Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection

Shweta Kharya
Bhilai Institute of Technology,
Durg C.G. India

Sunita Soni
Bhilai Institute of Technology,
Durg C.G. India

ABSTRACT
In this paper investigation of the performance criterion of a machine learning tool, Naive Bayes Classifier with a new weighted approach in classifying breast cancer is done . Naive Bayes is one of the most effective classification algorithms. In many decision making system, ranking performance is an interesting and desirable concept than just classification. So to extend traditional Naive Bayes, and to improve its performance, weighted concept is incorporated. Exploration of Domain knowledge based weight assignment on UCI machine learning repository dataset of breast cancer is performed. As Breast cancer is considered to be second leading cause of death in women today. The experiments show that a weighted naive bayes approach outperforms naive bayes.

Keywords
Data Mining, Breast cancer, Naive bayes classifier, Domain based weight, Weights, Posterior probability, UCI machine learning repository, Prediction.

1. INTRODUCTION
Data mining [11] is the set of techniques and tools applied to the non-trivial process of extracting and presenting implicit knowledge, previously unknown, potentially useful and humanly comprehensible from large datasets[2, 8].Medical data mining has great prospective for exploring hidden patterns in data sets. The applications of data mining in medical and health research have proved itself to be effective, showing great development potentialities. The model created in data mining can be Predictive and Descriptive in nature. A predictive model makes a prediction about values of data using known results found from different data. In this work classification technique of predictive model is used. Classification is supervised learning which maps data into predefined groups or classes. *Prediction* involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. *Description,* on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data-mining activities into one of two categories.

Breast cancer is the most common cancer in women worldwide [6]. It is also the principle cause of death from cancer among women globally. The most effective way to reduce breast cancer deaths is its early detection. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. Breast cancer is the most frequently diagnosed cancer and is the leading cause of cancer death among women worldwide [7].

- Every 19 sec, somewhere around the world a case of breast cancer is diagnosed among women.
- Every 74 sec, somewhere in the world, someone dies from breast cancer.

In this paper evaluation of the performance of weighted naive bayes classifier model using standard UCI datasets for prediction of presence of breast cancer is build. Naive Bayes is one of the most effective classification algorithms. And to improve its performance, ranking of attributes by assigning weights is the interesting idea [4]. And GUI is designed o accept the patient's screening result and predict the probability of having breast cancer in her future with more accuracy.

2. RELATED WORK
In Naive Bayes Classifier [3] has been applied to Wiscosin Prognostic Breast Cancer(WPBC) dataset (UCI Machine Learning epository:http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/), concerning a number of 198 patients and a binary decision class :non-recurrent-events of 151 instances and recurrent-events of 47 instances .The input features contain 12 relevant attributes describing the characteristics of cell nuclei .The testing diagnosing accuracy was about 74.24% in accordance of other well known Machine Learning techniques.

In this paper authors present the comparison of different classification techniques [9] like Bayes Network, Radial Basis Function, Pruned Tree and Nearest Neighbors algorithm using Waikato to Environment for Knowledge Analysis (WEKA) on large dataset. The data used in their investigation is the breast cancer data. It has a total of 6291 data and a dimension of 699 rows and 9 columns. In this 75% of overall data is used training and the rest is used for testing the accuracy of classification technique. According to the simulation result, highest accuracy is 89.71% which belongs to bayes network with minimum time taken to build the model is 0.19 seconds and lowest average error is 0.2140 compared to others.

In this paper authors analyze the performance of supervised learning algorithm such as Naive Bayes, SVM Gaussian RBF kernel, RBF neural networks, Decision tree J48 and simple CART [13] .These algorithm are used for classifying the breast cancer datasets WBC, WDBC, Breast tissue from UCI Machine learning Repository (http://archive .ics.uci.edu/ml) .They conducted their experiments using WEKA tool. In which the accuracy percentage of Naive Bayes algorithm for WBC dataset yields to be 96.50%,for Breast tissue dataset comes to be 94.33% and for WDBC dataset it is 92.61%.

Intelligent and Effective Heart Disease prediction system designed by the author [15] using weighted concept on Associative Classifiers provides improved accuracy as compared to other already existing associative classifiers. The accuracy found to be 81.51% for WAC.

An author proposes a framework, weighted associative classifier (WAC) that assigns different weights to different attributes according to their predicting capability. They used maximum likelihood estimation (MLE) theory to calculate weight of each attribute using training data.[14] Experiments have been performed on benchmark data set to evaluate the performance of WAC in terms of accuracy, number of rules generating and impact of minimum support threshold on WAC outcomes. The result reveals that WAC is a promising alternative in medical prediction and certainly deserves further attention.

In this paper authors [1] present an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. The data used is the SEER Public-Use Data. The preprocessed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. They have investigated three data mining techniques: the Naive Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. The results show the Naive Bayes accuracy found to be 84.5%.

In [16] authors extended the classical association rule mining paradigm by allowing weights to be associated with items in transactions in order to reflect the interest/intensity of items in transactions. For example, 70% of people buying more than four bottles of beers will also be likely to buy more than three packs of potato chips. In their approach, they first calculated frequent item sets without considering the weights of items and then introduced weight during the process of rule generation. In particular, they segment the domain weight space of each frequent item set and then identify regions that contain transactions that are heavily populated with such segments in order to derive association rules. Authors demonstrated that their method not only improves the confidence of the rules, but also provides a mechanism for more effective targeted marketing by 21 categorizing customers on the basis of their level of loyalty or volume of purchases.

3. RESEARCH OBJECTIVES

The core objective of this research is to extend the traditional naive bayes classifier with a novel approach of weights assignment to its attributes and to develop more accurate probabilistic classifier for Breast cancer detection system that can be used by experts in decision making. This software can imitate like human diagnostic expertise for treatment of cancer ailment.

4. PREDICTIVE DATA MINING REVIEW

4.1 Classification

Is a Supervised Learning Technique .Classification is to build (automatically) a model that can classify a class of objects so as to predict the classification or missing attribute value of future objects (whose class may not be known). It is a two-step process. In the first process, based on the collection of *training data set*, a model is constructed to describe the characteristics of a set of data classes or concepts. Since data classes or concepts are predefined, this step is also known as *supervised learning* (i.e., which class the training sample belongs to is provided). In the second step, the model is used to predict the classes of future objects or data.

4.2 Classification based on Naive Bayes Approach.

Naive Bayesian Classifiers are statistical classifiers which can predict class membership probabilities such as the probability that a given sample will belong to a particular case. Naive Classifiers assumes that the effect of an attribute value on the given class is independent of the values of other attributes. Naive Bayesian Classifiers depends upon the BAYE'S THEOREM. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. It performs better in many complex real world situations like Spam Classification, Medical Diagnosis, and Weather forecasting. It is suited when dimensionality of input is high.

4.3 Naive Bayesian Approach in Breast Cancer Prediction.

The endeavor of the classification is to build classifier model based on bayes theorem .Then ,the classifier is used to predict the group attributes of new cases from domain based on the values of attributes. This prediction technique assigns patients to either a "benign "group that is non-cancerous or a "malignant" group that is cancerous.

4.4 Weighted Naive Bayesian approach in Breast Cancer Prediction

The traditional NBC was designed considering the fact that items have same importance and in the database simply their presence or absence is mentioned. In the medical domain all the symptoms does not equally contribute in predicting a particular disease. For example in medical domain predicting the probability of heart disease, the prior-stroke attribute is having more impact than the BMI attribute. In the proposed framework called *Weighted Naive Bayes Classifier (WNBC)* that assigns different weights to different attributes according to their predicting capabilities by consulting the domain expert doctors. Domain based weights are used to assign weight of each attribute using expert knowledge. Experiments have been performed on benchmark data set to evaluate the performance of WNBC in terms of accuracy. The result reveals that WNBC is a promising alternative in medical prediction and certainly deserves further attention.

5. METHODOLOGY

Traditional Naive Classifier is statistical classifier with mutual independency amongst the attributes which is not suitable for medical data mining. So weighted approach is applied on attributes of cancer datasets. Here domain based weights are assign to the attributes. Weighted Naive Bayes classifier uses Bayes Theorem, Weighted count and weighted probability to build the classifier model.WNBC has been proposed as a new technique to build the classifier. The major steps are shown out in figure 1 and the working technique is described below.

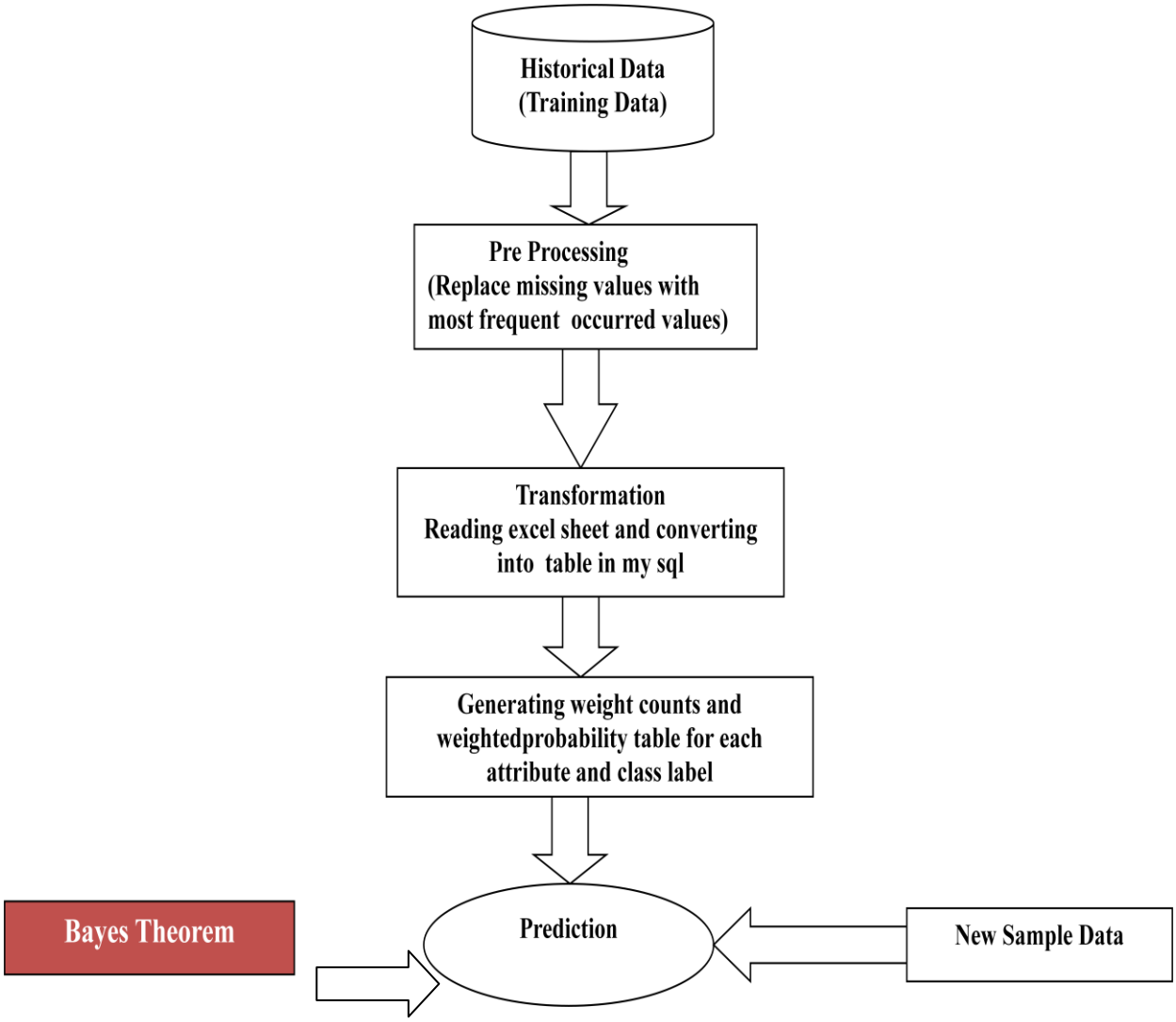


Figure 1.Main steps of Prediction using Weighted Naive Bayes Classifiers

1) Initially, the breast cancer disease data warehouse is pre - processed in order to make it suitable for the mining process.

2) Each attribute is assigned a weight ranging from 1 to 10 to reflect their importance in prediction model. Attributes that have more impact will be assigned a high weight (nearly 9)and attributes having less impact are assigned low weight(nearly 2) as shown in Table 1.

3) Once the preprocessing gets over, bayes theorem is applied to generate interesting pattern. This algorithm uses the concept of Weighted Count and Weighted Probability framework instead of tradition count and Probability as shown in Table 2.

4) Using this approach, Classifier model is built with training dataset .

5) And then this model is applied on test data set .

6) Accuracy is calculated.

7) Whenever a new patient's record is provided, GUI based Prediction System used to predict the class label.

6. BREAST CANCER PREDICTION SYSTEM USING WEIGHTED NAIVE BAYES CLASSIFIER.

A. Weighted Naive bayes classifier

A weighted naive bayes classifiers consists of training dataset $T=\{r_1, r_2, r_3 \ldots r_i \ldots\}$ with set of weight associated with each {attribute, attribute value} pair. Each i^{th} record r_i is a set of attribute value and a weight w_i attached to each attribute of r_i tuple / record. In a weighted framework each record is set of triple $\{a_i, v_i, w_i\}$ where attribute a_i is having value v_i and weight w_i, $1<w_j<=10$. Weight is used to show the importance of the item.

B. Domain Based Weights

In which weights are assigned on the basis of domain knowledge i.e experience of expert doctor can be utilized to assign weights to different symptoms in Medical domain. In Breast cancer datasets, all the nine attributes with 18 different values are assigned weight according to Domain expert knowledge. As every attributes values are having different predicting capability, weights are assigned in range 1 to 10.Attributes values those who are more responsible are given more value or weights as compared to less responsible attributes for the occurrence of breast cancer.

Definition 1.Bayes Theorem

Let D be the training set of tuple & their associated class labels. Each tuple is represented by N attributes such that a tuple will contain N values. Suppose there are m class labels from $C1, C2,\ldots Cm$ for any new tuple X, then the classifier will predict that $X\in$ the class having highest probability condition on X. It shows that X belongs to the i^{th} class then i is having highest probability i.e

$$\text{If } P\left(\frac{ci}{x}\right) > P\left(\frac{ci}{x}\right) where\ 1 \le j \le m$$

The class Ci for which (Ci/X) is maximized is called maximum posterior hypothesis.

As (X) is constant for all the classes it is not considered & the formulas becomes

$$P\left(\frac{ci}{x}\right) = P\left(\frac{x}{ci}\right) * P(ci)$$

In order to predict the class label of X, calculate $P\left(\frac{x}{ci}\right) * P(ci)$ is evaluated for each class Ci and the predictor class label is class Ci for which $P\left(\frac{x}{ci}\right) * P(ci)$ is maximum.

Definition 2. Attribute Weight

Attribute weight is assigned depending upon the domain. For example item in supermarket can be assigned weight based on the profit on per unit sale of an item. In web mining visitor page dwelling time can be used to assign weight .In medical domain symptoms can be assigned weight by expert doctor. Weight of different attribute in predicting the probability of breast cancer is given in Table 1.

Definition 3. Record weight/Tuple Weight

Consider the data in relational table, the tuple weight or record weight can be defined as type of attribute weight. It is average weight of attributes in the tuple. If the relational table is having n number of attribute then Record weight is denoted by $W(r_k)$ and given by

$$W(r_k) = \frac{\sum_{i=1}^{|rk|} Weight(ai)}{\text{Number of attributes in a record}}$$

Definition 4. Weighted Count:

Sum of Records weight having condition attribute column = "Class Value1(S)" given Class Label=S
Total Record Weights

Definition 5 .Weighted Probability:

Weight Count of particular attribute="Yes"/"No"
Weight Count of total "Yes"/"No"

B. Data Source

For training the system Wisconsin Datasets consisting of 699 records with 9 medical attributes with 2 class labels have been used i.e. benign and malignant. Dataset is available in [17] .num format. For the experiment the data in excel sheet is used directly. Table 1 shows different attributes with discritized and normalized values.

Table 1 Normalized Breast.D20.N699.C2 dataset with Values and Weights

S.No.	Attribute name	Attribute Values	Weights(1-10)
1	Clump Thickness[1-10]	1	4
		2	9
2	Uniformity of Cell Size[1-10]	3	5
		4	9
3	Uniformity of Cell Shape[1-10]	5	4
		6	9
4	Marginal Adhesion[1-10]	7	5
		8	9
5	Single Epithelial Cell Size[1-10]	9	3
		10	9
6	Bare Nuclei[1-10]	11	4
		12	8
7	Bland Chromatin[1-10]	13	4
		14	9
8	Normal Nucleoli[1-10]	15	5
		16	9
9	Mitoses[1-10]	17	3
		18	9
10	Output(Class label representing 2 type of breast cancer class)	19/20	

Table 2: Weighted Count and Weighted Probability

Attributes		Clump Thickness		Values	Uniformity of cell		values	Uniformity of cell		Values	Marginal Adhesion		Values	Single Epithelial cell	
	Values	Yes	No		Yes	No		Yes	No		Yes	No		Yes	No
weighted Count	a	0.361311	0.03	c	0.49241	0.006	e	0.450792	0.0028	g	0.488404	0.0496	i	0	0
	b	0.247961	0.36	d	0.116863	0.385	f	0.158481	0.3879	h	0.120869	0.3411	j	0.609273	0.391
weighted Probability	a	0.62	0.1	c	0.84	0.02	e	0.77	0.01	g	0.84	0.01	i	0	0
	b	0.43	1.18	d	0.2	1.2	f	0.23	1.26	h	0.21	1.11	j	1.04451	1.272
Attributes		Bland Chromatin		Values	Bare Nuclei		Values	Normal Nucleoli		Values	Mitosis		Values	Class label	
	Values	Yes	No		Yes	No		Yes	No		Yes	No		Total	
weighted Count	k	0	0	m	0.183314	0.003	o	0	0	q	0	0	s	0.583308	
	l	0.608718	0.39	n	0.420575	0.388	p	0.608718	0.3907	r	0.608718	0.3907	t	0.306937	
weighted Probability	k	0	0	m	0.32	0.01	o	0	0	q	0	0	s	0.59	
	l	1.04356	1.27	n	0.72	1.26	p	1.04	1.27	r	1	1	t	0.4	

Based on this weighted count and weighted probability, classifier model is build. Here classifier model is built with all 699 records and testing is also performed on 699 records.

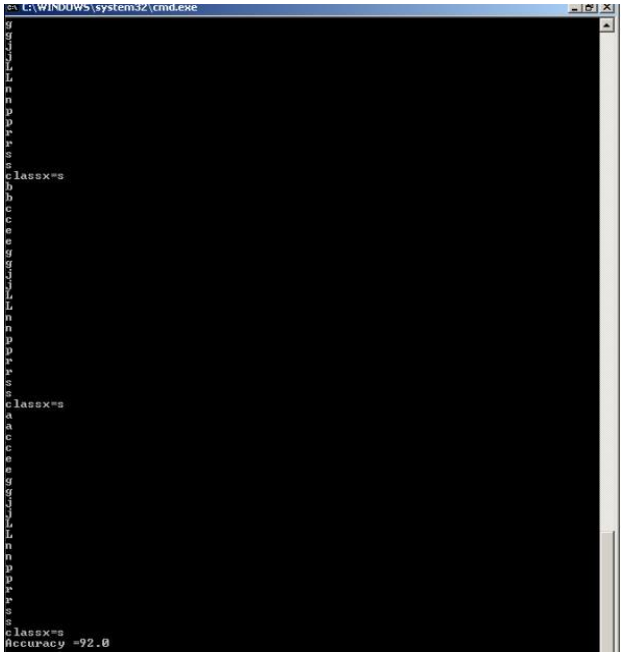
7. EXPERIMENTAL RESULTS

In order to evaluate the effectiveness of WNBC, benchmark Medical data set [18] i.e. breast.D20.N699.C2.num and Java

as front end and MySql as backend tool is used. For the training, all 699 records datasets has been used and testing has been performed on whole dataset.

A. Accuracy

Accuracy is one of the basic performance measures for classification algorithm. The breast.D20.N699.C2.num dataset is having almost 66% of cases with Benign case and remaining 34% is Malignant case. On building the Classifier Model for Breast Cancer Prediction based on Weighted Naive Bayes Classifier, the accuracy is found to be 92% for WNBC.



B. Comparative Analysis

Here is the comparison of different classifiers like WAC, FWAC, CBA, CMAR, CPAR, RBF, and Decision tree on WBC datasets as shown in Table 3. This comparative tabulation states that with weighted approach in NBC gives the very promising result with high accuracy.

Table 3.Accuracy Comparison of different Classifiers on Breast Cancer datasets

S.no	Data Sets	Classifiers	Percentag
1	WBC	Weighted Associative Classifier	90.41%
2	WBC	Fuzzy Weighted Associative Classifier	95.10%
3	WBC	CBA	93.70%
4	WBC	CMAR	88.82%
5	WBC	CPAR	92.84%
4	Large Data Sets	Radial Basis Function	87.42%
5	Large Data Sets	Decision Tree	85.71%
6	Large Data Sets	Nearest Neighbors	84.57%

This comparison shows that introducing the novel concept of Weights in field on Naive Bayes Classifier on Breast cancer detection yields very promising results with strong and efficient predictive results which outperforms others classifiers.

8. CONCLUSION AND FUTURE WORK

Naive Bayes Classifiers is a Predictive Tool. In this work, concentration is on the way to improve the accuracy in terms of efficiency and accurate prediction in the field of Medical Data mining especially Breast Cancer. Further summarization of results is presented in the paper and also description of future work in this area. Here Weighted concept is applied on NBC for Breast Cancer disease and proposed a new predictive model i.e. WNBC. The Framework implementation and experiments have been performed using benchmark dataset to compare its performance with the other non-weighted NBC and recently available model like WAC, FWAC, and RBF etc. To incorporate the weighted concept in NBC, weighted count and weighted probability framework is build.

As compare to recently available advanced model ,the proposed model is found to be equipped with number of advantages such as –an easy to implement the model, readable, modifiable ,efficient training mechanism regardless of the size of the training set, training sets with high dimensionality can be handled easily, interpretability as against to the black box method and finally it can be used as an alternative ,computerized decision tool which can be used to assist physicians in the diagnosis of various disease.

A. Scope of Future Work

- Further, Fuzzy concepts can be introduced in WNBC.As here attributes values are discretized, which causes the sharp boundary problem. A fuzzy logic based preprocessing can be performed to deal with this problem.
- Automated weight assignment can be done using MLE, MCA, PCA, or Information Gain.
- And also in future Bayesian Belief Network can be incorporated as this classifier is best suited for medical prediction.

9. ACKNOWLEDGEMENT

Author acknowledge the guidance and support received from Mrs.Sunita Soni, Sr.Associate Professor, Head of Department(Computer Applications) BIT , Durg for motivating me for my research work.I acknowledge my thanks to Dr. Rajeev Chandrakar (M.D) for his expert knowledge in assigning weights to breast cancer datasets. At last but not the least I am thankful to Bhilai Institute of Technology management and CSIT management for timely support and encouragement in the field of research and development.

10. REFERENCES

[1] Abdelghani,Bellaachia.,Erhan,Guven.2006. Predicting Breast Cancer Survivability Using Data Mining Techniques . Scientific data mining workshop in conjuction with SIAM conference on Data Mining.

[2] Chen,M., Han,J., and Yu,P. 1997. IEEE Trans. Knowledge and Data Eng.8(866) .

[3] Diana, D. 2009. Prediction of recurrent events in breast cancer using the Naive Bayesian Classification. Annals of University of Craiova, Math. Comp. Sci. 36(2):92-96 ISSN: 1223-6934.

[4] Harry, Z.,Shengli,S. 2004.Learning weighted Naive Bayes with accurate Ranking. 4th IEEE International Conference on Data Mining.567-570,ISBN-0-7695-2142-8.

[5] Item Intensities. Knowledge and Information Systems, 6(2):203–229.

[6] Kharya ,S.2012. Using data mining techniques for diagnosis and prognosis of cancer disease. International Journal of Computer Science, Engineering and Information Technology 2(2):55-66.

[7] Kharya, S., Agrawal, S., and Soni,S.2014. Naive Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer. International Journal of Computer Applications (0975 – 8887) Volume 92 (10):26-31.

[8] Mannila, H.1996.Methods and problems in data mining..Proc. of Int. Conf. on Database Theory.

[9] Mohd,F.,Thomas,M.,2007.Comparison of different classification techniques using WEKA for Breast cancer.IFMBE proceedings 15:520-523.

[10] Perichinsky,G., and R, Garc´ıa-Mart´ınez.2000 .Proc. Workshop Comput. Sc. Researchers (La Plata University Press, Buenos Aires. 107

[11] Perichinsky,G., R, Garc´ıa-Mart´ınez., and A. Proto.2000 .Knowledge Discovery Based on Computational Taxonomy And Intelligent Data Mining, CD of the VI Comput. Sc. Argentinean Congr.

[12] Perichinsky,G., R,Garc´ıa-Mart´ınez., A, Proto., A,Sevetto., and D, Grossi.2001. Data Mining: Supervised and Non-Supervised Intelligent Knowledge Discovery, Proc. II Workshop Computes Sc. Researchers

[13] S, Aruna., Dr S.P. Rajagopalan .,and L.V. Nandakishore.2011. Knowledge based analysis of various Statistical tools in detecting breast Cancer.CCSEA. 02:37-45.

[14] Soni,S.,Vyas,O.P 2013.Building Weighted Associative Classifiers using Maximum Likelihood Estimation to Improve Prediction Accuracy in Health Care Data Mining..Journal of Information & Knowledge Management. 12(1) 1350008 (14 pages)

[15] Soni. J. Ansari. Uzma., Sharma, D., and Soni ,S.2011.Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers.IJCSE, 3(6),ISSN:0975-3397

[16] Wang W., Yang, J., and Yu, P.S. 2004. WAR: Weighted Association Rules for item intensities.

[17] Link1-Retreived from http://csc.liv.ac.uk/~frans/KDD/software/LUCS-KDD-DN/datasets/dataSet.html.

[18] Link-2 Retrieved from UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml/]. Irvine, CA: University of California, Center for Machine Learning and Intelligent Systems.