International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

An Integrated Cancer Prediction System Using Data Mining Techniques

Pallavi Mirajkar[1], Dr. G. Prasanna Lakshmi[2]
[1]Research Scholar Faculty of Computer Science Pacific Academy of Higher Education and Research University Udaipur, Rajasthan, India
[2]Guide, (WOS-A) Andhra University, Andhra Pradesh, India

ABSTRACT

Cancer identification and prediction are huge challenge to the researchers. The use of various techniques of data mining techniques has revolutionized the whole process of cancer Diagnosis and Prognosis. We are proposing integrated system which is based on combination of various data mining techniques such as analytical hierarchy process, rule based association, classification etc. that is helpful to predict the patient's disease status. Cancer disease risk can be discovered by analyzing and identifying various factors and symptoms of the patient before recommending treatments. The vital aim of our system is to help oncologist and medical practitioners in diagnosing the patient by analyzing available data and relevant information.

Keywords: Ensemble Learning, Prediction, Cancer, Data Mining, Rule-based algorithm, Integration.

I. INTRODUCTION

Cancer study is generally medical and/or genetic in nature, data driven statistical research has become a common complement. Predicting the disease is one of the most important and tricky tasks where to implement data mining applications. As the utilization of computers controlled with automated tools, expansive volumes of medical data are being gathered and made accessible to the medicinal research groups.

Data mining plays a significant role in the medical field by predicting various diseases. It is the process of selecting, exploring large amounts of data to find out previously unknown patterns [1]. Data mining is used to analyze different type of data by using available data mining tool. There are different data mining techniques, such as classification, regression, clustering and association rules that are applied on data sets for prediction results [2].

II. LITERATURE REVIEW

Subrata Kumar Mandal (2017) have applied different techniques such as data cleaning, feature selection, feature extraction, data discretization and classification for predicting breast cancer as perfectly. They stated that Logistic Regression Classifier gives the maximum accuracy with reduced subset of features and time complexity of this algorithm is least compared to other two classifiers.

R.Kaviarasi, Dr.A.Valarmathi (2017) has applied two types of clustering. The hierarchical clustering is produced dendrogram results are produced using Euclidean distance and Ward.D linkage. The K-Means clustering are produced WSS values against number of cluster K values. The proposed method finally validate to the two type of clustering fit values. The validation measurement result is helped to the distance are measured in two clustering values. The results are helped to at the beginning of Non Small

Cell Lung Cancer prevention through human way of life handled risk factors characteristics. This paper work is very helped to the cancer research center and hospitals to prevent the Non Small Cell Lung Cancer.

Arpit Bansal et, al.(2017) have proposed technique for a modification in K-Means Clustering Algorithm. The proposed modification in the K-Means clustering will vanish off the two vital drawbacks of K-Means clustering that are accuracy level and calculation time consumed in clustering the dataset. Although when they have also used small datasets these two factors accuracy level and calculation time may not matter much but when they used large datasets that have trillions of records, then little dispersion in accuracy level will matter a lot and can lead to a disastrous situation, if not handled properly, it can be stated that proposed modification can be more extended to achieve the full accuracy level up to 100%, with very little time and with more quality clusters.

R.Senkamalavalli and Dr.T.Bhuvaneswari (2017) have proposed novel algorithm was experimented on the Breast cancer database. It has been proved that this approach achieved a very high accuracy rate than the existing methods. They also demonstrated a certain level of accuracy in the classifier, and for finding accurate results there must be sufficient preprocessing of data done. They have also demonstrated accuracy in diagnosing breast cancer disease using the K-means classifier, adaboost and Support Vector Machines.

Sumalatha.G, Archana.S (2017) has analyzed cancer patient data using ZeroR method and J48 algorithm of data mining techniques. They stated that this prediction system may give easy and a cost effective approach for screening cancer and may play a significant role in earlier diagnosis process for different types of cancer. This system can also be used as a source of record with detailed patient history in hospitals as well as help doctors to premeditate on particular therapy for any patient.

Megha Rathi et.al. (2016) have developed a software tool for the prediction of disease which helps in decision making for the treatment method. This tool will be helpful in diagnosing the disease type and to help out for decision support in medical system. They studied hybrid classification technique is used for classification of medical data sets and is applicable in healthcare domain. SVM and bootstrap is integrated to improve classification accuracy. This tool helps doctors or patient to decide in a short time whether the person is suffering from disease and is generic to all types of disease.

Prabhakar Chalise et,al.,(2014) have proposed Cluster analysis aims to highlight meaningful patterns or groups inherent in the data that will be useful in identifying the subtypes of the diseases. Different types of clustering algorithms have been proposed that use various assays of molecular variation of cells most of which are designed for one type of data at a time. These types of methods have been successfully implemented in several disease classification studies.
Rahul Patil et.al (2016) has used various clustering techniques of data mining such as partition based method, hierarchical based clustering. They stated that this hybrid method will help society to analyze and understand disease and their health status.

Tanu Minhas and Nancy Sehgal (2017) concluded that prediction analysis is an efficient technique for the complex data analysis. They have applied k-means clustering algorithm with Boltzmann algorithm to increase accuracy of data clustering. They stated that the performance of proposed algorithm is tested in MATLAB and it has been analyzed that accuracy is increased up to 20 percent.

III. PROPOSED WORK

Cancer prediction is one of the major challenges in the health care industry. Motivated by the world-

wide increasing mortality of cancer patients, researchers are using different data mining techniques in the diagnosis of cancer disease. In thesis study, three data mining techniques were used to propose a medical diagnostic scheme for prediction of cancer. Figure 1 show a graphic representation of our overall approach, the framework of the integrated multi mapping method, which consisted of three steps.

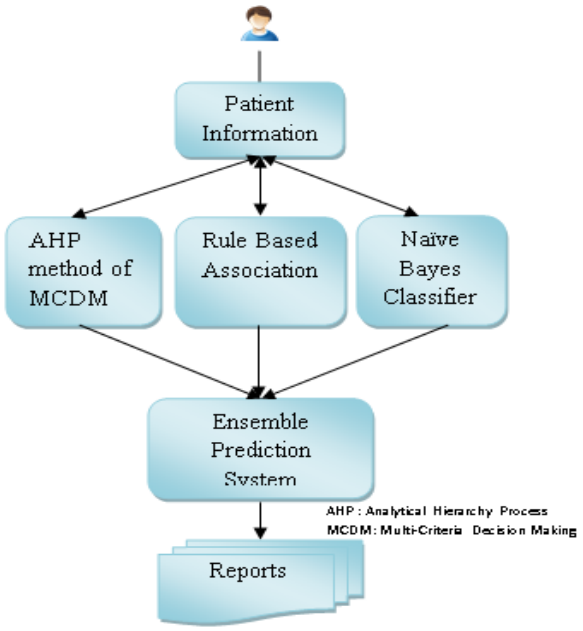

Figure 1. Proposed Integrated Model of

Each technique used in an algorithm contains certain functions which are helpful to predict the cancer disease at an early stage. For perfect prediction of cancer disease, the outputs of each algorithm is integrated and compared. Here combination of output is considered "Integration". By applying integrated data mining techniques can confirm promising results in the diagnosis of cancer disease.

The first stage of the proposed prediction system consists of Analytical hierarchy process method of multi-criteria decision making which is used to develop rational and objective findings. Multi-criteria decision making is used to develop a rational and objective finding. The advantage of the analytical hierarchy process is simple and maintains consistency as compared to other decision support system. In the second stage, we have used rule based classifier of data mining techniques. Rule based analysis is a technique to uncover how items are associated to each other. Association rule are developed by identifying data for frequent if/then patterns and identification of the most important relationship by using the criteria support and confidence. The proposed algorithm produces ranking result for all frequent patterns. After studying rule based association, in the third stage we have proposed prediction of cancer risk using Naïve Bayes classifier. Naïve Bayes classifier is most effective statistical and probabilistic classification algorithm. The proposed algorithm learns the probability of an object with certain features belonging to a particular group/class.

As the user enters into the cancer prediction system, they have to answer the questions, Then the prediction system calculate the risk score. Based on the predicted risk values the range of risk will be assigned. The result can be shown to the user through data base.

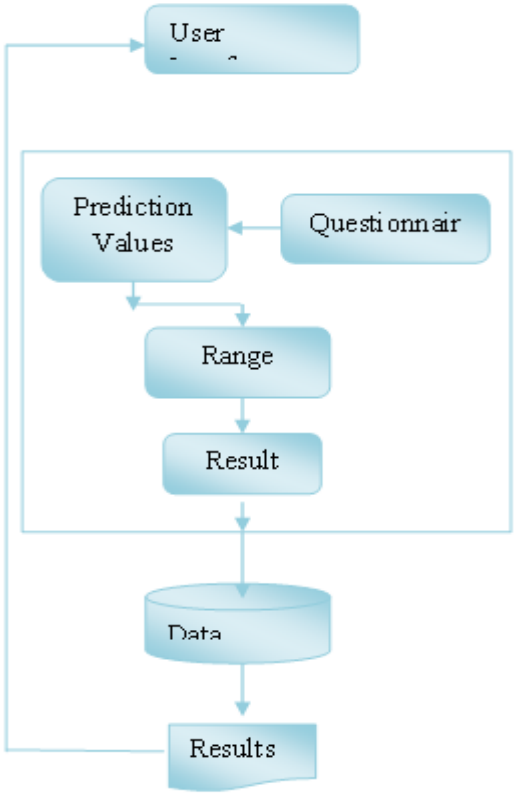

Figure 2. Architecture of Cancer prediction

Working of the Integrated Model:

Step 1: Enter the personal information of the patient.

Step 2: Prediction system will checks for the various conditions.

Step 3: Select the various risk factors Ri

Where R= Risk Factors, i= (1, 2, 3… n) and also select its severity as given in table no. 1

Step 4: Select the different cancerous food habits Fi Where i= (1, 2, 3… n) as given in table no. 2

Step 5: Choose the organ of the symptoms.

Step 6: Display Symptoms related to the organ

Step 7: Design a matrix for selected symptoms by the patients

	Weightage W			
S1				
S2				
.				
.				
.				
Sn				

Step 8: Design a matrix for selected risk factors by the patients

	Weightage W			
	Moderate	Strongly	Very Strongly	Extreme
RF1				
RF2				
.				
.				
.				
RFn				

Step 8: In favor of, all selected symptoms, select the weight of the symptoms and determined the risk value.

Step 9: Calculate the Risk Score

Risk Score= $\sum_{j=1}^{n}(Sij) + \sum_{j=1}^{n}(RFij) * Wsv$

Where W_{sv} = Weightage of Scaling Variable

Step 10: Display the Risk Score.

IV. CONCLUSION

Cancer as an unceasing disease is expensive for medical systems in many countries. Appropriate cancer management and suitable control programs are very important especially in developing countries with limit resources and require correct information. Data analysis techniques are essential for effective usage of stored data and provide useful knowledge for health care personnel to improve decision support, increase quality of preventive; treatment and diagnosis planning.

Prediction system will help to oncologist to predict cancer disease at an early stage. Report and feedbacks generated can decrease physician resistance and improve their attitudes and trust.

Automation tools are required to improve diagnosis and prognosis of oncologist or medical practitioners to save affected cancer patients.

V. REFERENCES

[1]. Dave Smith, "Data Mining in the Clinical Research Environment". Available at. http://www.sas.com/

[2]. Damtew A., "Designing a predictive model for heart disease detection using data mining Techniques" A Thesis Submitted to the School of Graduate Studies of Addis Ababa University, 2011.

[3]. https://www.cancer.gov/

[4]. Subrata Kumar Mandal "Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naive Bayes, Logistic Regression and Decision Tree" International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 6 Issue 2 Feb. 2017, Page No. 20388-20391.

[5]. R. Kaviarasi, Dr. A.Valarmathi "Near The Beginning of Non Small Cell Lung Cancer Avoidance in Human Way of Life Risk Factors

Classification Using Clustering Algorithm in the R Environment" International Journal of Advanced Research in Computer Science, 8 (5), May-June 2017,1023-1026.

[6]. Arpit Bansal, Mayur Sharma "Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining", IJCA (0975-8887) Vol. 157-No 6, January 2017.

[7]. Prabhakar Chalise, Devin C.Koestler, Milan Bimali, Qing Yu, Brooke L.Fridley to "Integrative clustering methods for high – dimensional molecular data" , Transl Cancer Res 2014.

[8]. Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR (2013) "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence". J Health Med Inform 4: 124. doi:10.4172/2157-7420.1000124

[9]. V.Kirubha , S.Manju Priya "Survey on Data Mining Algorithms in Disease Prediction" International Journal of Computer Trends and Technology (IJCTT)-Volume 38 Number 3 - August 2016.

[10]. P. Saranya , B. Satheeskumar "A Survey on Feature Selection of Cancer Disease Using Data Mining Techniques" International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016, pg. 713-719.

[11]. R.Senkamalavalli and Dr.T.Bhuvaneswari "Improved Classification Of Breast Cancer Data Using Hybrid Techniques" International Journal of Advanced Research in Computer Science Volume 8, No. 8, September-October 2017 ISSN No. 0976-5697.

[12]. Sumalatha.G, Archana.S "A Study on Early Prevention and Detection of Breast Cancer using Data Mining Techniques" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Website: www.ijircce.com Vol. 5, Issue 6, June 2017.

[13]. Megha Rathi , Vikas Pareek "Disease prediction tool: an integrated hybrid data mining approach for healthcare" IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol.6, No.6, Nov-Dec 2016.

[14]. Rahul Patil, Pavan Chopade, Abhishek Mishra, Bhushan Sane, Yuvraj Sargar "Disease Prediction System using Data Mining Hybrid Approach" Communications on Applied Electronics (CAE)-ISSN : 2394-4714 Foundation of Computer Science FCS, New York, USA Volume 4-No.9, April 2016.

[15]. Tanu Minhas , Nancy Sehgal "Prediction Analysis Technique using Hybrid Clustering and SVM Classification" International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Website: www.ijirset.com Vol. 6, Issue 7, July 2017.

[16]. K. Suneetha "Early Prediction and Detection of Lung Cancer using Data Mining" International Journal of Advanced in Management, Technology and Engineering Sciences Volume 7, Issue 12, 2017 ISSN NO : 2249-7455.

[17]. Chih-Jen Tseng, Chi-Jie Lu, Chi-Chang Chang, Gin-Den Chen, Chalong Cheewakriangkrai "Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence" Contents lists available at Science Direct Artificial Intelligence in Medicine journal homepage: www.elsevier.com/locate/aiim/ Artificial Intelligence in Medicine 78 (2017) 47–54.

[18]. Huang M-W, Chen C-W, Lin W-C, Ke S-W, Tsai C-F (2017) " SVM and SVM Ensembles in Breast Cancer Prediction" PLoS ONE 12(1):e0161501. doi:10.1371/journal.pone.0161501.

[19]. Florije Ismaili , Luzana Shabani , Bujar Raufi , Jaumin Ajdari , Xhemal Zenuni "Enhancing breast cancer detection using data mining classification techniques" 2nd World Conference on Technology, Innovation and Entrepreneurship May 12- 14, 2017