RESEARCH ARTICLE OPEN ACCESS

Classification of Cancer Dataset in Data Mining Algorithms Using R Tool

P.Dhivyapriya [1], Dr.S.Sivakumar [2]
Research Scholar [1], Assistant professor [2]
Department of Computer Science [1]
Department of Computer Applications [2]
Thanthai Hans Roever College, Perambalur
Tamil Nadu - India

ABSTRACT

Cancer is a big issue all approximately the world. It is a disease, which is mortal in many cases and has affected the lives of many and will continue to affect the lives of many more. The most effective way to reduce cancer deaths is to detect it earlier. Early diagnosis needs an accurate and reliable diagnosis procedure that can be used by physicians to distinguish benign cancer from malignant ones without going for surgical biopsy. The objective of these predictions is to assign patients to one of the two group either a "benign" that is noncancerous or a "malignant" that is cancerous. The calculation problem is the lasting care for the virus for patients whose cancer has been surgically removed. Predicting the outcome of a disease is one of the most attractive and challenging tasks where to enlarge data mining applications. The objective of this paper is to predict the presence of two types of life threatening diseases such as Leukemia and Breast cancers by analyzing the clinical datasets. Naïve Bayes and Support Vector Machine prediction models are built for the prediction classification. The performance of the models is then compared in terms of accuracy, time complexity and iterations.

Keywords:- Cancer, naive Bayes, support vector Machine

I. INTRODUCTION

Medical Data Mining (MDM) deals with the problem of scientific decision-making for the diagnosis and treatment of a disease by extracting useful knowledge from large medical databases. Clinical databases have large quantity of in order about patients and their medical circumstances. Relationships and patterns within this data could provide new medical knowledge. Data mining methods assist physicians in many ways right from the understanding of compound diagnostic tests, merging information from multiple sources and providing support for differential diagnosis and providing patient-specific prognosis.

Classification algorithms of data mining often used in the prediction are medical data analysis. Many researchers have been working on improving the presentation of presented algorithms in terms of minimizing the time taken to build the model and maximizing the prediction accuracy of the proposed model.

II. REVIEW OF LITERATURE

Saleema et al [1] found the effect of sampling techniques in classifying the prognosis variable and proposed an ideal example method based on the result of the testing. They compared three example techniques: random, stratified, and balanced stratified. The model has been tested with the SEER data sets. The SEER public use cancer database provides various prominent class labels for prognosis prediction. The categorization model for experimentation had been built using the breast cancer, respiratory cancer and mixed cancer data sets with three traditional classifiers namely Decision Tree, Naïve Bayes and K-Nearest Neighbour. The three prediction factors survival, period and metastasis had been used as class labels for experimental comparisons. The results showed a steady increase in the prediction accuracy of balanced stratified model as the sample size increases, but the traditional approach fluctuates before the optimum results.

Kaishi Li, et,al.,[2] discussed the feature extraction of microarray genes has a greater impact on its categorization and clustering as it is taken as input to any network. The use of gene appearance data in discriminating two types of very similar cancers acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) presented in Classification results are reported in using methods other than neural networks. This paper explores the role of the feature vector in classification. In order to achieve best results in knowledge algorithm, feature

subset collection method should be applied on to the dataset.

Soltani Sarvestani, A. A. Safavi et al [3] composed datasets for breast cancer information detection and invoked various data mining technique to find out the proportion of disease development. Thus, the result helped in selecting a reasonable treatment of the patient. This work also indicated that statistical neural networks can be effectively used for breast cancer diagnosis to help oncologists.

Senthil et al. [4] analyzed the liver cancer DNA sequence data using the generalization of Kimura Models and Markov Chain. The reasonable results verify the validity of our method. The study focused at the level of biological modules, rather than individual genes, results produced by this approach were biologically interpretable and statistically robust. The study tried to use biological knowledge in developing analytic techniques. From the point of view of long-term averages, over a long time period the random variable should spent about 25.96% of the time in state A, about 28.56% of the time in state G, about 34.69% of the time in state C and 10.79% of the time in state T. Finally the result revealed that the percentage is approximately same for all the states. Hence In future, the following symptoms were observed it may lead to liver cancer.

XiangchunXiong et al [5] discussed on three methods to diagnose breast cancer. Mammography, FNA Fine Needle Aspirate and surgical biopsy.They used FNA with a data mining & Statistics method to get an easy way to achieve a best result. They combined some statistical methods with data mining methods to find the unsuspected relationships. They explored that statistics and data mining techniques can offer great promise in helping us uncover patterns in the data.

Table 2.1. Obtained accuracy of reviewed articles

Citation	Algorithms	Accuracy in % (Reviewed Articles)
[6]	Classification and Regression	83%
[11]	Naïve Bayes	83%
[8]	Naïve Bayes Classifier	84.4%
[4]	Decision Tree	86.7%
[10]	Decision Tree	93%
[3]	Decision Tree	98.40%
[13]	Bayes Net	90.95%
[15]	Multilayer Perceptron	96.5%
[16]	Decision Trees(Average)	96.57%
[17]	Decision Table	80.0%

III. MATERIALS AND METHODS

3.1 Naive Bayes Classifier

A Naive Bayes classifier is a easy probabilistic classifier based on applying Bayes theorem from Bayesian data with strong Naïve independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model In simple terms, a Naive ayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature.

3.2 Support Vector Machine

Support Vector Machine is set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classification. A special property of Support Vector Machine is, Support Vector Machine simultaneously minimize the empirical classification error and maximize the geometric margin. So Support Vector Machine called Maximum Margin Classifiers. Support Vector Machine is based on the Structural risk Minimization (SRM).

Support Vector Machine map input vector to a advanced dimensional space where a maximal unraveling hyper plane is constructed. Two parallel frenzied planes are constructed on each side of the hyper plane that separate the data. The separating hyper plane is the hyper planes that maximize the distance between the two parallel hyper planes. An assumption is made that the larger the margin or distance between these parallel hyper planes the better the generalization error of the classifier will be more

One of the primary goals of this thesis is to make a comparative analysis of two classic classification algorithms such as naïve Bayes and Support Vector Machine in terms of predicting the presence of cancer disease in two specific organs breast and bone marrow (Leukemia). The breast dataset is downloaded from UCI Machine Learning

Repository and the Leukemia dataset is downloaded from Bioinformatics Research group.

3.3 Breast cancer dataset description

Malignancy is determined by taking a sample tissue from the patient's breast and performing a biopsy on it. A benign analysis is confirmed either by biopsy or by episodic examination, depending on the patient's choice. All groups in the file are separated as the groups with a line beginning with the number of points in that group. There are 10 attributes per data point, with one data point per line. Attribute are separated by a commas.

- Number of Instances: 699
- Number of Attributes: 10 plus the class attribute
- Number of Missing attributes: 16

The attribute names and the description of the breast cancer dataset are described in Table 3.1.

TABLE3. 1. Breast Cancer Dataset Attribute Description

S.No	Attribute Name	Domain
1	Sample code number	id number
2	Clump Thickness	1-10
3	Uniformity of Cell Size	1-10
4	Uniformity of Cell Shape	1-10
5	Marginal Adhesion	1-10
6	Single Epithelial Cell Size	1-10
7	Bare Nuclei	1-10
8	Bland Chromatin	1-10
9	Normal Nucleoli	1-10
10	Mitoses	1-10
11	Class:	(2 for benign, 4 for malignant)

3.4 Leukemia Cancer Dataset Description

The Leukemia cancer dataset consists of 38 samples: 27 samples of Acute Myeloid Leukemia (AML) and 11 samples of Acute Lymphoblastic Leukemia (ALL). The source of the gene expression measurement is taken from 22 bone marrow samples and 16 peripheral blood samples. Gene expression levels in these 38 samples are measure using high density oligonucleotide microarrays. Each sample contains 7129 gene expression levels.

- Number of Instances: 38
- Number of Attributes: 7129 plus the class attribute
- Number of Missing attributes: None

The domain of the attributes is set to real as the gene expression may fall under real values.

IV. RESULT AND DISCUSSION

4.1 Breast Cancer Dataset

The quality of the classification is measured using two quality measures such as precision and recall have calculated. The precision for a class is the number of true positives divided by the total number of elements labeled as belonging to the positive class. Recall in this context is defined as the number of true positives divided by the total number of elements that really belong to the positive class.

In information recovery, a perfect precision score of 1.0 means that every result retrieved by a search was relevant whereas a perfect recall score of 1.0 means that all relevant documents were retrieved by the search of quality measures

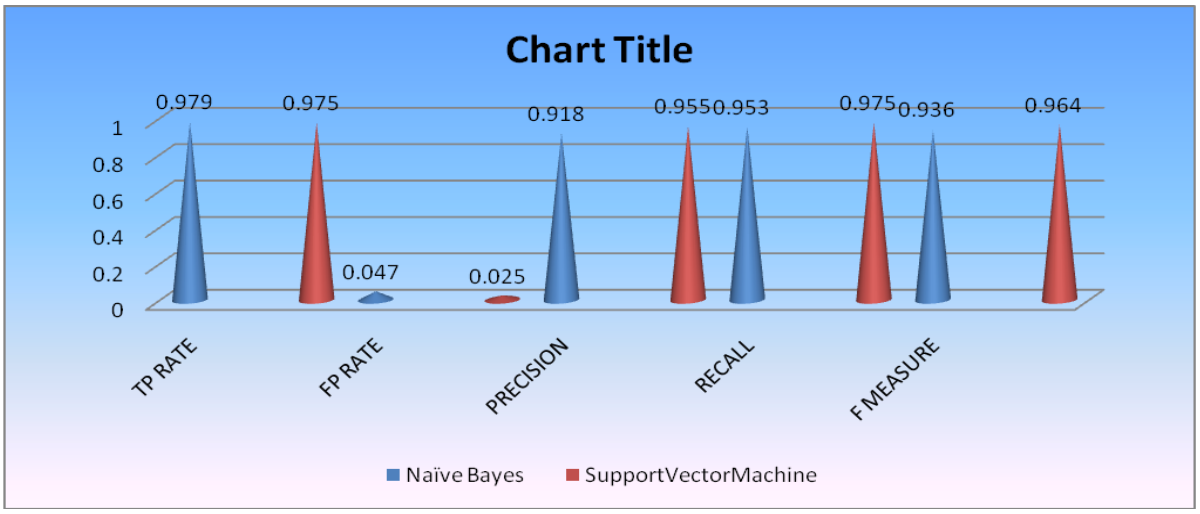

Figure 4.1 Graphical Representation of the quality performance of Naive Bayes and Support Vector Machine (Breast Cancer)

The graphical visualization of the comparative analysis on the performance of Naive Bayes and Support Vector Machine algorithms over breast cancer dataset is shown in Figure 4.1This bar chart displayed in red represents Naive Bayes algorithm and the bar that displayed in aqua represents Support Vector Machine algorithm.

4.2 Leukemia Cancer Dataset

The experiment is extended to compare the performance analysis of naive Bayes and Support Vector Machine classification algorithms with yet another leukemia cancer dataset for verifying the consistency of the qualitative performance of both the algorithms.

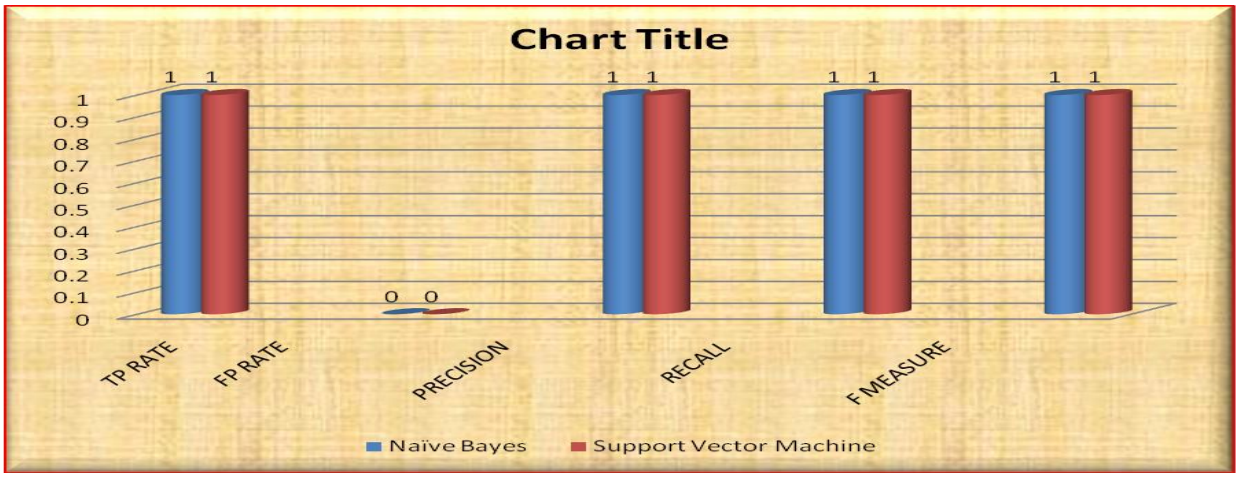

Figure 4.2. Graphical Representation of the quality performance of Naive Bayes and Support Vector Machine (Leukemia Cancer)

The graphical visualization of the comparative analysis on the performance of Naïve Bayes and Support Vector Machine algorithms over Leukemia cancer dataset is shown in Figure. 4.2. The bar that is displayed in red represents Naive Bayes algorithm and the bar that displayed in aqua represents Support Vector Machine algorithm.

The analysis of quality measure, an attempt is directly made to compare the time complexity of a Naive Bayes with the Support Vector Machine algorithm. Five trials of twofold cross-validation of both algorithms were

executed and subjected to exactly the same sequence of splits. This procedure allowed us to run statistical significant tests between the time differences in the Naive Bayes and Support Vector Machine performances. Table 4.1 depicts the number of iterations and time complexity of both algorithms.

TABLE 5.1 Comparative Analysis of Time Complexity and Iterations of naïve Bayes Vs. Support Vector Machine

S.No	Dataset	Iterations		Time taken to build model	
		NB	SVM	NB	SVM
1	Breast Cancer	1253	2781	0.7169	0.8570
2	Leukemia Cancer	3412	4758	0.8375	0.9576

As it is seen not only does the Naïve Bayes run faster than the Support Vector Machine , but for the two datasets the number of iterations also equal or lesser.

V. CONCLUSION

In this research the ability of Naïve Bayes and Support Vector Machine classifiers is compared in terms of accuracy over classifying two different cancer datasets. The best results are achieved using Naive Bayes classifier and Support Vector Machine. Believe that the results are promising, and that with further data preprocessing and adjustment of the classifiers they can be improved. The evaluation of the experiments on Naive Bayes classifier, we found the improvement of 96% for breast cancer and 100% for leukemia cancer .After evaluating same experiments on Support Vector Machine was found the improvement of 97% for breast cancer and 100% for leukemia cancer.

REFERENCES

[1] J.S.Saleema, N.Bhagawathi, S.Monica, P.DeepaShenoy, K.R.Venugopal andL.M.Patnaik," Cancer Prognosis Prediction using Balanced Stratified Sampling" International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.3, No. 1, February 2014..

[2] Kaishi Li, Meixue Yang, Gaurav Sablok, Jianping Fan, Fengfeng Zhou 2013 "screening features to improve the class prediction of acute myeloid leukemia and myelodysplastic syndrome", ELSEVIER, (2013) 348–354

[3] A. Soltani Sarvestani, A. A. Safavi, N.M. Parandeh, M.Salehi 2010 "Predicting Breast Cancer Survivability Using Data Mining Techniques", IEEE (2010) 978-1-4244-8666-3.

[4] K.Senthamaraikannan, N. Senthilvel Murugan, V. Vallinayagam and T. Viveka, "Analysis of Liver Cancer DNA Sequence Data using Data Mining " International Journal of Computer Applications (0975 – 8887) Volume 61– No.3, January 2013.

[5] Xiangchun Xiong, Yangon Kim, Yuncheol Baek, Dae Wong Rhee, Soo-Hong Kim 2005 "Analysis of Breast Cancer Using Data Mining & Statistical Techniques" (2005) IEEE, 0-7695-2294-7/05.