ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Novel Classification based approaches over Cancer Diseases

Mr. P. Thangaraju[1], R. Mehala[2]

Assistant Professor, Department of Computer Application, Bishop Heber College (Autonomous), Tiruchirapalli, India[1]

M. Phil., Research Scholar, Dept., of Computer Science, Bishop Heber College (Autonomous), Tiruchirapalli, India[2]

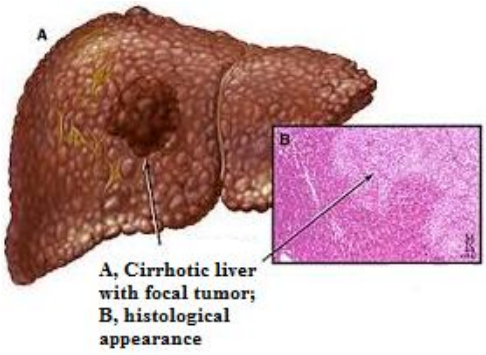
Abstract: One of the crucial applications of data mining is medical data mining (MDM) that diagnoses almost all medical syndromes. MDM incorporates with early predictions, existence and depth of any disease. In fact, MDM supplements the partial assistance of physicians. Though there were many works carried out in MDM, liver cancer is still considered to be a life threatening and has lowest survivability. Predicting the existence of liver cancer at the early stage is highly challengeable for the doctors. The goal of this review is to study the related works on cancer especially liver cancer, an experiment was carried out with the help of dataset to check out the performance of seven classification algorithms with the respect to time and accuracy. This paper also discussed some research key factors that could help the research fraternity to be focused on.

Keywords: Data Mining, Cancer, Liver Cancer, Classification.

I. INTRODUCTION

Cancer disease is the abnormal growth of cells in particular organ that spreads over to invade other tissues of the body. Cancer can be triggered by both internal and external factors. Internal factors could be hormonal issues, mutation inheritance and body resistances. Consumption of tobacco, external radiation and chemicals ingesting are the main external factors that causes cancer. So far, more than 100 typical types of cancers are identified and named as the organ they formed. Few familiar cancer diseases are lung cancer, liver cancer, breast cancer etc. In general, cancers are formed on the functional organ cells of the human body. The life cycle of the cells begins by growing in a controlled way and divide to create more cells to keep the body fit. The cells die and replaced with new ones when they become old or damaged. However, occasionally this cycle fails or gets disturbed due to many reasons. At the point when this happens, cells don't kick the bucket of course and new cells are structured actually when the body does not require them. These additional cells may structure a mass of tissue called a tumor. Tumors can be either benevolent or Harmful. A few diseases don't structure tumors. For instance, leukemia is a tumor of the blood that does not structure tumors. [1]

The medical term for liver cancer is Hepatocellular carcinoma. It is one of the most common malignancies in adults, and is more common for men than women. The existence of hepatocellular carcinoma varies extensively dependent on geographical location. Hepatocellular carcinoma is a malignant tumor composed of cells resembling hepatocytes;however, the similarity varies with the degree of differentiation. This type of liver cancer is possibly treatable by surgical resection. However, only those patients with localized disease are surgical candidates. Liver function impairment and degree of tumor localization determine patient prognosis proliferation. Clinical trials offer alternative treatment options for patients who are not candidates for resection. Figure 1depicts the diagram of the liver cancer



Figue 1. Appearance of liver Cancer

II. LITERATURE SURVEY

Nagaraju and Suresh [1] made a comprehensive comparative analysis of 14 different classification algorithms and their performance has been evaluated by using 3 different cancer data sets. The results indicate that none of the classifiers outperformed all others in terms of the accuracy when applied on all the 3 data sets. Most of the algorithms performed better as the size of the data set is increased. This study focused on finding the right algorithm for classification of data that works better on diverse data sets. However, it is observed that the accuracies of the tools vary depending on the data set used. They recommended the users not to stick to a particular classification method and should evaluate different classification algorithms and select the better algorithm. It should also be noted that classifiers of a particular group also did not perform with similar accuracies. Overall, the results indicated that the performance of a classifier depends on the data set, especially on the number of attributes used in the data set and one should not rely completely on a particular algorithm for their study. So, they recommended that users should try their data set on a set of classifiers and choose the best one.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Priyanga and Prakasam[2] proposed a cancer prediction system based on data mining technology by examining a number of user-provided genetic and non-genetic factors. The system estimated the risk of the breast cancer in the earlier stage. The system was validated by comparing its predicted results with patient's prior medical information and it was analyzed using weka system. The main aim of this model is to provide the early warning to the users, and it is also cost effective to the user. Once the user enters into the cancer prediction system, they need to answer the queries, related to genetic and non-genetic factors. Then the prediction system assigned the risk value to each question based on the user responses.

Once the risk value was predicted, the range of the risk can be determined by the prediction system. We have four levels of risk low level, intermediate, high level and very high level. Based on the predicted risk values, the range of risk will be assigned. The result was shown to the user through data base. The above mentioned technique can be successfully applied to the data sets breast cancer as it was successfully verified on the breast cancer. Finally this prediction system was validated is through a weka tool, it provided the better accuracy compare to the existing system. The main aim of this model was to provide the earlier warning to the users, and it was also cost and time benefit to the user.

Saleema et al [3] found the effect of sampling techniques in classifying the prognosis variable and proposed an ideal sampling method based on the outcome of the experimentation. They compared three sampling techniques: random, stratified, and balanced stratified. The model has been tested with the SEER data sets [21]. The SEER public use cancer database provides various prominent class labels for prognosis prediction.

The classification 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 prognosis factors survival, stage 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.

ShwetaKharya[4] discussed some effective techniques that can be used for breast cancer classification. Among the various data mining classifiers and soft computing approaches, Decision tree was found to be best predictor with 93.62% Accuracy on benchmark dataset [22] and also on SEER dataset.

She suggested the predictor can be used to design a web based application to accept the predictor variables and automated system Decision Tree based prediction can be implemented in remote areas like rural regions or country sides, to imitate like human diagnostic expertise for prediction of ailment. The Bayesian network is also found to be a popular technique in medical prediction Particular it has been successfully utilized for Breast cancer prognosis and diagnosis. In future they intended to design and implement such system for web based applications.

Cheng-Mei Chen andChien-Yeh Hsu [5] established a survival prediction model for liver cancer using data mining technology. They collected the data from the cancer patient's registration database of a medical center in Northern Taiwan between the years 2004 and 2008. A total of 227 patients were newly diagnosed with liver cancer during this time. They extracted nine variables pertaining to liver cancer survival were analyzed using t-test and chi-square test through literature survey and expert consultation. Six variables showed significant. Artificial neural network (ANN) and classification and regression tree (CART) were adopted as prediction models.

The models were tested in three conditions; one variable (clinical stage alone), six significant variables, and all nine variables (significant and non-significant). 5-year survival was the output prediction. The results showed that the ANN model with nine input variables was superior predictor of survival ($p<0.001$). The area under receiver operating characteristic curve (AUC) was 0.915, 0.87, 0.88, and 0.87 for accuracy, sensitivity, and specificity respectively. The ANN model is significant more accurate than CART model when predict survival for liver cancer and provide patients information for understanding the treatment outcomes.

Senthil et al[6] 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.

V.Krishnaiah [7] examined the potential use of classification based data mining techniques to massive volume of healthcare data. The data collected from the health care industry were not "mined" to discover hidden information.

They used One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2) for data preprocessing and effective decision making.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Diagnosis of Lung Cancer Disease can answer complex "what if" queries which traditional decision support systems cannot. Using generic lung cancer symptoms such as age, sex, Wheezing, Shortness of breath, Pain in shoulder, chest, arm, it can predict the likelihood of patients getting a lung cancer disease. They proposed a model for early detection and correct diagnosis of the disease.

Ada and Rajneet [8] used some techniques that were essential to perform medical image mining, Lung Field Segmentation, Data Processing, Feature Extraction, Classification using neural network and SVMs.

The methods used in this paper work states to classify digital X-ray chest films into two categories: normal and abnormal. The normal or negative ones are those characterizing a healthy patient. Abnormal or positive ones include types of lung cancer.

Different learning experiments were performed on two different data sets, created by means of feature selection and SVMs trained with different parameters; the results are compared and reported. In this paper they used classification methods in order to classify problems aim to identify the characteristics that indicate the group to which each case belongs.

Dursun et al [9] used two popular data mining algorithms (artificial neural networks and decision trees) along with a most commonly used statistical method (logistic regression) to develop the prediction models for breast cancer survivability using a large dataset (more than 200,000 cases).

They used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes.

The results indicated that the decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample, artificial neural networks came out to be the second with 91.2% accuracy and the logistic regression models came out to be the worst of the three with 89.2% accuracy.

Thangaraju et al [10] built a model based as a test case on the UCI repository dataset. The experiment had been performed with several data mining classification techniques and it was found that the Naive Bayes algorithm gives a better performance over the supplied data set with the accuracy of 83.4%[19].

They assured that the data mining can significantly help in the Lung Cancer research and ultimately improve the quality of health care of Lung Cancer patients. They also suggested some future to introduce new models can be used in the design of clinical decision support system for mining Lung Cancer.

TABLE I. ATTAINED ACCURACY OF CLASSIFICATION TECHNIQUES ON REVIEWED ARTICLES

Citation No	Disease Selected	Attributes	Instances	algorithm Used	Accuracy
[5]	Liver Cancer	9	277	Classification and Regression	83%
[10]	Lung Cancer	13	6728	Naïve Bayes	83%
[7]	Lung Cancer	57	32	Naïve Bayes Classifier	84.4%
[4]	Breast Cancer	11	699	C4.5	86.7%
[9]	Breast Cancer	72	433272	Decision Tree C5	93%
[3]	Breast Cancer	64	30000	Decision Tree	98.40 %

III. EXPERIMENTATION AND RESULT ANALYSIS

In this study, a dataset for breast cancer was derived and used for the experimentation from the UCI machine learning repository database. The dataset comprised of ten attributes with 287 instances. This dataset was given as an input to the most popular data mining tool WEKA 3.6[23] for analyzing the correct accuracy prediction of various classification algorithms[16]. Table 2. describes the attributes with its possible range[16.

S. No	Attribute Name	Description
1	Class	no-recurrence-events, recurrence-events
2	age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
3	Menopause	lt40, ge40, premeno
4	tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
5	inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32, 33-35, 36-39
6	node-caps	yes, no
7	deg-malig	1, 2, 3
8	Breast	left, right
9	breast-quad	left-up, left-low, right-up, right-low, central
10	irradiat	yes, no

TABLE 2. Breast Cancer Data Description

In this work, 7 different types of classification algorithms were chosen such as Bayes Net, Naïve Bayes, Naïve Bayes Simple, Naïve Bayes Updatable, Multilayer Perceptron, IBk, KStar, Decision Table, JRip, ZeroR, J48, REPTree. From the results, it was fount, among all classification algorithms multilayer perceptron had taken more time for building the prediction model with 10.69 seconds with the accuracy 73%. Here, the Accuracy was

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

measured in terms of correctly classified instances. Decision Table algorithm had the highest prediction accuracy as 79%. Time taken to build the model and correctly classified instances were depicted in TABLE 3.

TABLE 3. Accuracy Prediction over Classification Algorithms

	Algorithm	Accuracy
1	Bayes Net	74.48%
2	Naïve Bayes	74.82%
3	Multilayer Perceptron	73.07%
4	Decision Table	79.02%
5	S.No	72.38%
6	J48	75.87%
7	REPTree	77.97%

IV. CONCLUSION

This paper provides a study of various technical and review papers on lung, liver and Breast cancer data sets and explores that data mining techniques offer great promise to uncover patterns hidden in the data that can help the clinicians in decision making. From the above study it is observed that the accuracy for the diagnosis analysis of various applied Data mining Classification techniques, Implementation of the techniques are highly acceptable and can help the medical professionals in decision making for early diagnosis and to avoid biopsy. In the case of the above said data sets it is observed that the accuracy reached up to 100% when there is more number of attributes and the accuracy is decreasing as the number of attributes decreased.

REFERENCES

[1] Gopala Krishna Murthy Nookala, Bharath Kumar Pottumuthu, NagarajuOrsu, Suresh B. Mudunuri, " Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification" , (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5, 2013.

[2] A.Priyanga andDr,S.Prakasam " The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness" International Journal of Computer Science and Engineering Communications- IJCSEC. Vol.1 Issue.1, December 2013.

[3] 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.

[4] ShwetaKharya, "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[5] Cheng-Mei Chen, Chien-Yeh Hsu, Cheng-Mei Chen andChien-Yeh Hsu," Prediction of Survival in Patients with Liver Cancer using Artificial Neural Networks and Classification and Regression Trees" Seventh International Conference on Natural Computation 2011.

[6] K.Senthamaraikannan,N.SenthilvelMurugan, 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.

[7] V.Krishnaiah, Dr.G.Narsimha and N.Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 – 45.

[8] Ada andRajneetKaur, "A Study of Detection of Lung Cancer Using Data Mining Classification Techniques",International Journal of Advanced Research in Computer Science and Software Engineering 3(3), March - 2013, pp. 131-134.

[9] DursunDelen, Glenn Walker, Amit and Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods" Department of Management Science and Information Systems, Oklahoma State University, 700 North Greenwood Venue, Tulsa, OK 74106, USAReceived 13 January 2004.

[10] Thangaraju P1, Barkavi G2 and Karthikeyan T, "Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques",International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014.

[11] R. D´ıaz-Uriarte and A. de André's.Gene selection andclassification of microarray data using random forest. BMCbioinformatics, 7(1):3, 2006.

[12] R.S. Michal ski and K. Kaufman. Learning patterns in noisy data:TheAQ approach. Machine Learning and its Applications, Springer-Verlag, pages 22–38, 2001

[13] R. Linder, T. Richards, and M. Wagner.Microarray data classifiedby artificial neural networks.METHODS IN MOLECULARBIOLOGYCLIFTON THEN TOTOWA-, 382:345, 2007.

[14] Murat Karabhatak, M.CevdetInce 2008.Expert system fordetection of breast cancer based on association rules and neuralnetwork. Journal: Expert systems with Applications.

[15] Han, J. and M. Kamber, 2001. Data Mining: Concepts andTechniques. San Francisco, Morgan Kauffmann Publishers.

[16] T. Karthikeyan and P.Thangaraju,"Analysis of Classification Algorithms Applied to Hepatitis Patients", International Journal of Computer Applications, vol. 62, no. 5, Jan. 2013.

[17] Maria-LuizaAntonie, Osmar R. Za¨ıane, AlexandruComanApplication of Data Mining Techniques for Medical ImageClassification.Page 97

[18] Heckerman, D., A Tutorial on Learning with BayesianNetworks.1995, Microsoft Research.

[19] T. Karthikeyan and P.Thangaraju, "PCA-NB Algorithm to Enhance the predictive Accuracy", International Journal of Engineering and Technology, vol. 6, no. 1, Feb-Mar 2014.

[20] Neapolitan, R., Learning Bayesian Networks. 2004, London:Pearson Prentice Hall.

[21]. http:// seer.cancer.gov/data/

[22]. https://archive.ics.uci.edu/ml/datasets.html

[23] www.cs.waikato.ac.nz/ml/weka/