European Journal of Business & Social Sciences

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

Heart Disease Prediction System Using Data Mining Classification Techniques (K-Means, C4.5, Cart)

D.Bharathi,
M.Phil. Scholar, Department of Computer Science,
National College (Autonomous), Trichirappalli, Tamilnadu, India.
P.Sundari,M.Sc.,M.Phil.,SET.,
Asst. Professor, Department of Computer Science,
National College (Autonomous), Trichy Tamilnadu, India.

ABSTRACT

The Healthcare industry is generally "information rich", but unfortunately not all the data are mined which is required for discovering hidden patterns & effective decision making. Data mining techniques are used to notice knowledge in database and for medical research, mainly in Heart disease prediction. This paper has analyzed prediction system for Heart disease using more number of input attributes. The system uses medical terms such as sex, blood pressure, cholesterol Family history, Smoking , Poor diet , High blood pressure , High blood cholesterol , Obesity , Physical inactivity , Hyper tension etc like 13 attributes to predict the likelihood of patient getting a Heart disease. This research thesis added two more attributes i.e. obesity and smoking. The data mining classification techniques, namely K-Means, Cart, C4.5 are analyzed on Heart disease database. The show of these techniques is compared, based on accuracy. As per our results accuracy of C4.5, Cart, and K-means respectively. Our analysis shows that out of these three datamining models c4.5 predicts Heart disease with highest accuracy.

I. INTRODUCTION

Data mining is a process of extracting interesting knowledge or patterns from large databases. There are several techniques that have been used to discover such kind of knowledge, most of them resulting from machine learning and statistics. The greater part of these approaches focus on the discovery of accurate knowledge. Though this knowledge may be useless if it does not offer some kind of surprisingness to the end user. The tasks performed in the data mining depend on what sort of knowledge someone needs to mine. Data mining technique are the result of a time-consuming process of study and product development.

The heart is important organ of human body part. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it.

A number of factors have been shown that increases the risk of Heart disease: Family history, Smoking, Poor diet, High blood pressure, High blood cholesterol,Obesity,Physical inactivity, Hyper tension, Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on patient's current test results & doctor's experience. Thus the diagnosis is a complex task that requires much experience & high skill.

II LITERATURE SURVEY

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

In [1] discuss the show analysis of classification data mining technique for heart disease prediction. The three algorithms used in this work are Naïve Bayes, WAC and Apriori. The performance valuation is based on classification matrix, it display the frequency of correct and incorrect prediction. The analyze model is view in Lift charts, Bar charts and Pie charts. This struggle can be further enhanced and expanded by using other data mining techniques like Time series, Clustering and Association rules. Instead of categorical data, the continuous data can be used.
In [4] the fuzzy specialist system is planned for heart disease diagnosis with reduced number of attribute. The author find that how genetic algorithm and fuzzy logic combine together for efficient and cost effectual diagnosis of heart disease. The GA and two models of fuzzy system Mamdani and Takagi-sugeno were used to find the cost. The dataset were taken in Cleveland clinic establishment dataset. The input field is a set of all the certain kind and the output of the system is to get a value '1' or '0' that indicates the presence or absence of disease. It is additional improved by using art classifiers like Decision tree, Naïve bayes, Classification via clustering and SVM classifier.
In [5] data mining classification technique namely RIPPER classifier, decision tree, Artificial Neural Network and Support vector machine are used for predicting cardiovascular heart disease. The arrangement factor used for compare these technique are sensitivity, accuracy, specificity, error rate, true positive rate and false positive rate. To gauge the unbiased estimate of prediction models the author used 10 fold cross validation method. This model was industrial by using data mining classification tool weka version 3.6. It contains 14 attributes and 303 instances. Error rates for RIPPER, Artificial Neural Networks, Support vector machine and Decision tree are 02.756, 0.2248, 0.1588 and 0.2755 respectively. The accuracy of RIPPER, Artificial Neural Networks, Support Vector Machine and Decision tree are 81.10%, 80.07%, 84.13% and 79.15% respectively. Four classification models, the Support Vector Machine has given least error rate and highest accuracy. The writer concluded that the Support Vector Machine is the best technique for predicting the cardiovascular disease. In future in order to improve the efficiency of the classification techniques by creating meta models.

In [6] heart attack symptom is predicted using biomedical data mining technique. The creator used data classification which is based on supervised machine learning algorithms. For data classification the Tanagra tool is used. Using entropy based cross validations and partitioned techniques, the data is evaluated and the results are compared. The algorithms used in these techniques are Knearest neighbours, K-means and Mean Clustering Algorithm (EMC) is the extension of the K-mean algorithm for clustering process which reduces the number of iterations. In this paper result the author analyzed that the mean clustering algorithm perform well when compare to other algorithm. To run the data the time taken is very fast and it gives the result of accuracy about 83.25%. An enhanced by applying unsupervised machine learning algorithm.

In [8] association rule mining method is used for predicting heart attack. In this paper creator future a novel method CBARBSN for
association rule mining based on sequence numbers and clustering the transactional database for predict heart attack. The two important step of this process first the medical data is transformed into binary and the planned method is applied to the binary transactional data.

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

The data is collected from Cleveland database. The medical data contains 14 attributes. From the results author concluded that the proposed algorithm performs better than the existing ARNBSN ( ) algorithm. The performance of the algorithms is compared based on the execution time.

In [9] the coronary artery disease was efficiently diagnosed by rotation forest algorithm in order to support clinical decision-making process. It utilizes the Artificial Neural Networks with Levenberg-Marquardt back propagation algorithm of rotation forest ensemble method as base classifier. The algorithm is implemented in matlab. From an experiment, the author diagnosed the disease by comparing the performance of base classifiers in terms of sensitivity, accuracy, AUC and specificity on two things i) without rotation forest classifier, the greatest performance of classifiers and ii) with rotation forest algorithm it actually improve the performance of classifier. Result it is experiential that Levenberg-Marquardt was the best classifier with or without random forest. The accuracy is improved to 94.2% of original classification accuracy which is an improvement of 8%. In prospect the proposed work may be used to develop efficient expert systems for the diagnosis of heart disease.

.

III METHODOLOGY

3.1 K-MEANS Algorithm:

K-means cluster is a partition based clustering technique of classifying/grouping items into k groups (where k- is user specified number of clusters).

Algorithm: Original K-means(S, k), S= $\{x_1, x_2 \ldots x_n\}$.

Input: The no of clusters k and a dataset contain n objects x_i.

Output: A set of k clusters C_i that minimize the squared-error criterion.

Begin

1. m=1;
2. initialize k prototypes; //arbitrarily chooses k objects as the initial centers.
3. Repeat

for i=1 to n do

begin

for j=1 to k do

Compute $D(X_i, Z_j) = |X_i, Z_j|$;//$Z_j$is the center of cluster j.

if $D(X_i, Z_j) = \min\{D(X_i, Z_j)\}$ then

$X_i \in C_j$;

end;// (re)assign each object to the cluster based on the mean

if m=1 then

$J_c(m) = \sum_{j=1}^{k} \sum_{X_i \in C_j} |X_i - Z_j|^2$

m=m+1;

for j=1 to k do

$Z_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i^{(j)}$; //(re)calculate the mean value of the objectsfor each cluster

$J_c(m) = \sum_{j=1}^{k} \sum_{X_i \in C_j} |X_i - Z_j|^2$; //compute the error function

4. Until $J_c(m) - J_c(m-1) < \zeta$

End

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

3.2 C4.5

The decision tree approach is more powerful for classification problems. There are two steps in this techniques building a tree & applying the tree to the dataset. There are many popular decision tree algorithms C4.5. From these C4.5 algorithm is used for this system. C4.5 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications. The C4.5 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

Inputs: R: a set of non-target attributes, C: the target attribute,
S: training data.
Output: returns a decision tree
Start
Initialize to empty tree;
If S is empty then
Return a single node failure value
End If
If S is made only for the values of the same target then
Return a single node of this value
End if
If R is empty then
Return a single node with value as the most common value of the target attribute values found in S
End if
D ← the attribute that has the largest Gain (D, S) among all the attributes of R
{djj = 1, 2, ..., m} ← Attribute values of D
{Sjwith j = 1, 2, ..., m} ←The subsets of S respectively constituted of djrecords attribute value D
Return tree whose root is D and the arcs are labeled by d1, d2, ..., dm and going to sub-trees ID3 (R-{D}, C, S1), ID3 (R-{D} C, S2), .., ID3 (R-{D}, C, Sm)E

3.3 CART

Decision Trees are commonly used in data mining with the objective of creating a model that predicts the value of a target (or dependent variable) based on the values of several input (or independent variables). In today's post, we discuss the CART decision tree methodology.
Classification And Regression Trees (CART) algorithm is a classification algorithm for building a decision tree based on Gini's impurity index as splitting criterion. CART is a binary tree build by splitting node into two child nodes repeatedly. The algorithm works repeatedly in three steps

1. Find each feature's best split. For each feature with K different values there exist K-1 possible splits. Find the split, which maximizes the splitting criterion. The resulting set of splits contains best splits (one for each feature).
2. Find the node's best split. Among the best splits from Step i find the one, which maximizes the splitting criterion.
3. Split the node using best node split from Step ii and repeat from Step i until stopping criterion issatisfied.As splitting criterion we used Gini's impurity index, which is defined for nodetas:

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

$$i(t) = \sum_{i,j} C(i|j)p(i|t)p(j|t),$$

3.4 WEKA TOOL

Weka is a landmark system in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time. The Weka or woodhen (Gallirallus australis) is an endemic bird of New Zealand. It provides many different algorithms for data mining and machine learning. Weka is open source and freely available. It is also platform-independent.

Advantages

As weka is fully implemented in java programming languages, it is platform independent & portable.

It is freely available under GNU General Public License.

Weka s/w contain very graphical user interface, so the system is very easy to access.

There is very large collection of different data mining algorithms.

Disadvantages

Lack of possibilities to interface with other software

Performance is often sacrificed in favour of portability, design transparency, etc.

Memory limitation, because the data has to be loaded into main memory completely

IV. EXPERIMENT AND RESULTS

In this study mainly classification algorithms such as K-means and CART algorithm is used for predicting the Heart Disease from the given data set instances and the proposed algorithms are applied on type Heart Disease dataset in the WEKA tool and the performance is measured. The heart is main organ of human body part. It is nothing more than a pump, which pump blood through the body. If circulation of blood in body is incompetent the organ like brain suffers and if heart stops working altogether, death occurs within minutes. The term Heart disease refers to disease of heart & blood vessel system within it. A number of factors have been shown that increases the risk of Heart disease: Family history, Smoking , Poor diet , High blood pressure , High blood cholesterol , Obesity , Physical inactivity , Hyper tension etc.,Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on patient's current test results & doctor's experience. Thus the diagnosis is a complex task that requires much experience & high skill.

4.1 DATA SOURCE

The publicly available heart disease database is used. The Cleveland Heart Disease database consists of 1000 records & Statlog Heart Disease database consists of 970 records [12]. The data set consists of 3 types of attributes: Input, Key & Predictable attribute which are listed below.

Attribute Information:

Only 14 attributes used:

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

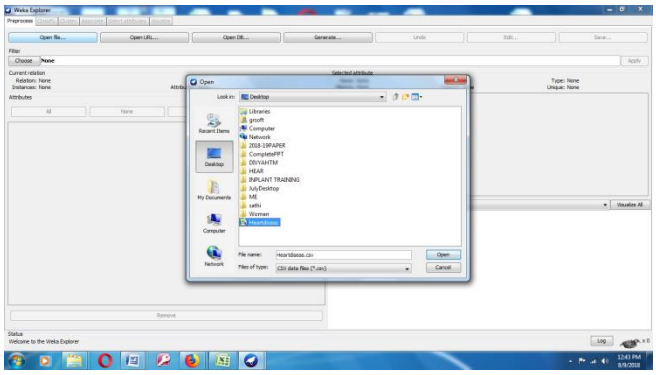


Figure 4.1.1 Upload Heart Disease Dataset

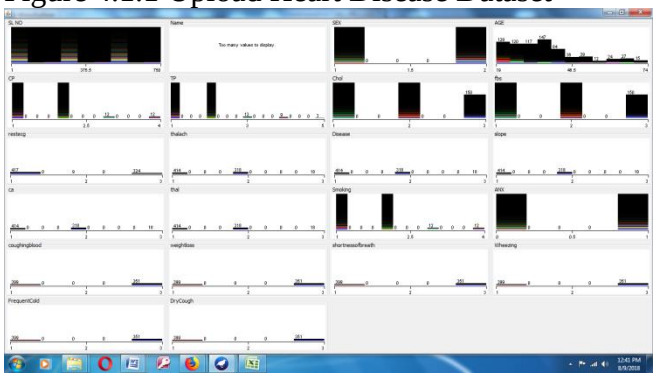


Figure 4.1.2 View Heart Disease Chart

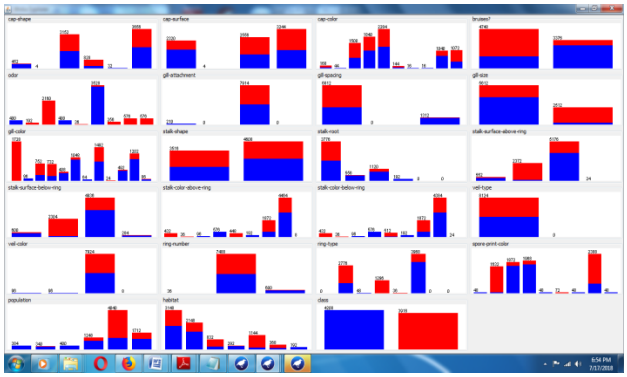
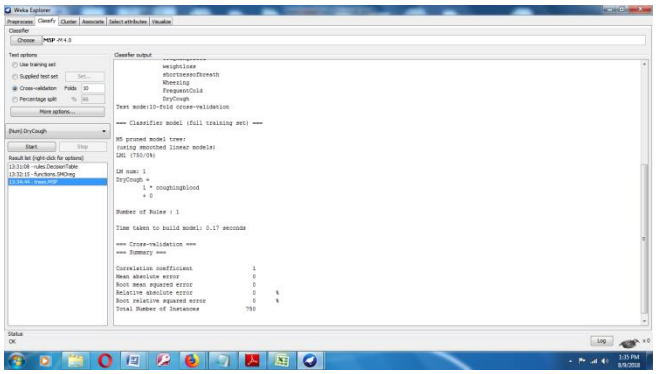


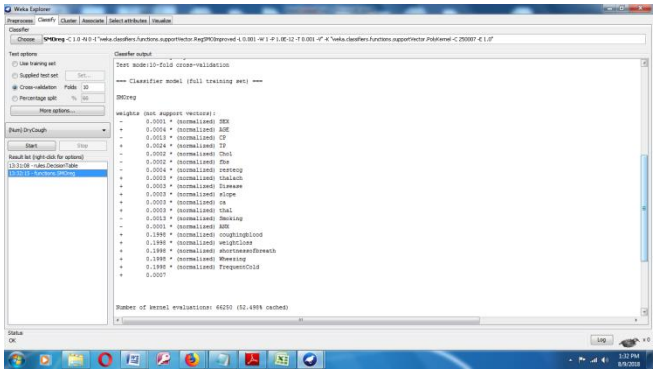
Figure 4.1.3 Over All Chart

European Journal of Business & Social Sciences

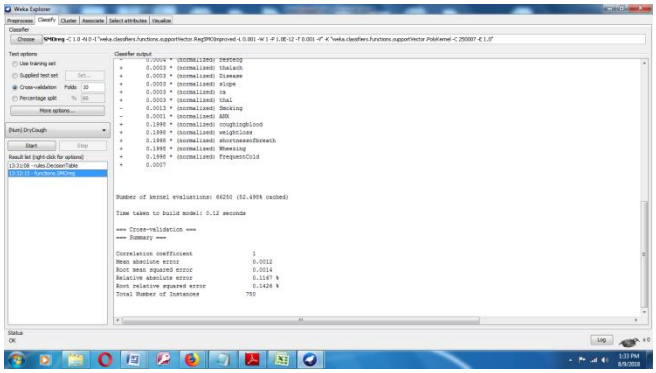
ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue



4.1.4 K-Means Heart Disease Dataset Execution Second



4.1.5 C4.5 Heart Disease Dataset Execution Second



4.1.6 CART Heart Disease Dataset Classification Techniques

4.2 Performance Calculated Using Correlation Coefficient

Experiments are performed on the heart disease datasets. Machine with configuration of windows 7 system and 2-GB of RAM is used. The results were compared to experiments with WEKA implementations of K-Means, CART and C4.5 the techniques run to ensure that the results are comparable for Correlation coefficient.

Table 4.2.1: The table shows Correlation coefficient

ALGORITHM	CORRELATION COEFFICIENT
K-Means	0.7485
CART	0.8961
C4.5	1

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

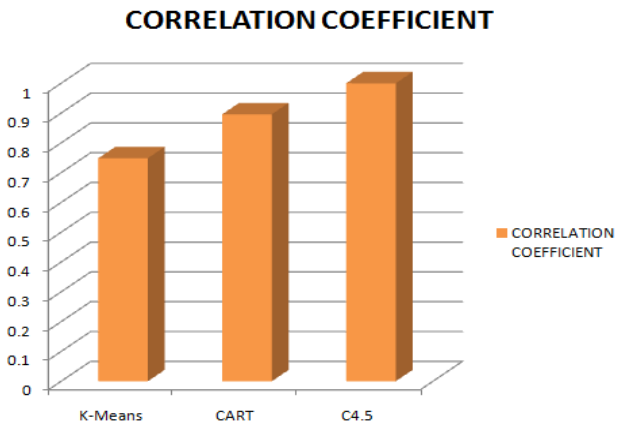


Fig 4.2.1 Evaluation measures of Correlation coefficient

4.2.2 Relative Absolute Error

Experiments are performed on the heart disease datasets. Machine with configuration of windows 7 system and 2-GB of RAM is used. The results were compared to experiments with WEKA implementations of K-Means, CART and C4.5 the techniques run to ensure that the results are comparable for Relative absolute error.

Table: 4.2.2 The table shows relative absolute error

ALGORITHM	RAE
K-Means	0.7632
CART	0.6971
C4.5	0.1707

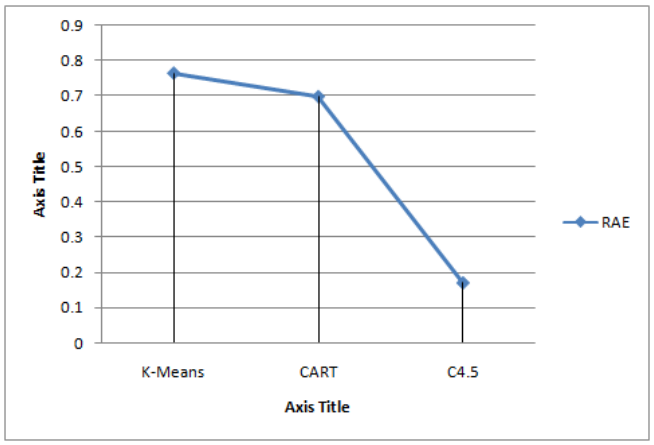


Fig: 4.2.2 Evaluation measures of Relative absolute error

4.2.3 EXECUTION TIME SECOND

Experiments are performed on the heart disease. Machine with configuration of windows 7 system and 2-GB of RAM is used. The results were compared to experiments with WEKA implementations of k-Means, Cart and C4.5 the techniques run to ensure that the results are comparable for Execution Time Second.

Table 4.2.3 Execution Time Second

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

ALGORITHMS	TIME TAKEN (IN MIL SECS)
K-Means	0.09
CART	0.17
C4.5	0.02

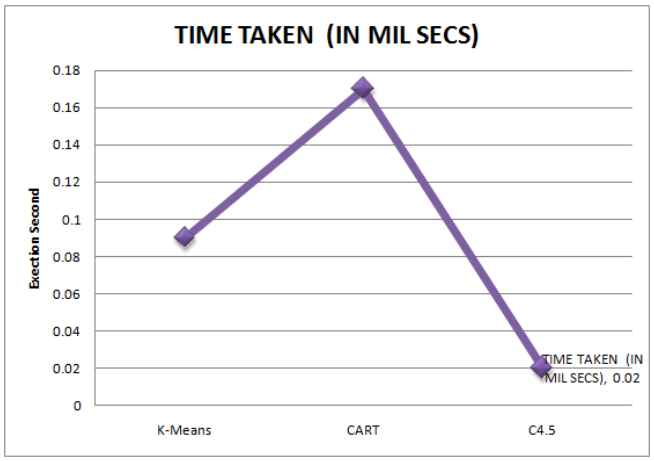


Figure: 4.2.3 Representing Dataset measured in Execution Time Second (ms)

V. CONCLUSION

Around 18 million people 7% of the Indians are affected by heart disease. Heart disease is mostly affected the person under the age of 65. This paper mainly focuses on three different categories of heart diseases namely cardiovascular disease, coronary artery disease and cardiomyopathy. The overall objective of work is to predict more accurately the presence of heart disease. In this thesis more input attributes obesity and smoking are used to get more accurate results. Three data mining classification techniques were applied namely CART, K-MEANS & C4.5. From results it has been seen that C4.5 provides accurate results as compare to CART & K-Means.

REFERENCES

[1] Frawley and G. Piatetsky -Shapiro, Knowledge Discovery in Databases: An Overview. Published by the AAAI Press/ The MIT Press, Menlo Park, C.A 1996.

[2] Yanwei, X.; Wang, J.; Zhao, Z.; Gao, Y., "Combination data mining models with new medical data to predict outcome of coronary heart disease". Proceedings International Conference on Convergence Information Technology 2007, pp. 868 – 872.

[3] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

[4] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1 (January - June 2007).

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

[5] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.
[6] Shantakumar B.Patil, Y.S.Kumaraswamy "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network". ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656.
[7] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.
[8] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer, Vol:345, pp: 721- 727, 2006.
[9] Franck Le Duff, Cristian Munteanb, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics, Vol. 107, No. Pt 2, pp. 1256-9, 2004.