PROGNOSIS OF BLOOD CARCINOMA USING DATA MINING TECHNIQUES

Durgalakshmi.R,
M.E., Multimedia Technology
Anna University Regional Campus
Coimbatore, India.

Mannar Mannan.J,
Department Of Information Technology,
Anna University Regional Campus
Coimbatore, India.

Abstract: in the present Information age, Data mining plays a vital role in various research areas like marketing, customer behavior analysis and various medical researches such as cancer prediction system which provides effective preventive strategy. Data mining algorithms are used on medical database in order to predict survivability of cancer patients. Blood cancer causes serious issues and hence Data mining techniques are broadly utilized to predict blood cancer from test data collection. The Data mining algorithms along with semantic knowledge using Ontology is used for predicting blood cancer. Earlier system used genetic and environmental factors in order to predict cancer. This paper uses data mining technologies such as classification, clustering along with semantic analysis in order to predict leukemia. The parameters from complete blood count (CBC) and peripheral smear are taken to analyze possibilities of leukemia's presence. Data mining is used for analyzing Semantic relationship among parameters in Datasets. The gathered data is preprocessed fed into the database and classified to obtain significant patterns using Data mining algorithms. Then the data is clustered in order to separate malicious and non-malicious leukemia. Then finally by analyzing Ontology based semantic relationship, prediction of leukemia can be easily done, before it becomes severe. The experimental results show the better performance compared with existing techniques.

Keywords: Data mining, Clustering, Classification, Ontology

I. INTRODUCTION

Computer plays a major role in Medical research area. Information processing, exchanging, interpreting is a key points to discuss. Data mining plays a vital role in Medical research in order to predict disease early. Krishnaiah.V et al., [5] developed a lung cancer prediction system using Data mining techniques. This is most effective model to predict patients with Lung cancer appears to be Naïve Bayes followed by IF-THEN rule,Decision Trees and Neural Network. So people suggests Data mining techniques in order to predict Cancer. Blood cancer causes severe social issues that challenges the government and medical practitioners. Early prediction of cancer is still not yet optimal and there is no appropriate methodology to predict cancer early. Durairaj.M and Deepika.R[6] Reviews several articles related in diagnosing myeloid dysplastic syndrome (MDS) and acute myeloid leukemia (AML) causes. They address the issue of applying medical data mining, an upcoming research trend which helps in finding accurate solutions in many fields. This prediction techniques sometimes fails to predict based on stability and accuracy in classification algorithms. There is a need of modern and adaptive Data mining techniques is required to predict the blood cancer early Data mining is the process of abstracting patterns from data. Data mining is the search for hidden patterns that may exist in large databases. Data mining scans a large volume of data to discover patterns and correlations among patterns. Data mining involves the use of intellectual data analysis tools to discover previously unknown, proper patterns and relationships in large datasets. These tools can include statistical data, mathematical forms and machine learning methods.Data mining can be executed on data represented in quantitative, textual or other multimedia forms. Data mining applications can use a diversity of parameters to examine the data. They include association mining, sequence or path analysis, classification, clustering and forecasting. Data mining tools predict future trends and acts. Data mining tools contain different tasks. Data mining tool not to be inconstant, extensible, able of exactly forecasting responses between actions and results and capable of automatic implementations. Data mining is commonly used in a wide area of profiling practices such as marketing, close observation of an individual or group, fraud detection and health care.

A. Data Mining Process

Data from different source have provided in Data Warehouse in which mining algorithms are incorporated. By using Data mining techniques, Classification, clustering and patterns can be easily identified. Then the result is obtained by analyzing the patterns. The Schematic representation of Data mining process is shown in Figure 1.

B. Data Mining in Healthcare

In healthcare, Data mining is becoming more popular and its applications are greatly benefit all parties involved in the healthcare industry.Data mining can help healthcare insurers detect fraud and abuse. In Healthcare organizations customer relationship management decisions can make easily.Physicians

can identify effective treatments and best practices.Patients receive better and more affordable healthcare services. The large volume of data generated by healthcare transactions are too complex to be processed and analyzed by traditional methods. Data mining techniques provide the methodology and technology to transform these protected data into useful information for decision making.

C. Data Mining in Cancer Prediction

Data mining techniques can be used to build Cancer prediction System. This system provides effective preventive strategy. Data mining algorithms are used on medical Database in order to predict survivability of cancer patients. Some of the Data mining algorithms have been identified as Decision Trees, Support vector machine, artificial neural network, Naïve Bayes, fuzzy rule etc.,Clustering is a process of separating data into subgroups according to their characteristics. A cluster is a collection of data objects that are similar to each another within the same cluster and are dissimilar to the objects in differentclusters. Data Mining algorithms are implemented together to create a novel method to diagnose the presence of cancer for patient. In order to ignore data mining problem, it is initial necessary to bring all the data together into a set of instances. Integrating data from different sources usually presents many challenges. The data must be gathered, integrated, and cleaned up. After this methods, data can be used for processing through machine learning techniques.

D. Data Mining and Ontology in Cancer Prediction

Ontology plays a major role in the format of Semantic Web int egrationwhere in information is given described meaning and t he Search engines deploy ontology to find pages with words that are semantically similar but syntactically different. A data warehouse is the essential point of any decision support system which stores preprocessed and integrated data for knowledge discovery during the data mining systems. To improve the overall data warehouse mining process, one needs to adopt an intelligent data warehouse mining approach incorporated with a user preference ontology.

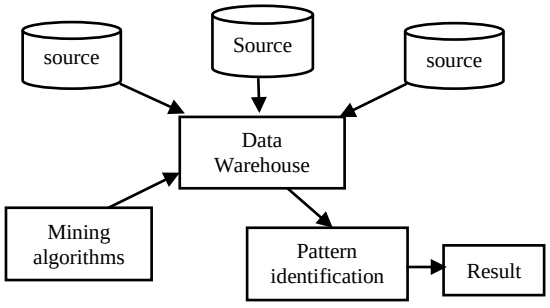

Figure 1: Data mining Process

II.RELATED WORK

Ramachandran.P et al., [1] have proposed Early Detection and prevention of cancer using K-means clustering algorithm to separate cancer and non cancer patient data that plays a important role in predicting cancer.Identification ofgenetic and environmental factors is very important in developing novel methods to detect and prevent cancer, but unfortunately genetic testing is cost and time consuming process.Finally, a prediction system is developed to analyze risk levels which help in prognosis. The most effective way is to detect cancer before it becomes severe. Similarly, Ritu Chauhan et al.,[2] focuses on clustering algorithm such as HAC and K-Means in which, HAC is applied on K-Means to determine the number of clusters. The quality of cluster is improved, if HAC is applied on K-Means. The experiment results showed that there are certain facts that are evolved and cannot be superficially retrieved from raw data. Work related to missing values cannot be Determined. It is an emerging field which is currently used in marketing, Surveillance fraud detection, human factor related issue, medical pattern detection and scientific discovery.

Raymond.T and Jiawei Han [8] explored whether clustering methods have a role to play in spatial data mining. To this end, they had developed a new clustering method called CLARANS which is based on randomized search and also develop two spatial data mining algorithms that use CLARANS. Furthermore, experiments conducted to compare the performance of CLARANS with that of existing clustering methods show that CLARANS is the most efficient. Alternatively, Arutchelvan.K and Periyasamy.R [7] Performs experimental evaluation of the effectiveness and efficiency of DBSCAN using synthetic data. The results of this experiment shows that DBSCAN is significantly more effective in discovering clusters of arbitrary shape than the well-known algorithm CLARANS and that DBSCAN outperforms CLARANS by a factor of more than 100 in terms of efficiency. The applications of DBSCAN to high dimensional features spaces should be investigated. In particular, the shape of the k-dist. graph in such applications has to be explored.

Lookman sithic.H and Umarani.R [9] have explored data mining techniques in healthcare management. Particularly, it talk about data mining and its various application in areas where people are mostly affected rigorously by cancer using tobacco, chemical water. This paper identifies the cancer level using clustering algorithms and finds meaningful hidden patterns which gives meaningful decision making to this socio-economic real world health venture. Similarly, Gopala Krishna Murthy Nookala et al., [13] made a comprehensive comparative analysis of 14 different classification algorithms and their performance has been evaluated by using 3 different cancer data sets and recommend the users not to stick to a particular classification method and should evaluate different classification algorithms and select the better algorithm.

Similarly, Krishnaiah.V et al., [5] have developed a prototype for lung cancer prediction system using Data mining classification technique.The most effective model to predict Lung cancerusing Naïve Bayes followed by set of rules and facts. For Diagnosis of Lung Cancer, Naïve Bayes observes better results and fared better than Decision Trees.Continuous data cannot be used instead of just categorical data. Decision Trees results are easier to read and interpret.Naïve Bayes fared better than Decision Trees as it could identify all the significant medical predictors.The emphasis of this work is to find the target group of people who needs further screening for Lung cancer, so that the prevalence and mortality rate could be

brought down. Dechang Chen et al., [3] haveproposed the algorithm EACCD developed which a two steps clustering method. In the initial step, a dissimilarity measure is learnt by using PAM and in the following step, the learnt dissimilarity is used with a hierarchical clustering algorithm to obtain clusters of patients. These clusters of patients form a basis of a prognostic system and presence of high percentage of censored observations is difficult to determine. A computer-based prognostic system for cancer patients that admit multiple prognostic factors by using this factors survival function provides a probability of survivability for a cancer patient.

In addition to this, Charles Edeki et al., [4]Suggests that none of the Data Mining and statistical learning algorithms applied to breast cancer dataset outperformed and hence it could be declared the optimal algorithm for the prediction of survivability rate of Breast cancer.This analysis does not include records with missing data.The preliminary results are promising for the application of the data mining methods into the survivability prediction problem in medical databases. The C4.5 algorithm has a much better performance than the other Data mining techniques.Similarly, G.Ravikumar et al., [15] aims to establish an accurate classification model for Breast cancer prediction, in order to make full use of the invaluable information in clinical data, especially which is usually ignored by most of the existing methods when they aim for high prediction accuracies and infer that the SVM are more suitable in handling the classification problem of breast cancer prediction and we recommend the use of these approaches in similar classification approaches.

Likewise,Durairaj.M and Deepika.R [6] havefocused the issue of applying medical data mining, an emerging research trend which helps in finding accurate solutions in many fields especially in diagnosing myeloid dysplastic syndrome (MDS) and acute myeloid leukemia (AML) pathogenesis. They have highlighted the various data mining techniques used for the diagnosis of cancer and compares the correct accuracy level of the pathogens. Similarly, Neha Sharma et al., [14] Proposes ED&P framework which is used to develop a data mining model for Early Detection and Prevention of malignancy of Oral Cavity. The database of 1025 patients has been created and the required information stored in the form of 36 attributes. This deliver the technology and knowledge that users need to readily: (1) organize relevant data, (2) detect cancer patterns (3) formulate models that explain the patterns, and (4) evaluate the efficacy of specified treatments and interventions with the formulations.

Atiya kazi et al., [10] have presented the problem of assessing a given ontology for a particular criterion of an application, by typically determining which of several ontologies would best suit for the current application domain and also focuses on techniques, which incorporate ontology during the data mining process. It further proposes a methodology for building an ontology on the basis of the output of data mining result. The effects of the generated ontology are under research in order to improve the data mining process. Similarly, Henrihs Gorskis et al., [11] have proposed certain data mining techniques in order to discover their potential when used with automated ontology building. The ultimate goal is the reduction in the time requirement for the construction of any given ontology and necessity for expert consultation. This can be achieved by combining data mining and ontology engineering which can be used for ontology building. The C4.5 and COBWEB algorithms that are based on these techniques are the basis for automated ontology building.Similarly, Martin Ladvinka[12]Introduced Java OWL Persistence API (JOPA) that tackles the problems by providing Object Ontological Mapping (OOM) driven by integrity constraints designed on top of OWL ontologies, together with transactional processing and various backend implementations. Optimizing CRUDoperations of "frames" (objects) on the storage is crucial for efficient object-oriented access to OWL ontologies. For this purpose we defined an API (similar to JDBC) that encapsulates these optimized operations together with common transaction support, and analyzed the complexity of these operations.

III. PROGNOSIS OF BLOOD CARCINOMA USING DATA MINING TECHNIQUES

The following is the model of the proposed work where the input is taken as Cancer dataset and the output is obtained whether it is possibility to get cancer or not. This can be achieved by performing Data cleaning, classification and Semantic analysis of data. The Steps that are followed are as: The Input is taken as Cancer Dataset and the Dataset are taken for Data Cleaning in order to remove noise, irrelevant data. Then the Data have taken for Classification depends on attributes such as Complete blood count results as well as Blood smear results. The Classified data are performed with Semantic Relationships between given input and analyzed data.Then the output is determined as whether possibilities of cancer or not is shown in Figure 2.

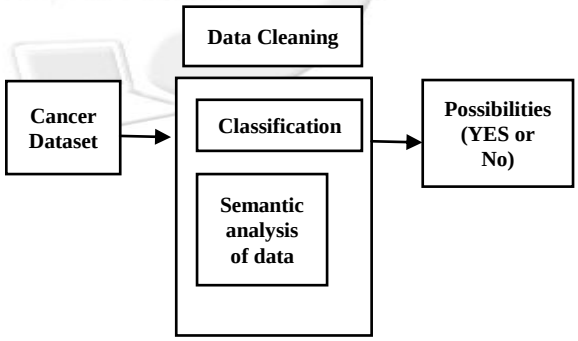

Figure 2: Proposed Work

IV. MATERIAL AND METHODS

A.Data Source

Massive related works, case studies and discussions with medical experts show that there are number of factors influencing leukemia.These factors are identified and taken as attributes for this study.

The data for this study was prepared by discussed some medical experts and oncologists, consisting of cancer and non-cancer patients data and they are preprocessed to work with this paper.

This data consists of more than 15 attributes consisting of Complete Blood count(CBC) Blood Smear results includes

Glucose,urea, Hemoglobin,Total WBC count,neutrophil,basophil,age,gender etc. These attributes are used to practice and develop the system. The attributes play vital role in predicting blood cancer. This data is stored in a knowledge base along with semantic relationship using Ontology are having ability to expand itself as new data which enters the through front end to system from which new knowledge is gained and thus the system becomes intelligent.

B. Pattern Mining and Classification

a) Decision Tree algorithm J48:

J48 classifier is a simple method for classification based on Decision tree. The tree is constructed to perform classification model. Once the tree is built, it is applied to each element in row in the database and results in classification for that row's element. While building a tree, the missing values are ignored by J48 i.e. the value by which prediction can be perform has taken as training sample. J48 Performs classification using Decision trees is incorporated in Weka toolis shown in Figure 3.

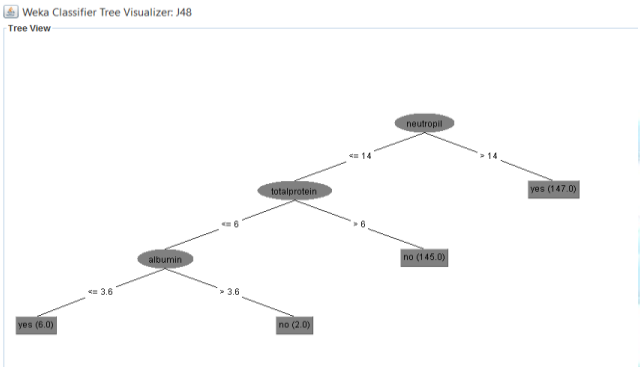

Figure 3:Decision Tree

b) Naïve Bayes classifier:

The Naïve Bayes algorithm is a simple classifier based on probability functions that calculates a set of probabilities which counts the frequency and combinations of values in a given dataset.

Naïve Bayes classifier is based on Bayes theorem and the total probability theorem. The probability that a document d with vector x=<x1, x2…xn>

$$P(h1|xi) = P(xi|hi).\frac{P(h1)}{P(xi|h1).P(h1)+P(xi|h2).P(h2)} (1)$$

Here, P(h1|xi) is posterior probability, while P(h1) is the prior probability associated with hypothesis h1.

$$P(xi) = \sum_{k=0}^{n} P(xi|hi)P(hi) \quad (2)$$

Thus, we have

$P(h1|xi) = P(xi|h1)P(h1)/P(xi).(3)$

The frequent pattern mined which satisfies the below condition are taken as significant Frequent Pattern

$Sw(i) = \sum_{i=0}^{n} m * t * n(4)$

WhereSw(i) is Significant Frequent Pattern; m is no. of data;t is no. of attributes;n is no. of times occurences of value.

Table 1. Risk scores for some important attributes that represent the significant patterns.

Attributes	Values	Risk score
Age	x<30 30 < x < 40 40< x < 60	3 4 5
Gender	Male Female Transgender	5 3 2
Glucose	X1<80 80<x1<140 X1>140	5 2 4
Hemoglobin	X2<13 13<x2<17 X2>17	5 2 4
Total WBC	X3<4000 4000<x3<1000 1000<x3<100000 X3>100000	2 2 3 5
Neutrophil	X4<10 10<X4<18 X4>18	3 1 4
Eosinophil	X5<10 10<x5<15 X5>15	2 1 5
monocytes	X6<10 12<x6<20 X6>20	2 1 4

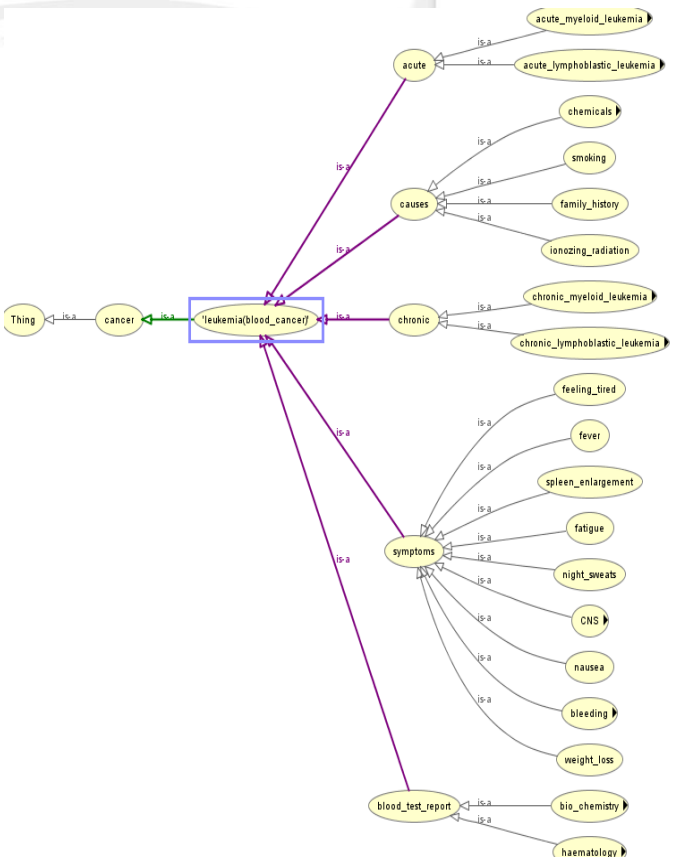

Figure 4: Blood cancer Ontology

The Figure 4, Shows that semantic relationships between attributes which provides the possibilities of blood cancer when this deployed with data mining algorithms.

V. CONCLUSION AND FUTURE WORK

The Data mining techniques are used for early prediction of blood cancer by classifying and clustering the data. Semantic relationship among the datasets are examined closely using Ontology. The dataset obtained from cancer test is compared with data which is already stored in database to predict the possibilities of cancer. By using CBC and blood smear results, one can predict that he/she is currently in possibilities of entrance stage of leukemia and further tests provide confirmation and types of leukemia. In Future by collecting original data from reputed oncology centers, prediction can be done easily.

VI.REFERENCES

[1] Ramachandran.P, Girija.N and Bhuvaneshwari.T, "Early Detection And prevention of cancer using Data mining techniques", International Journal of Computer Applications (0975-8887) Volume 97-No.13, July 2014.

[2] Ritu Chauhan,"Data clustering method for Discovering clusters in spatial cancer databases", International Journal of Computer Applications (0975-8887), Volume 10-No.6, November 2010.

[3] Dechang Chen, "Developing Prognostic Systems of Cancer Patients by Ensemble Clustering",Hindawi publishing corporation, Journal of Biomedicine and Biotechnology Volume 2009, Article Id 632786.

[4] Vikas Chaurasia, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", International journal of Computer Science and Mobile Computing (IJCSMC), Vol.3, Issue. 1, January 2014, pg.10-22, ISSN: 2320- 088X.

[5] Krishnaiah V.,"Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques",International Journal of Computer Science and Information Technologies, Vol. 4 (1) 2013, 39 – 45.

[6]Durairaj M and Deepika R., "Prediction of Acute Myeloid LeukemiaCancer using Data mining- A Survey",International Journal of Emerging technology and Innovative Engineering, Vol. 1, Issue 2, February 2015.

[7] Martin et al., "A Density-Based Algorithm for Discovering Clusters In Large Spatial Databases with Noise", Institute for Computer Science, University of Munich Oettingenstr. 67, D-80538 Munchen, Germany.

[8] Raymond T.Ng and Jiawei Han., "Efficient and Effective Clustering Methods for Spatial Data Mining", 20th VLDB Conference Santjago, Chile, 1994.

[9] Lookman Sithic.H and Uma Rani.R,"A Grouping of Cancer in Human Health using Clustering Data mining Technique",InternationalJournal of Innovative Engineering and Sciences, ISSN: 2319-9598,Volume-3 Issue-6, May 2015.

[10] Atiya Kazi et al., "An Ontology based Approach to Data mining", International Journal Engineering Development and Research, Volume 2, Issue 4, ISSN: 2321-9989.

[11] Henrihs Gorskis et al., "Ontology Building Using Data mining Techniques", Information Technology and Management Science, VERSITA, 10.2478/v10313-012-0024-5, 2012/15.

[12] Martin Ledvinka et al., "JOPA: Developing Ontology-Based Information Systems", Czech Technical University in Prague, Czech Republic.

[13] Gopala Krishna Murthy Nookala et al., "Performance analysis and Evaluation of Different Data mining Algorithms used for CancerClassification", International Journal Of AdvancedResearchinArtificial Intelligence, Vol. 2, No. 5, 2013.

[14] Neha Sharma and Hari Om,"Framework for Early Detection and Prevention Oral Cancer Using Data mining", International Journal Of Advances in Engineering and Technology, Sept 2012, ISSN: 2231-1963.

[15] Rajaraman Swaminathan et al.,"Cancer pattern and Survival in a Rural district in South India", International Journal of Cancer Epidemiology, Detection and prevention, November 2009, volume 33, Issue 5, Page 325-331.