is to find a decision tree that on the base of solution to problems about the non-listing elements forecast properly the value of the listing elements. Normally the listing element takes only the values i.e. {right, wrong} or {pass, fail} or something identical. In any instance one of its points will mean fail.

A. ID3 Algorithm

Itemized Dichotomize 3 algorithm also known as ID3 algorithm was put forward by J.R Quinlan in the year 1970. It is a materialistic algorithm that will pick up the next element based on the data that been collected related to the element. The data obtain is calculated by entropy, ID3 algorithm tender that the initiate tree is minuscule with lower elements that are put closer to the tree. ID3 algorithm is a sample of symbolic learning and rule induction. It is also an administer learner which means it's an example like a training data set which makes the decisions. J. Ross Quinlan introduced it late back in 1979. It is like a decision tree on numerical evaluation.

Fig.1 Decision Tree

A decision tree classifies data using its elements. It is upturned process. Decision nodes and leaf nodes are included in the tree. In Fig 1, "link From Academia" element is a decision node and the "author" attribute is the leaf node. The leaf node has equivalent data which means additional classification is not required. ID3 algorithm constructs same decision trees until all the leaf nodes are equivalent.

B. CART Algorithm

Classification and regression trees is a non- parametric technique that produces either classification or regression trees, built upon on whether the subordinate variable is absolute or numeric, respectively. Trees are formed by a collection of rules based on values of certain variables in the modelling data set. Rules are selected based on how well splits based on variables' values can differentiate observations based on the dependent variable once a rule is selected and splits a node into two, the same logic is applied to each "child" node (i.e. it is a looping procedure). Splitting stops when CART encounter no further gain can be made, or some pre-set stopping conditions are met. The basic idea of tree growing is to choose a split among all the possible splits at each node so that the resulting child nodes are the "purest". In this algorithm, only unilabiate splits are considered. That is, each split built upon on the value of only one predictor variable.

SURVEY

The purpose of the current study is the growth and examination of a clinical decision support system for the treatment of patients with heart disease and diabetes. One of the study shows, heart disease is the major cause of death in the universe every year. In the United States, almost 930,000 people die and its price is about 393.5 billion dollars. Heart disease also knows as coronary artery disease (CAD), is a major term that can refer to any state that affects the heart. Diabetes mellitus is a chronic disease and a broad public health challenge altogether. According to the international diabetes federation, there are about 246 million diabetic people worldwide, and this number is predicted to increase to 380 million by 2025.

A. Experimental Data

Table 1: Description of the features in the heart disease dataset

No	Name	Description
1	Age	age in years
2	Sex	1 = male ; 0 = female
3	Cp	chest pain type (1 = typical angina; 2 = atypical angina ; 3 = non-anginal pain; 4 = asymptomatic)
4	Trestbps	resting blood pressure(in mm Hg on admission to the hospital)
5	Chol	serum cholestoral in mg/dl
6	Fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	Restecg	resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or define left ventricular hypertrophy by Estes' criteria)
8	Thalach	maximum heart rate achieved
9	Exang	exercise induced angina (1 = yes; 0 = no)
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	the slope of the peak exercise ST segment (1 = upsloping; 2 = flat ; 3= downsloping)
12	Ca	number of major vessels (0-3) colored by flourosopy
13	Thal	(3 = normal; 6 = fixed defect; 7 = reversible defect)
14	Num	Diagnosis classes (0 = healthy; 1 = patient who is subject to possible heart disease)

Table 2: description of the features in the diabetes dataset

No	Attribute Name	Description
1	Number of times pregnant	Numerical values
2	Plasma glucose concentration	glucose concentration in a 2 hours in an oral glucose tolerance test
3	Diastolic blood pressure	In mm Hg
4	Triceps skin fold thickness	Thickness of skin in mm
5	2-Hour serum insulin	Insulin (mu U/ml)
6	Body mass index	(weight in kg/(height in m)^2)
7	Diabetes pedigree function	A function – to analyse the presence of diabetes
8	Age	Age in years
9	Class	1 is interpreted as "tested positive for diabetes and 0 as negative

B. ID3 Algorithm

Itemized Dichotomize 3 algorithm also known as ID3 algorithm was put forward by J.R Quinlan in the year 1970. It is a materialistic algorithm that will pick up the next element based on the data that been collected related to the element. The data obtain is calculated by entropy ID3 algorithm tender that the initiate tree is minuscule with lower elements that are put closer to the tree. ID3 algorithm is a sample of symbolic learning and rule induction. It is also an administer learner.

1) Training Data and Set: ID3 algorithm is an administer learner. It requires training data sets to make settlement. The training set lists the elements and their feasible values. ID3 doesn't deal with constant, numeric information which means we have to approximate them. Elements such as age which can have values like 1 to 100 are listed like old or young.

TABLE I
TRANING SET

Attributes	**Values**
Age	Young, Middle aged, Old
Height	Tall, Short, Medium
Employed	Yes, No

The training data is the list of data containing actual values

TABLE II
TRANING DATA

Age	**Height**	**Employed**
Young	Tall	Yes
Old	Short	No
Old	Medium	No
Young	Medium	Yes

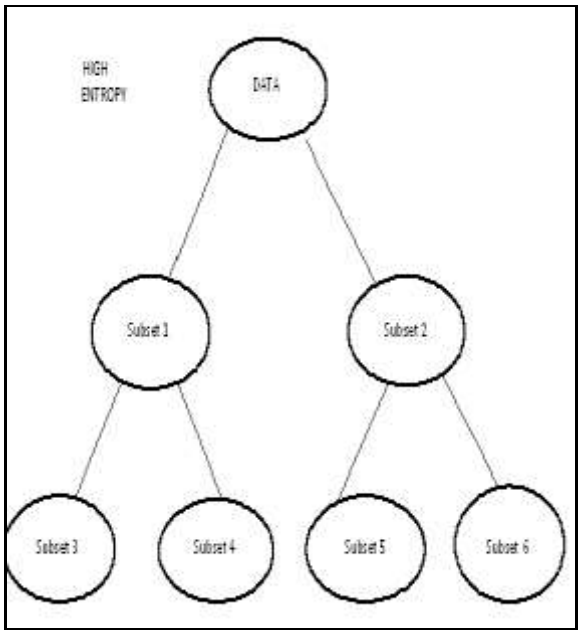

Fig. 1 Entropy

2) Entropy:

Entropy introduce to the non-coherence of the information. It ranges from 0-1. Data sets with entropy 1 means it is as non-coherence which is of similar kind. In Fig [2], the root of the tree has a collection of Data. It has big entropy which means the data is coherent. The set of data is properly divided into subsets 3, 4, 5 and 6 where it is now of same kind and the entropy is 0 or close to 0.

Entropy is calculated by the formula:

E(S) = - (p+)*log2(p+) - (p-)*log2(p-)

"S" represents the set and "p+" are the number of data in the set "S" with right values and "p –"are the numbers of elements with wrong values.

The aim of ID3 algorithm is to divide data using decision trees, such that the concluding leaf nodes are all similar with 0 entropy.

3) Steps for ID3 algorithm : ID3 algorithm works in the following steps :-

• Create a root node for the tree
•If all examples are positive, Return the single-node tree Root, with label = +.
• If all examples are negative, Return the single-node tree Root, with label = -.
• If number of predicting attributes is null, then Return the single node tree Root, with label = most common value of the target attribute in the examples.

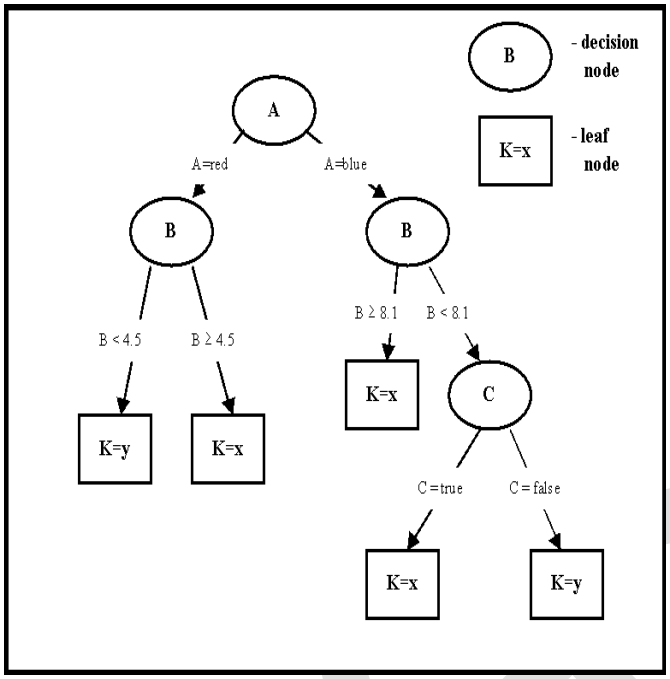

Fig. 2Solving example using ID3 algorithm

C. CART Algorithm

Classification and regression trees is a non- parametric technique that produces either classification or regression trees, built upon on whether the subordinate variable is absolute or numeric, respectively. Trees are formed by a collection of rules based on values of certain variables in the modelling data set. Rules are selected based on how well splits based on variables' values can differentiate observations based on the dependent variable once a rule is selected and splits a node into two, the same logic is applied to each "child" node (i.e. it is a looping procedure). Splitting stops when CART encounters no further gain can be made, or some pre-set stopping condition is met. The basic idea of tree growing is to choose a split among all the possible splits at each node so that the resulting child nodes are the "purest". In this algorithm, only unilabiate splits are considered. That is, each split built upon on the value of only one predictor variable.

The construction of CARTs (classification and regression trees) is best described in breiman84 and has become a common basic method for building statistical models from simple feature data. CART is dominant because it can compromise within complete data; multiple types of features (floats, enumerated sets) both in input features and predicted appearance, and the trees it generates often contain rules which are humanly readable

Decision trees contain a binary question (yes/no answer) about some feature at each node in the tree. The leaves of the tree contain the prime prediction based on the training data. Decision lists are a compressed form of this where one answer to each question leads precisely to a leaf node. A tree's leaf node may be a part of some class as a single member, a probability density function (over some discrete class), a predicted mean value for a stable feature or a Gaussian (mean and standard deviation for a continuous value).

All possible splits consist of possible splits of each predictor. CART innovations include:

• Solving the "how big to grow the tree"- problem;
• Using closely two-way (binary) splitting;
• Incorporating automatic testing and tree verification and
• Giving a completely new approach for handling missing values.

D. Genetic Algorithm:

Genetic Algorithm is an evolutionary algorithm which offers multi criterion optimization for higher dimensional space problems [11].It's a popular stochastic search method used for feature selection. It is based on Darwin's theory of natural selection and 'survival

of the fittest' [11]. Genetic algorithm search initially starts with the least number of attributes. Every set of individuals are called population and each individuals are called as chromosomes. These chromosomes are constituted of many genes which are most binary value indicating the presence of the element in the set. The search of the best result is based on the objective function called as Fitness Function [11].

$$\text{Fitness} = \frac{\text{Total No of Correctly classified Instances}}{\text{Total No of Training Samples}}$$

The selected solutions with highest fitness value have more influence than that of the new solutions with less fitness value. This function plays a key role in the selection of the best solution of the problem. In Genetic algorithm, each iteration is known as generation [11]. Fittest individuals are selected from eachgeneration and pooled out to form base for new populations. A new population is created based on the compliance to the fitness function. Off springs are generated based on the genetic operator's cross over and mutation. Threshold for fitness function will be the maximum accuracy at which the system converges. This process continues till the Fitness threshold is met.

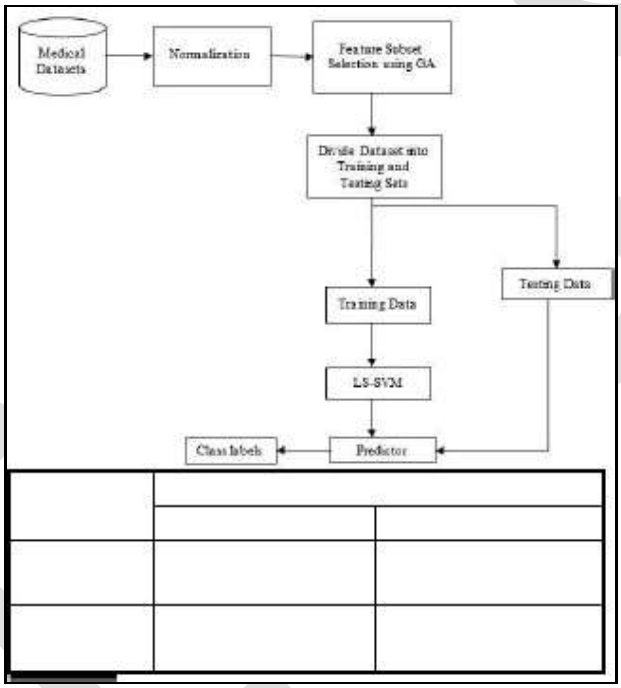

Fig 3. Proposed System.

E. LS-SVM Algorithm.

Least Square Support Vector machine (LS-SVM) is a kind of Support vector machine based on the structural risk minimization principle of statistic all earning theory [12]. Support vector machine (SVM) was introduced by Fisher [13] and has been used successfully in most regression and classification problems. The key role of SVM in Classification problem is to divide the data into two distinct classes with maximum margin and minimum classification error rate. In order to solve constrained quadratic programming problems, SVM requires higher computational load, which is a major drawback in using in high dimensional problems. In order to overcome this problem, LS-SVM was introduced by Suykens and Vandewalle [12], which uses linear equations to solve the problems.LS-SVM are used for classification problems to find a hyper plane, which could separate various classes with higher margin. An optimal hyper plane is obtained using maximum Euclidean distance to the nearest point. It maps the input vector into higher dimensional space for non-separable data. Then the optimal separating hyper plane is found. X IЄR p and YIЄ {0, 1} where XI is 'p' dimensional input vector and YI is the corresponding class label.

$$F(X) = \text{sign}\left(\sum_{i-1}^{N} Y_i \alpha_i K(X, X') + b\right)$$

Where f(X) is the output of new input vector XIЄ R p(Equation 3).XIis the support vectors belongs to training set. The dataset used for training the classifier are training set.Iα

is the Lagrange multipliers and b be the realconstant.LS-SVM performance depends mainly on two key parameters. Choosing best value for these parameters is important to maintain the classifiercharacteristic. The two kernel parameters are C and Gamma(γ) . C be the box constraint and Gamma be the regularization factor[12].Input data sets are distributed in nonlinear dimensional space. These are converted into high dimensional linear feature space by using kernels. Radial Basis Kernel is used for such mapping for our medical datasets, which is given in (4).

RBF Kernels:

$$K(x, x') = \exp(-(x-x)^2/\sigma^2).$$

Performance Evaluation:

The proposed System is examined by 10 fold cross validation methodology. The performance of the system is evaluated using four measures: Confusion Matrix, Sensitivity, Specificity and Classification Accuracy.

Confusion Matrix:

Confusion matrix [13] (COM) is a 2×2 matrix which shows the predicted and actual classification given in Table 1.

Table 1
Confusion Matrix

Predicted	Actual
Positive	
Negative	Positive
TP (true positive)	FP (false positive)
Negative	
FN (false negative)	TN (true negative)

-TN is the correct predictions of an instance as negative
.
-FN is the incorrect predictions of an instance as positive
.
-FP is the incorrect of predictions of an instance as negative
.
-TP is the correct predictions of an instance as Positive

Classification Accuracy:

Performance of classifier is commonly measured using classification accuracy (CA). It provides the rate of correctly predicted instances to the overall instances in the dataset. CA can be calculated from Confusion Matrix [13] using the equation (5).

$$\text{Accuracy} = \frac{TN + TP}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + TN}$$

$$\text{Specificity} = TN$$

TN + FP.

IV. EXISTING SYSTEM

A vital question faced by the healthcare organizations (hospitals, medical centers) is the provision of standard services at inexpensive price. Value services suggest diagnosing patients properly and examining treatments that are fruitful. Deprived clinical decisions leads to disastrous results that cannot be accepted. Hospitals should also decrease the price of clinical test. They can attain this upshot by assigning appropriate computer-based data and decision support systems.

In today's world many hospitals manage details and data related to patients. Numbers, text, charts and images are generated by the decision support system. Woefully, these data are not used to support clinical decision making. There is a opulence of unseen information in these data that is largely not utilized. This objects an important question: "How the useful information can be extracted from the untapped data?" This is the main purpose for this paper.

Diabetes	Classification Accuracy
SVM [15]	77.73
Grid Algorithm [9]	76.47
ACO –SVM [16]	67.11
DSS [17]	75.73
GA-LSSVM (Proposed)	81.33

Fig 4.Accuracy Comparison

V. SCOPE

Data mining utilization in healthcare can have tremendous potential and usefulness. Nevertheless, the success of healthcare data mining depends on the availability of clean healthcare statistics. In this respect, it is fault-finding that the healthcare industry explores how data can be better captured, stored, planned and mined. Practicable guidance includes the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining utilization.

Future Directions of Health care system through Data Mining Tools

As healthcare data are not finite to just quantitative data (e.g., doctor's notes or clinical records), it is required to also explore the use of content mining to expand the scope and nature of what healthcare data mining can directly do. This is specially used to blended all the data and then mining the content. It is also useful to look into how images (e.g., MRI scans) can be import into healthcare data mining utilization. It is noted that progress has been made in these areas.

B. UNIQUENESS

Human medical data are most profitable and challenging of all biological data to mine and study. Most closely watched category on the earth. Human subjects can provide information that cannot be gained easily from animal examination, such as optical and audible sensitivity, the recognition of pain, soreness, illusion and exposures. Most animal knowledge are limited, and therefore cannot record longstanding disease processes of medical concern, such as preneoplasia. With human statistics, there is no issue of having to deduce animal analyse to the human species.

1) Intermixture of medical statistics: Intermixture of medical statistics the major areas of intermixture of medical statistics are Volume and complication of medical statistics (crude medical data are colossal, voluminous and composite. These are gathered via interviews with the patients, various images from diagnostic approach and clinical response, analysis), Importance of physician's interpretation (the physician's perception of diagnostic decisions, even specialists from similar practice vary in describing a patient's condition hence it becomes problematic to mine). Awareness and particularity analysis (precision in nearly all interpretation and medications in medicine is subject to flaws, there is need to differentiate between a test and diagnosis in medicine. The medical condition of the patient is described by various inspection and based on these inspection disease is diagnosed. Both inspection and diagnosis are subject to awareness and particularity analysis.), Poor mathematical simulating (medical data cannot be put into formulas equations and figure because the theoretical structure of medicine consists of word characterization and images).

2) High-principled, legal and social problems: Since the medical statistics are gathered on human beings, it involves high-principled, legal and social problems that are constructed to avoid the wrongdoing of patients and squandering of their data. These issues may refer to data hold, fear of law threads, isolation and preservation of human data, expected benefits and administrative problems.

3) Analytical ideology: Analytical ideology of medical data is contrast from non-medical data. Medical data is collected to provide gain to the respective patients. Sometimes the patient who has permission to be involved in the research projects may not get any gain but the data collected from such patients is narrowly observed and is regulated by legal and high-principled attention.

4) Appropriate status of medicine Medical science: Appropriate status of medicine Medical science like special status in everyday life. Medical care is a vital part; it gives life to the patient and ambition for a beneficial life. Every patient await recovery after going to the medicine person and demands belief, understanding and care but no one recognize the uncertainty which medicine person face. Everybody wants to enjoy the benefits of research but no one is ready to contribute into it. The collected medical data when published is used for social causes without harming the dignity of the patients. Though medical science enjoys special status yet it has to face many barriers. It is still doubtful as to what questions to be request from patients, what attempt to be performed on the patients and what conclusions may not be drawn. That is why experiments on animals are conducted and results from these experiments are considered reliable.

VI. CONCLUSION

The decision-tree algorithm is one of the popular productive classification methods. The data will determine the algorithm in terms of its efficiency and correction rate. We used 10-fold cross authentication to figure out confusion matrix of each model and then check out the performance by using precision, recall, F measure and ROC space. As conventional, bagging algorithms, especially CART, showed the best performance among the tested algorithms. The results showed here make clinical application more practicable, which will provide strong leading in curing CAD, hepatitis and diabetes. The analysis is made on the decision tree algorithms ID3 and CART towards their steps of refining data and complexity of ongoing data. Finally it can be declare that between the two algorithms, the CART algorithm is the best in performance of rules generated and accuracy. This showed that the CART algorithm is better in induction and rules generalization than the ID3 algorithm. Finally, the results are reserved in the decision support depository. Since, the knowledge base is currently concentrated on a narrow set of diseases. The approach has been authenticated through the case study; it is feasible to widen the scope of formed medical knowledge. Additionally, decision support can be more enhance by considering interactions between the different medicaments that the patient is on.

In this paper, a decision support system based on GA-LSSVM is proposed for the diagnosis of the diabetes disease. A Gaussian radial basis function issued as a kernel of LS-SVM [14].

The robustness of the proposed system were analyzed with metrics like classification accuracy, using 10-fold cross-validation and confusion matrix. The accuracy of the system for the PID dataset was found to be 81.33% with GA as a feature selection method. In future, this system can be used for the diagnosis of real life medical data of patients [14].

The LSVM Algorithm is not a very feasible solution to decision making in medical field. The algorithm is difficult to implement and the time and space complexity is also high.

The genetic Algorithm is a very optimal solution to the medical field for its Diagnosis. Genetic algorithm due to its fitness function can give accurate results and provides optimal solution.

REFERENCES:

[1] UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html.

[2] American Diabetes Association, "Standards of medical care in diabetes—2007," Diabetes Care, vol. 30, no. 1, pp. S4 S41, 2007.

[3] J. Du and C.X. Ling, "Active Learning with Generalized Queries," Proc. Ninth IEEE Int'l Conf. Data Mining, pp. 120-128, 2009

[4] Jiawei Han and Micheline Kamber, "Data Mining Concepts and techniques", 2nd ed., Morgan Kaufmann Publishers, San Francisco, CA, 2007.

[5] H.W. Ian, E.F., "Data mining: Practical machine learning Tools and techniques," 2005: Morgan Kaufmann.

[6] R. Detrano, A.J., W. Steinbrunn, M. Pfisterer, J.J. Schmid, S.Sandhu, K.H.Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for The diagnosis of coronary artery disease," American Journal Of Cardiology, 1989. 64: p. 304-310.

[7] G. John, "Models if incremental concept formation," Journal Of Artificial Intelligence, 1989: p. 11-61.

[8] A. L. Gamboa, M.G.M., J. M. Vargas, N. H. Grass, and R. E. Orozco, "Hybrid Fuzzy-SV Clustering for Heart Disease Identification," in Proceedings of CIMCA-IAWTIC'06. 2006.

[9] D. Resul, T.I., S. Abdulkadir, "Effective diagnosis of heart Disease through neural networks ensembles," Elsevier, 2008.

[10] Z. Yao, P.L., L. Lei, and J. Yin, "R-C4.5 Decision tree Model and its applications to health care dataset, in Proceedings of the 2005 International Conference on Services Systems and Services Management," 2005. p. 1099-1103

[11] R.Yuan, B.Guangchen,"Determination Of Optimal SVM Parameters by Using Genetic Algorithm/Particle Swarm Optimization", Journal of Computers, No.5, pp.1160-116, 2010

[12] *Ersen*, Y, An Expert System Based on Fisher Score and LS-SVM for Cardiac Arrhythmia Diagnosis" Computational and Mathematical Methods in Medicine, pp.1-6, 2013.

[13] Duygu, C. and Esin, a New Intelligent Hepatitis Diagnosis System: PCA LSSVM Expert Systems with Applications, 2011.

[14] Reference Paper of IJESRT for Medical Diagnosis for Decision Support System.