ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Data Mining For Prediction of Disease Domains

P.Jagdish Kumar
Department of Computer Science and Engineering,
AVN Institute of Engineering and Technology,
Koheda, Ibrahimpatnam, T.S-501510, India.

ABSTRACT

Due to the sensitivity of the information required to treat a disease of a patient efficiently, the healthcare industry collects huge amounts of healthcare data, the volume and the details of this data increased in recent years, which is due to the patient information acquiring strategies called Electronic Health Records (EHR) and Electronic Medical Records (EMR). As Data mining is vital and important in an exploratory analysis because of nontrivial information in large volumes of data, it plays an important and key role in health care industry. In particular data mining is having vital scope in health care industry needs such as disease prediction, estimation of scenarios that leads to possibility of a disease. In this regard we explored a statistical mining model called disease prediction by degree of disease possibility, which is aimed to estimate the degree of disease possibility threshold from given electronic medical records for training. With the motivation gained from our model that explored in our earlier work, here we devised a novel and improved mining strategy for disease prediction. In this mining strategy a feature correlation is considered to estimate the degree of disease possibility. The experimental results indicating that the feature correlation is having significant impact towards disease prediction by degree of disease possibility.

Keywords: ***Health Mining, Associability, Disease prediction, HITS algorithm, weighted associative classifiers***

1. INTRODUCTION

Data mining is the identification of formerly unidentified, implicit information or facts having immediate implications from enormous quantities of raw data or large volumes of databases [1]. It is a process of mutual interaction of human and computers and a strategic balance of the human skills defining the problems and goals combined with the computers search and analysis abilities to produce effective outcomes.

Data mining is useful for examining routinely collected large amounts of data. It is most useful for exploratory analysis of large volumes of data containing nontrivial information. The objectives of data mining are prediction and description. Prediction is used to find unfamiliar or unidentified variables applying known variables on the data set and Description primarily gives insights into actual, expected and obvious details in a familiar way for humans to understand. Data mining performs a crucial role in Disease Prediction. The technique helps in forecasting several diseases like Hepatitis, Lung cancer, Liver disorder, and Breast cancer, Thyroid disease, Diabetes and many more.

Patient's data recorded by medical professionals is very useful to find information that is relevant and is of immediate application. Techniques such as statistical analysis and data mining, studied widely by researchers today, are primarily focused on helping medical professionals in predicting diseases especially heart related diseases. Several such analyses categorize heart disease by factors such as age, obesity, lack of physical activity [2], smoking habit [3], blood pressure, hypertension, total cholesterol [4], diabetes [5], and family history [6]. The criteria help medical professionals in giving timely diagnosis and treatment to patients with high risk factors.

Cite this article as: P.Jagdish Kumar, "Data Mining for Prediction of Disease Domains", International Journal & Magazine of Engineering, Technology, Management and Research, Volume 5 Issue 1, 2018, Page 63-71.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Decision tree, naïve bayes, neural network, bagging, kernel density and support vectl or machine are some of the statistical techniques being investigated by researchers on various heart disease datasets [7]-[12]. They have showed in [13] correct classified accuracy of approximately 77% with logistic regression and comparatively more effective accuracy is observed by applying model R-C4.5 based on C4.5. Rules generated for this model C4.5 provides medical experts with better insights into the disease [14].

The common interest of all of these said models is the feature selection and feature optimization towards scalable predictions. Hence here in this paper, we devised a statistical mining model to predict the degree disease possibility using the correlation of the features extracted from given patient record, which is an extension to our earlier model [15].

2.RELATED WORK

Jyoti Soni et. al [16] introduced three types of supervised machine learning algorithms for studying the heart problems dataset [17]. The data is classified using Tanagra data mining tool evaluated with 10 fold cross validation and comparing the outcomes. The algorithms proposed are Decision List algorithm, Naive Bayes and K - NN. The algorithms details are as below;

i)Decision tree requires simple steps for initiation, handles large amount of dimensional information, is perfect for exploratory knowledge discovery and finally the results gained from Decision Tree are easier to read and interpret [1] resulting it to be used widely.

ii)Naive Bayes is a statistical classifier where the characteristics have no dependency assigned to them, requires posterior probability to be maximized for identifying the group and does not require any Bayesian approach for execution. The efficiency in execution is high and the algorithm is simple compared to all other machine learning algorithms.

iii) k-NN algorithm has loud features that highly reduces the performance and this conclusion is reached after training and executing 3000 instances with 14 different attributes.

Jyoti Soni et a; [16] additionally proposed connection rule data-mining technique for predicting one's heart disease. However they introduced varied rules in large number for organization guidelines used over health related datasets, most of them clinically negligible for the data. Subsequently to decrease the number of guidelines the authors implemented four restrictions i.e., item filtering, feature group, optimum product set size as well as antecedent/consequent principle restrictions. Another problem was without relevance the organization rules are applicable to the entire data-set and to overcome these limitations and to reduce the large number of guidelines another algorithm was introduced. In the new algorithm the verification of the validation outcome is done with set queries, organization guidelines and evaluation set. A new parameter "lift" is applied in place of help and assurance and another parameter Raise is used for measuring and assessing the dependability and medical importance of organization guidelines. The validation of the results by physicians is defined by susceptibility and specificity. The accuracy of determining high risk patients based on chance and susceptibility is further defined by the correct determination of a healthy individual based on uniqueness.

There are three steps in the formula for determining the predictive association rules in health-related dataset [18] where;

(i) Conversion of medical data-set comprising of numeric and categorical attributes into trade data-set should be done.

(ii) Predictive association guidelines with medically valid features should be obtained by incorporating all previously reviewed four constraints into the search process.

(iii) Test and train strategy for validating the organization rules ought to be utilized.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Genetic algorithms reduce [19] the original size of data to a most effective sub-set of sufficient size, enough for forecasting heart disease. Classification is a strategy of supervised learning to model datasets of valid definition.The most important role of feature selection is to minimize the dimensionality and improve classification. Data mining in medical domain is not the same as in that of other domains due to multiple attributes.

Generally predicting a disease involves diagnosis with multiple tests that usually incur huge costs, data mining techniques in contrast analyzes specific attributes of the disease and forecasts the disease probability and eliminates the requirement of most of the tests and its expenses. Here in healthcare domain the attributes are many and the approach of dimensionality reduction is a major strategy for minimizing the number of attributes and accurately predicting the diseases. The feature correlation based heart rate variability was checked in [21], which was done by a classification approach that supported by statistical analysis on correlated categorical features. The experiments explored in [21] shown that the usage of support vector machine (SVM) as a classifier delivered accuracy in classification and operational process. The heart disease prediction using classifiers such as Decision Trees, Navy Byes and Nearest Neighbor were explored in [22]. The results illustrated that the classification accuracy is proportionate to feature optimization. In this regard we devised a mining statistical mining method in our earlier work [15], which is in the aim of predicting the scope of the disease by degree of disease possibility.

This model is estimating the impact of the each feature as individual towards disease possibility, but in vigorous experimental analysis, the consistency of the model is violated as more than one feature together reflecting divergence in estimating the degree of possibility, which is due to their categorical values. Henceforth here in this paper we explored a novel model that attempts to estimate the impact of multiple features together to explore the degree of disease possibility.

3.Disease Prediction by Estimating Feature Correlation Towards Degree of Disease Possibility

A metric called closeness support was devised in our earlier model [15], which is aimed to estimate the impact of each feature towards degree of disease possibility. In this regard we considered a set of medical records of divergent heart disease patients and set of features representing the values in those medical records. Further a bipartite graph was build between features and medical records, and then a refined HITS algorithm was applied to estimate the waits of the features and records. Further these values were used to measure the degree of disease possibility. Experimental results on benchmark dataset also indicated that the model is effective towards predicting the disease possibility. But the experimental results on increasing set of medical records indicating that often the disease possibility estimation is contradict in regard to more than one feature as a set, which is due to the categorical values of these features. Henceforth here we refined the said model such that the correlated feature-set will be taken as one unit rather than an individual feature to estimate the degree of disease possibility threshold. In this DDP-FC proposal a bipartite graph is build between medical records and set of correlated features.

3.1Estimating Correlation between Features:

Pearson correlation coefficient [20] and mean-square contingency coefficient [20] are two bench mark models to assess the correlation between any two attributes with continuous and categorical values respectively. As described in our earlier work [15] the attributes taken in to account of disease prediction are categorical. Henceforth here we use mean-square contingency coefficient [20] to estimate the correlation between attributes. A and B such that $\{a_1, a_2, a_3, \ldots..a_m\}$, are categorical values of A and B respectively. The size of the set of values appeared for A is m and B is n. Then the mean square contingency coefficient between attributes A and B can be measured as follows:

$$\rho_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n} 1 - \frac{1}{o(a_i, b_j)} \quad \ldots. \text{(Eq1)}$$

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Here in this equation Eq1, ρ_{ij} is the fraction of co occurrence of a_i, b_j

$$\rho_i = \sum_{i=1}^{m} 1 - \frac{1}{o(a_i)} \ldots. \text{(Eq2)}$$

Here in this equation Eq2, ρ_i is the fraction of occurrence of a_i

$$\rho_j = \sum_{j=1}^{n} 1 - \frac{1}{o(b_j)} \ldots.. \text{(Eq3)}$$

Here in this equation Eq3, ρ_j is the fraction of occurrence of b_j

$$\chi^2_{(A \leftrightarrow B)} = \frac{1}{\min(m,n)-1} * \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(\rho_{ij} - (\rho_i . \rho_j))^2}{\rho_i . \rho_j} \ldots.. \text{(Eq4)}$$

Here in this equation Eq4, $\chi^2_{(A \leftrightarrow B)}$ is the mean square contingency coefficient that indicates the correlation between attributes A and B.Since the attributes used in disease prediction are mostly contains values that are categorical, henceforth k-medoid clustering technique can be used to group the attributes based on their correlation.

3.2Degree of Disease Possibility by Feature Correlation (DDP-FC)

A.Assumptions:

Let set of medical records $mr_1, mr_2, mr_3, \ldots\ldots mr_n$

Formed by the set of features $\left\{ FS \forall \bigcap_{i=1}^{fc} f_i \in FS \right\}$ …. (Eq5)

Here in the above equation feature set FS contains all features $\{f_1, f_2, \ldots.. f_n\}$ that are used to form each medical record.

Let $SCRFS$ as set of correlated feature sets such that each correlated feature set $\{crfs \forall crfs \in SCFRS\}$

The property of a correlated feature set is $\left\{ \bigcup_{i=1}^{|crfs|} \{ f_i \in crfs \forall crfs \subseteq FS \} \right\}$ …. (Eq6)

Here in above description $SCRFS$ represents the set of correlated feature sets found by using the approach described in section 3.1. $crfs$ is a correlated feature set which is belongs to $SCRFS$ and it represents set correlated features such that these features are subset of FS.

Process

In the process of detecting the closeness of each correlated feature set with medical records, initially we build a bi-parted graph between medical records and all correlated feature sets.

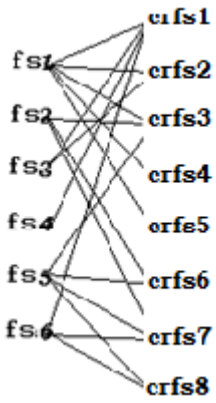

Fig 1: bipartite graph between medical records and correlated feature sets

If an edge found between a medical record mr_i and a correlation feature set $crfs_j$ then the weight of that edge $ew_{(mr_i \leftrightarrow crfs_j)}$ can be found as follows

$$ew_{(mr_i \leftrightarrow crfs_j)} = \frac{|\, mr_i \cap crfs_j \,|}{|\, mr_i \,|} \ldots. \text{(Eq7)}$$

Let consider a set of medical records $\left\{ MR \forall \bigcup_{i=1}^{|MR|} mr_i \in MR \right\}$ as a database such that each medical record mr contains one or more features that belongs to feature set FS. Then the medical records MR can be represented as bipartite graph G as

$$G = (MR, FS, E)$$

Here $E = \{(mr, crfs) \forall crfs \in SCRFS, crfs \subseteq mr, mr \in MR\}$

The graph representation (see fig 1) of the set of medical records gives us the idea of applying link-based ranking

models for the evaluation of connected sets. In this bipartite graph, the closeness support of a medical record is proportional to degree of all its correlated feature sets weight. However, it is crucial to have different closeness weights for different medical records in order to reflect their different importance. The evaluation of medical record influence should be derived from these weights.

Here comes the question of how to acquire weights in a set of medical records. Intuitively, a medical record with high closeness weights should contain many of the correlation feature sets those belongs to the same medical record with high closeness support; at the same time, a medical record with high closeness support should be contained by less or zero other medical records with high closeness weights. The reinforcing relationship of medical records and correlation feature sets is just like the relationship between hubs and authorities in the HITS model [23].

Regarding, the medical records as "pure" hubs and the correlation feature sets as "pure" authorities, we can apply HITS to this bipartite graph. The following explored the process:

Table 1: matrix A as follows that represents the connection weights between a feature and each medical record mr

	$crfs_1$	$crfs_2$	$crfs_3$	$crfs_4$	$crfs_5$	$crfs_6$	$crfs_7$	$crfs_8$
m	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$
m	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$
m	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$
m	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$
m	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$
m	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$	$ew_{(m}$

Table 2: Transpose matrix of matrix as fallows that represents the connection between a case and each medical record .

	mr_1	mr_2	mr_3	mr_4	mr_5	mr_6
$crfs$	$ew_{(mr_1}$	$ew_{(mr_2}$	$ew_{(mr_3}$	$ew_{(mr_4}$	$ew_{(mr_5}$	$ew_{(mr_6}$
$crfs$	$ew_{(mr_1}$	$ew_{(mr_2}$	$ew_{(mr_3}$	$ew_{(mr_4}$	$ew_{(mr_5}$	$ew_{(mr_6}$
$crfs$	$ew_{(mr_1}$	$ew_{(mr_2}$	$ew_{(mr_3}$	$ew_{(mr_4}$	$ew_{(mr_5}$	$ew_{(mr_6}$
$crfs$	$ew_{(mr_1}$	$ew_{(mr_2}$	$ew_{(mr_3}$	$ew_{(mr_4}$	$ew_{(mr_5}$	$ew_{(mr_6}$
$crfs$	$ew_{(mr_1}$	$ew_{(mr_2}$	$ew_{(mr_3}$	$ew_{(mr_4}$	$ew_{(mr_5}$	$ew_{(mr_6}$
$crfs$	$ew_{(mr_1}$	$ew_{(mr_2}$	$ew_{(mr_3}$	$ew_{(mr_4}$	$ew_{(mr_5}$	$ew_{(mr_6}$
$crfs$	$ew_{(mr_1}$	$ew_{(mr_2}$	$ew_{(mr_3}$	$ew_{(mr_4}$	$ew_{(mr_5}$	$ew_{(mr_6}$
$crfs$	$ew_{(mr_1}$	$ew_{(mr_2}$	$ew_{(mr_3}$	$ew_{(mr_4}$	$ew_{(mr_5}$	$ew_{(mr_6}$

Let matrix representation of medical records and correlated feature sets as a matrix 'A'(see Table 1). The value represents the edge weight between medical record and correlation feature set that calculated by using Eq7. Consider the matrix that representing each hub initial value as 1 (see fig 2).

Fig 2: Initially consider the each hub weight as 1 by default as fallow and represent them as matrix u.

1
1
1
1
1
1

Transpose the matrix A as A'(see Table 2)

Find Authority weights by multiplying A' with u as $v = A' \times u$ (Matrix multiplication between A' and u gives a matrix v that representing the authority weights)

Now find the original hub weights through matrix multiplication between $A and v$.

$$u = A \times v$$

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Then the Cs of correlation feature set $crfs$ can be measured as follows

$$cs(crfs) = \frac{\sum_{i=1}^{m} \{u(mr_i) : (crfs \rightarrow mr_i) \neq 0\}}{\sum_{i=1}^{m} u(mr_i)}$$

$$ddp(mr_i) = 1 - \frac{\sum_{j=1}^{m} \{cs(crfs_j) : (crfs_j \rightarrow mr_i) \neq 0\}}{ec(mr_i)}$$

Here in the above equation $ec(mr_i)$ represents the total number of edges connected to medical record mr_i

Then the degree of disease possibility threshold can be found as follows:

$$ddpt = \frac{\sum_{i=1}^{|MR|} ddp(mr_i)}{|MR|}$$

Here in the above equation $|MR|$ indicates the total number of medical records

The degree of disease possibility range can be explored as follows

Lower threshold of $ddpt$ range is

$$ddpt_l = ddpt - \left(\frac{\sum_{i=1}^{|MR|} dv(mr_i)}{|MR|} \right)$$

Here in this equation $dv(mr_i)$ indicates the deviation of the ddp value of medical record mr_i

Higher threshold of $ddpt$ range is

$$ddpt_h = ddpt + \left(\frac{\sum_{i=1}^{|MR|} dv(mr_i)}{|MR|} \right)$$

Medical record mr can be said as safe if and only if

$$ddp(mr) < ddpt_l$$

Medical record mr can be said as disease possibility risk is high if and only if

$$ddp(mr) \geq ddpt_l \;\&\&\; ddp(mr) < ddpt_h$$

Medical record mr can be confirmed as effected by disease if $ddp(mr) \geq ddpt_h$

4.EMPIRICAL ANALYSIS OF THE PROPOSED MODEL

We explored the credibility of the proposed model on heart disease dataset.

The above said data set contains 5136 samples, out of that 4136 samples were used to devise the degree of disease possibility threshold and its upper and lower bounds. Further we used the rest 1000 records to predict the disease scope. Interestingly, the empirical study delivered promising results. The statistics explored in Table 3

Table 3: Statistics of the experiment results

Total Number of Records	5136
Range of fields count in a record	14
Total number of Features Found in dataset	174
Total number of correlation sets found	24
Total number of bipartite edges found	74448
Degree of Disease possibility threshold found:	0.2442175329126821
Degree of Disease possibility threshold Upper Bound	0.3121509592964571
Degree of Disease possibility threshold Lower Bound	0.1920588722068104

Total records Tested 1000

Total number of records found with DDP less than lower bound 23 (false negative 3 and true negative 20)

Total number of records found with DDP greater than lower bound 977 (true positives 956 and false positives 21)

As per the results explored in Table 3, the proposed model is accurate to the level of 81.7%. The failure percentage is 18.3%, which is nominal.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The experiments also conducted on the same data set with earlier method which is not considering the correlation factor of the features, and the results are as follows:
Total records Tested 1000
Total number of records found with DDP less than lower bound 292 (false negative 205, true negatives 87)

Total number of records found with DDP greater than lower bound 708 (false positives 134 true positives 574)
As per these results, the accuracy of the degree of disease possibility without correlation feature set factor is less significant since we observed that the prediction success limited to 70.0%. The failure percentage is approx 30%, which is not a negligible factor.

Hence it is obvious to conclude that the correlation feature sets are more significant compared to independent features towards to measure the degree of disease possibility

Performance Analysis

We used disease prediction accuracy (the percentage of valid predictions by the proposed) as the main performance measure. In addition to measuring accuracy, the precision, recall, and F-measure were used to measure the performance; these are defined using following equations.

$$pr = \frac{t_+}{t_+ + f_+}$$

Here in above Equation the pr indicates the precision, t_+ indicates the true positives and f_+ indicates the false positive

$$rc = \frac{t_+}{t_+ + f_-}$$

Here in above Equation, the 'rc' indicates the recall, 'f_-' indicates the false negative.

$$F = \frac{2 * pr * rc}{pr + rc}$$

Here in the above Equation, 'F' indicates the F-measure.

Table 4: Precision, recall and F-measure values found from the results of the empirical analysis.

	Precision	recall	f-measure
DDP	0.810734	0.868381	0.886754
DDP-FC	0.978506	0.979508	0.920603

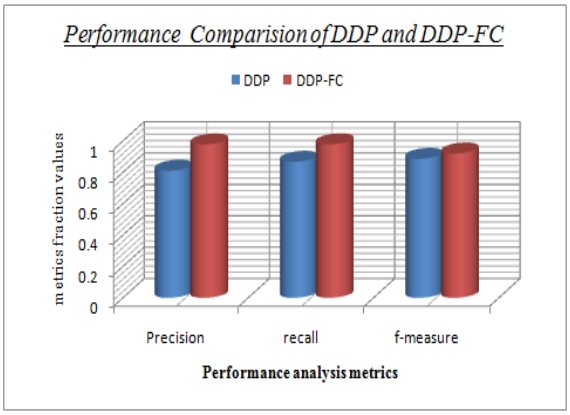

Fig 3: performance analysis of the estimating Degree of Disease Possibility with and without Correlation analysis of the features.

CONCLUSION:

Correlation analysis of the features and clustering them, which is based on their mean square contingency correlation coefficient, is a significant approach towards improving the accuracy of measuring degree of disease possibility. The vigorous experiments with divergent size of the medical records, it is observed that our earlier mining approach [15] that measures degree of disease possibility threshold is not significant against divergence in feature values. Henceforth here in this paper we devised a feature correlation analysis by mean square contingency threshold, which is used further to cluster these feature by k-medoid approach. Then these clusters with highly correlated features are used along with medical records to build a bipartite graph. The experimental results indicating that the model devised here in this paper is significantly improved the accuracy of measuring degree of disease possibility threshold compared to our earlier model [15]. In future the cluster refinement and optimization can be devised, which is necessary in situation if maximal number of clusters found due to broad range of categorical values of the features.

REFERENCES

[1] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006

[2] Data mining: Introductory and Advanced Topics" Margaret H. Dunham

[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" IJCSE Vol. 3 NO. 6 June 2011

[4] Carloz Ordonez, "Association Rule Disco very with Train and Test approach for heart disease prediction", IEEE Transactions on Information Technology in Biomedicine, Volume 10, No. 2, April 2006.pp 334-343.

[5] M. ANBARASI, E. ANUPRIYA, N.CH.S.N. IYENGAR, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376

[6] G. Parthiban, A. Rajesh, S.K.Srivatsa "Diagn osis of Heart Disease for Diabetic Patients using Naive Bayes Method"

[7] Choi J.P., Han T.H. and Park R.W., "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis", J Korean Soc Med Inform, 2009, pp. 49-57

[8] Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques,"Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining," 2006.

[9] Lundin M., Lundin J., BurkeB .H.,Toikkanen S., Pylkkanen L. and Joensuu H. , "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer", Oncology International Journal for Cancer Resaerch and Treatment, vol. 57, 1999.

[10] Delen Dursun , Walker Glenn and Kadam Amit , "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine ,vol. 34, pp. 113-127 , June 2005.

[11] Ruben D. Canlas Jr.,"Data Mining In Healthcare: Current Applications And Issues", August 2009

[12] Michael Feld, Dr. Michael Kipp, Dr. Alassane Ndiaye and Dr. Dominik Heckmann "Weka: Practical machine learning tools and techniques with Java implementations"

[13] Jon M. Kleinberg Authoritative Sources in a Hyperlinked Environment http://www.cs.com ell.edu/home/kleinber/auth.pdf

[14] K.P Soman, Shyam Diwakar, V.Vijay "Insight into Data mining theory and practice"

[15] Asha Rajkumar, G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm", Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver. 1.0 September 2010.

[16]Shantakumar B.Patil, Y.S.Kumara swamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.

[17]Wingo PA, Tong T, Bolden S, "Cancer statistics", 1995, CA Cancer J Clin 45 (1995), no. 1,8-30.

[18]Fentiman IS, "Detection and treatment of breast cancer", London: Martin Duntiz (1998).

[19] http://ftp.ics.uci.edu/pub/machine-learning-databases/heart- disease/reprocessed.hungarian.data

[20]Ramana .N, Dr.C.V.Guru Rao "Contem porary Affirmation of the Recent Literature on Disease

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Prediction Using Data Mining Techniques"; International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June 2013 ISSN 2229-5518.