MISMATCH CANCER COLOUR PREDICTION ANALYSIS ON BIG DATA

M.Alamelu | S.Karthikeshwar | V.Dhileepan | C.T.Gowtham

[1](Asst Prof, Dept of IT, Kumaraguru College of Technology, Coimbatore,India.)
[2](Department of IT, Kumaraguru College of Technology, Coimbatore,India.)
[3](Department of IT, Kumaraguru College of Technology, Coimbatore,India.)
[4](Department of IT, Kumaraguru College of Technology, Coimbatore,India.)

Abstract*—-"Leukemia" is one kind of the cancer spotted in the blood and bone-marrow of the human causing immense production of the infected white blood cells in the body, which targets the red blood cells to impair the whole blood cells of the system. The cause of the disease is still undetermined and if not identified at the early stage the probability of the risk is one the verge. The only method of diagnosis is the "blood test" at the regular interval with the watch full waiting. Since it requires a huge data base storage to maintain the records of the each patient to clearly determine the stages and provide the necessary medication for the patient. To recover this our proposed model cancer prediction rating system focuses to provide a solution to bring the awareness to the youth by predicting the cancer for the unaffected patient using the mismatch color prediction and identify the stage for the affected patients using the stage ranking algorithm. This also eliminates the data loss occurring during the data migration the pre-processing phase by fixing the 95% accurate data values in the record.*

Keywords— Stage Ranking Algorithm (SR Algorithm) , Color Prediction Rating system(CPR System) and Mismatch Color Prediction Algorithm(MCP Algorithm).

1. INTRODUCTION

Blood cancer, one major kind of the cancer which affects that affects the blood, bone marrow and lymphatic system of the human body. Being wild to attack the white blood cells and stops to balance the immune system and the healthy blood system in the body. The major issue faced during the treatment is that the regulated records that are needed to be maintained during each blood test and the lack of awareness among the youths who belong to the age group of 20- 45 who become the threat to this vulnerability. The objective of the proposed model is to overcome the data mismatch or the data loss occurring in the crucial medical reports due to data migration. The practice of manipulating the data set with the random values is replaced with the systematic method with the most accurate values and then identify weather he or she is affected with the Leukemia. The sample dataset contains the weekly blood-test report of the patient which contains the attributes such as the RBC count, WBC count, hemoglobin etc. In the proposed work, we determine the cancer percentage for the unaffected patient using the color mismatch prediction algorithm and the stage ranking algorithm for the affected patients in order to determine the weather the patient is in stage 1 , 2 or 3 and provide the medication based on which stage they fall. Color mismatch prediction algorithm uses the information gain, an statistical property measures the highest information and the entropy measures the amount of information in the attribute and constructs the decision tree and thus proves to increase the accuracy level of the classifier. The stage ranking algorithm compares the each and every values in the attribute with the range values of the different stages. This model is used in-order to provide more effective way of providing the result of the cancer percentage of the patient.

2. LITERATURE SURVEY:

Zakaria Suliman zubi (2014) et al, proposes that Neural Network Classifier and pattern recognition is been used to recognize Lung Cancer. The classifier algorithm helps to improve the treatment given to the patients. Here the behavioral patterns and the cancer symptoms are identified with the help of the test reports. The cancer affected patients details are obtained from the database which consists of the scan report and the blood samples of the patients. Based on the scan report the intensity level of cancer is figured out

Ada & Rajneet kaur (2013) et al, Lung Cancer is been detected by using the Data Mining algorithms. Certain Patterns are identified which helps to identify Lung Cancer. Here the author uses Data Mining algorithms like Support Vector machine and Neural Network classification algorithms. These two techniques give most accurate results even though several data mining techniques are applied to the same dataset. The classification technique helps to differentiate the cancer affected patients and the normal healthy patients who are not affected by Cancer.

S.M.Halawani (2012) et al, proposed that the clustering algorithm is been applied to the data that is been collected to understand the usage of mammogram. By using both the Data Mining techniques i.e Classification and clustering an acceptable outcome is achieved by using K-means clustering. Here Naïve Bayesian classification Algorithm is used since the dataset is of large quantity. The main reason

to use K-means Clustering algorithm is details of both the people who are affected by cancer at a high rate and people who are in the early stage are clustered and kept in the database. So they have some kind of similarities and they are clustered based on those aspects and the prediction is done.

Krishnaiah V (2013), proposed that Naïve Bayes algorithm helps to find the Lung Cancer patients. Decision tree algorithm is also used to predict the probability of patients affected with Lung Cancer. But when compared to Naïve Bayes, Decision tree algorithm result is quite difficult to understand. Thus the Naïve Bayes algorithm is used to get the maximum positive results and achieve the expected accuracy in predicting Lung cancer.

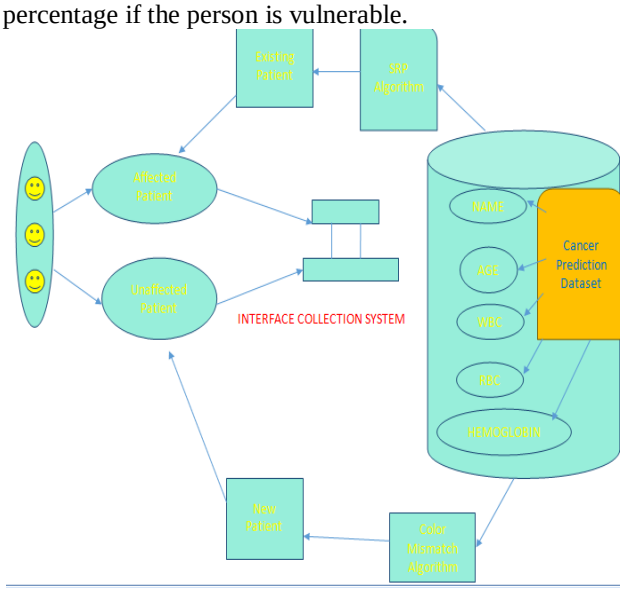
Rajashree Dash (2010) et al, proposed that hybridized K-means Clustering algorithm helps to cluster the large dataset. Here Clustering is done appropriately by reducing all the noise. The similar attributes are clustered correctly and based on the centroids the attributes fits in to the clusters. When the complexity of the dataset differs the Dimension also varies and therefore Clustering is done by using K-means algorithm.

Ritu Chauhan (2010) et al, proposed that application of HAC on K-means Algorithm helps to identify the clusters. The occurrence of Cancer is predicted accurately by applying the Hierarchical Agglomerative clustering with K-means algorithm. Here the Author carried out the experiment by using a software tool called TANGARA. This tool gives the statistical result after applying K-means and HAC algorithm. Results are analyzed by studying the statistical Graph Representation.

Dechang Chen (2009) et al, proposed the Hierarchical Clustering algorithm to predict the occurrence of Cancer. Patient's records are obtained and they are grouped by using the agglomerative Algorithm. Certain similar patterns are been identified and the results are generated by applying data mining techniques. The percentage of survival is finally predicted by identifying the patterns from the clustered group.

3. CANCER PREDICTION RATING SYSTEM:

Cancer being one of the most deadliest disease , there is no one major solution for the treatment. Thus the proposed CPR system initially identifies weather the patient is already affected or unaffected by the Leukemia disease.The affected patients for whom the stage is to be identified with the stage ranking algorithm and for the unaffected patients we determine the weather affected or not and then predict the cancer percentage using the color mismatch prediction algorithm. With the reference to the figure1.a the System user is broadly classified into the affected patients and the unaffected patients. The affected patients performs the SRP algorithm after mismatch is detected and fixed with the values, then the according to the stages the medication is prescribed. The unaffected patients undergoes the MCP algorithm to fix the values and then predict the cancer percentage if the person is vulnerable.



A. Data User Collection:

The user collection report contains the weekly report of the patients that are collected during the blood test. The dataset have been collected from both the affected patients and the unaffected patients and contains the input data with the 24 attributes of the weekly report of the blood test, which mainly includes:

- Name of the Patient
- Age
- Test dates
- White Blood Cells Count (WBC)
- Red Blood Cells Count (RBC)
- Hemoglobin(HGB)
- Hematocrit(HCT)
- Mean Corpuscular Value(MCV)
- Mean Corpuscular Hemoglobin(MCH)
- Platelet(PLT)
- Dextrose in water (DW)
- Mean Platelet Value(MPV)

B.System Users- Unaffected Patients:

The system user (Unaffected Patient), login-in using the provided username and password. Then, upload the predicted dataset into the CRP system. The main task performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. We examine the decision tree learning algorithm color prediction and implement this algorithm using C# programming

Finally, check the parameters Red blood cells(RBC),White blood cells(WBC), Haemoglobin(H), Platelets(p) If(RBC or WBC or H or P)=!NULL, Display the Box as RED.We first implement basic color prediction in which we dealt with the target function that has discrete output values. We

also extend the domain of color prediction to real valued output, such as numeric data and discrete outcome rather than simply Boolean value

.

A. *COLORMISMATCH PREDICTION ALGORITHM:*

The mismatch analysis has certain requirements has the same attributes must describe each example and have a fixed number of values must already be defined, that is, they are not learned by color prediction classes must be sharply delineated.

Given a collection S of c outcomes

Entropy(S) = S -p(I) log2 p(I)

where p(I) is the proportion of S belonging to class I. S is over c. Log2 is log base 2.

The range of entropy is 0 ("perfectly classified") to 1 ("totally random").

Gain(S, A) is information gain of example set S on attribute A is defined as

Gain(S, A) = Entropy(S) - S ((|S_v| / |S|) * Entropy(S_v))

Where:

S is each value v of all possible values of attribute A

S_v = subset of S for which attribute A has value v

|S_v| = number of elements in S_v

Then the data set is considered as a single matrix fixed with all the values then information gain is identified that is having the highest value and then the entropy condition is applied in order to construct the decision tree. By using the minimum and maximum range value calculated with the repeated processing of random values then the value is randomly fixed, the cancer is detected and if yes the percentage is identified .

C. System Users- Affected Patients:

The system user(Affected Patient), login-in using the given username and password. Once the weekly blood report is uploaded into the CRP system, mismatch analysis is performed using the inductive methods and the all the values are fixed in the data set for the further proceeding.

Using the SR algorithm, a classification analysis is performed on the dataset with the attributes and each with the values fixed for each week the percentage can be calculated and the stage can be identified for the improvement analysis for the particular week selected.

A. STAGE RANKING ALGORITHM:

- The ranges for each stages are previously identified by fixing the pivot value for the each stage of the WBC count.
- The classification is performed based on the three stages, namely:

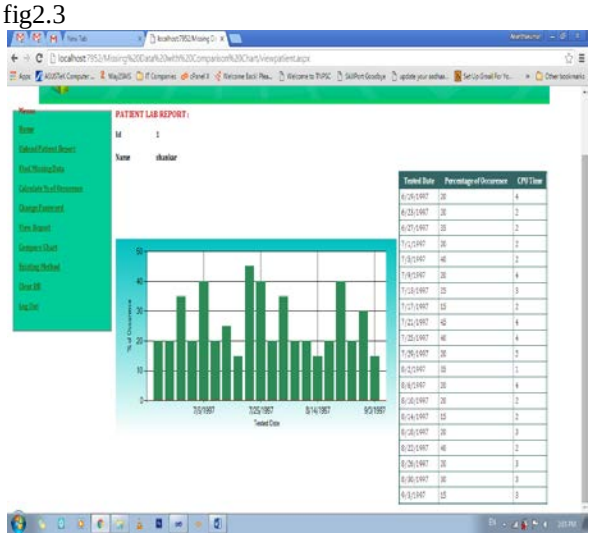
Stage 1 : Low risk (WBC is slightly high than normal)
Stage 2 : Intermediate risk (WBC count is high and Lymh node is more than 2)
Stage 3: High risk (PLT is very low and WBC is very high)
3.3.4 Once the stage is identified depending on the WBC count of the patient, the pre-defined medication is
provided and the improvement can be identified for the precautionary measures.

The CPR System, is then loaded with the excel sheet having the 21 weekly blood test reports of the patient having the 24 attributes. Then it is then identified with mismatch values by red color in fig2.1.
fig2.1.

RESULTS:

The red color is then turned to green when the mismatch values as referred in fig2.2 the dataset is fixed with the 95% accuracy values. Once the system is processed all the values, the cancer percentage ratio is obtained,
fig2.1
both for the whole 21 weeks and also to the appropriate week date(fig2.3).

fig2.3



3. CONCLUSION:

The eleventh most common cancer “Leukemia” can kills a quarter lakh people every year. So there is a wide range of necessity to analyse the causes and the reason to become aware about the deadly disease.The future enhancement for the proposed model are the prototype workbench which has been developed to provide an integrated approach to the application. The design rationale and the potential use of the system are justified. Finally, future directions and further enhancements of the workbench are : Can implement for web based application, handshakes with Inductive learning algorithm,improvisations can be done in the performance Evaluation, prediction can be done for all kind of diseases and finally in case of huge range of data set, data load balancing can be done

REFERENCES:

A Literature Review of Data Mining Techniques used in Healthcare Databases.

[1]. Ada and Rajneet kaur “Using some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient ” International Journal of Computer Science and Mobile Computing IJCSMC,Vol. 2,Issue.4,April 2013, pg 1-6,ISSN 2320-088X

[2]. V.krishnaiah “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, ISSN:0975-9646

[3]. Charles Edek “Comparitive Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability ” Mediterranean journal of Social Sciences Vol 3 (14) November 2012, ISSN:2039-9340

[4]. Rajashree Dash “A Hybridized K-means Clustering approach for high dimensional dataset” International Journal of Engineering, Science and Technology Vol.2, No.2,2010,pp.59-66

[5]. Ritu Chauhan “Data Clustering method for Discovering Clusters in Spatial Cancer databases” International Journal of Computer Applications Volume 10-No-6,November 2010

.