# Decision support system for cancer diseases using text and opinion mining techniques.

Ashok Kumar M
PG Student, Advance Information Technology
REVA UNIVERSITY, Bangalore, INDIA
Email: mashokm@gmail.com

A.Ananda Shankar
Associate Professor, C&IT,
REVA UNIVERSITY, Bangalore, INDIA
Email: anandshankar@reva.edu.in

***Abstract* - This paper describes the implementation and the application of decision support system to detect the blood cancer diseases with minimal blood test results. Opinion Driven Decision Support System refers to the use of large amounts of opinions to facilitate detection of any pattern by normal user. The main objective is to help the doctor to get the expert doctor opinion based on fast or historical data that are collected from the cancer patient. Opinion mining, will be used to analyze the opinion of people about a particular topic or data. Basically we need to automate the process of extracting the emotions behind the text, written in natural language, by the expert which will be useful in classifying the expert's attitude towards the topic.**

**Key Words: Blood Cancer, Decision support, Opinion mining**

## I. INTRODUCTION

Blood cancer is among the common cancer disease found in India with various types and stages. It would be better if we detect the blood cancer malignancies at early stages. Initial symptom is bleeding and serious infection. Patient need immediate medical attention for symptoms such as uncontrolled bleeding, severe swatting, breathing difficulty, blue or pale fingernails or lips, fast hear rate etc.

There are three types of blood cancer: [2]

1. Leukemia: Cells start multiplying and affects the bone marrow and blood production rate will reduce and results in white blood cell count
2. Lymphoma: Is the cancer that affects the lymphocytes. It is group of blood cell tumors that generate from lymphatic cells
3. Myeloma: In this type blood plasma is affected by the cancerous cells

There are four stages of blood cancer is divided based on metastasis. Mainly these are classified as

Stage 1: In this stage lymph nodes will get enlarged, this happens because of sudden increase of the number of the lymphocytes. The risk at this stage is very less.

Stage 2: In second stage spleen, liver and lymph nodes will get enlarged, also the growth of the lymphocytes is very high in this stage.

Stage 3: In third stage anemia develops and above mentioned cell found enlarged. Here we may see more than two organs get affected in this stage.

Stage 4: In fourth stage rate of blood platelets will rapidly decreases. This will start affecting the lungs along with the other organs which already affected in the earlier stage of cancer.

## II. DATA MINING

Data mining uses its strong predictive models and algorithms which help in exploring, selecting and discovering the unknown/hidden information from a set of large data[3]. According to some literature reports that to predict cancer diseases and to make cancer disease decision support systems, developer/researches use predictive models of data mining.

Data mining methodologies

To begin with we have to have clarity on the problem definition and make sure amount of data collected is sufficient for analysis. During the analysis if it we find data to be insufficient, then the process has to be reiterated. For reiteration we may have to have more test data with different age group to get more information and clarity.

Once we have sufficient data, we have to create a data model. Data modelling is organizing the data elements and identifying the relationship among the elements. Since a computer software runs over the data collected this steps becomes important.

Sentiment analysis: Sentiment analysis, also known as opinion mining, is to analyze the understanding and opinion of people about a particular topic or data. Basically we need to automate the process of extracting the opinion behind the text, written in natural language, by the expert which will be useful in classifying the expert's attitude towards the topic.

Two basic types of sentiment analysis are

Subjectivity/Objectivity Identification - The given text is classified into either Objective or Subjective.

Feature/Aspect based - Identifying the opinion expressed on different features or aspects of entities.

Different levels of analysis are

Document Level - Gives the overall sentiment for the entire document.

Sentence Level - Provides overall sentiment to each sentence.

Entity and Aspect Level - Granular level of analysis considers each entity in the sentence for sentiment analysis.

Approaches for sentiment analysis can be classified as

Machine learning approaches - This is comparatively accurate and adaptable. This is again classified into supervised and unsupervised techniques. More reliable is the supervised technique. Different stages in this technique is - Data Collecting, Pre Processing, Training data, Classification and Results.

Semantic orientation approaches - This approach works by considering positive and negative sentiment words and phrases.

Lexicon based approaches - This uses a predefined and pre compiled list of sentiment terms, keywords maintained as dictionary. The matches are taken from this dictionary.

Other unsupervised approaches - This should consider the advantage of each of the approaches and give an accurate output.

## III. PROBLEM ANALYSIS

In modern days most of the doctors are very busy with their schedule, hence they may not able to provide their valuable time to study the patient blood test report and provide their suggestion. This is really a main concern to many patient, hence I thought of developing a solution that will give the expert doctor opinion based on patient blood test report. Also this system supports what should be next course of action.

Work Carried:

A. Collect Cancer detection test methodologies such as Complete blood count (CBC), Blood protein testing, Tumor marker tests, Tumor marker tests

B. Collect all test cases performed on each methodologies and collect the test range to detect the level of diseases. In blood cancer most of the clinical tests are performed on blood sample, different tests that are performed on blood is detailed in next section.

C. Create a corpus of cancer detection with different doctor's opinion. Store this in database with plain text. All identified test with actual data with doctor opinion is stored in database. Here I have used windows based desktop user interface to collect the data from user and MS SQL database for storage.

•Create decision support system that will analyze the any input blood test data and analyze this using predefined corpus that contains expert doctor opinion.

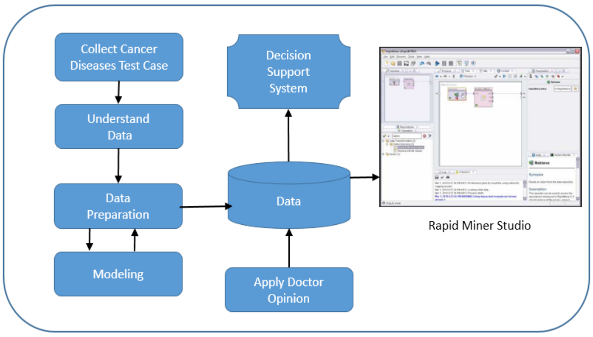
•Calculate the performance of this system with huge amount of data that is more than 50K record.

•With help of Rapid Miner tool expose this newly collected corpus data for any test data

•Evaluate the test data by using Rapid Miner tool using text mining algorithm –Text Processing

## IV. IMPELMENTATION AND ALOGORITHM

- Accept user provided blood sample test results
- Calculate the final results or levels by comparing blood test result corpus.
- Provide the doctor opinion by comparing corpus created by doctor opinion.
- Finally provide the rank for each opinion along with what should be the next step.
- Further data will be analyzed using K-means algorithm using Rapid miner tool.

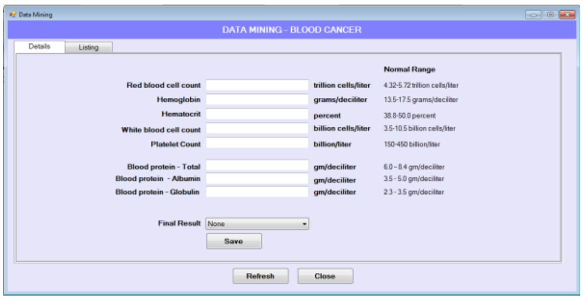


Sample Blood test methods and its valid ranges

| Complete blood count (CBC) | |
|---|---|
| Red blood cell count | 4.32-5.72 trillion cells/liter |
| Hemoglobin | 13.5-17.5 grams/deciliter |
| Hematocrit | 38.8-50.0 percent |
| White blood cell count | 3.5-10.5 billion cells/liter |
| Platelet count | 150-450 billion/liter |
| Blood protein testing | |
| Total | 6.0 – 8.4 gm/deciliter |
| Albumin | 3.5 - 5.0 gm/deciliter |
| Globulin | 2.3 - 3.5 gm/deciliter |

## V. RESULTS AND CONCLUSIONS

Data Collection User Interface: Using this user can enter a clinical test results, this tool will utomatically give the final result such as Negative, Level 1, Level 2, Level 3 OR Level 4. Here doctor can give their opinion using text input. This text data will be further used to get the expert opinion using text mining method.

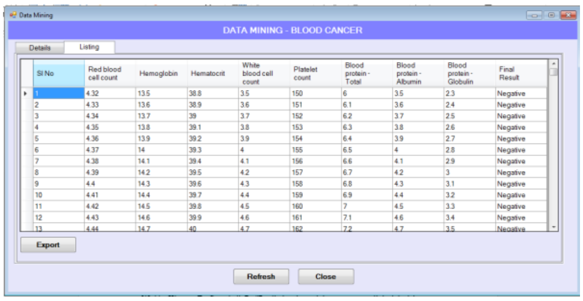


DATA MINING - BLOOD CANCER

| | | | Normal Range |
|---|---|---|---|
| Red blood cell count | | trillion cells/liter | 4.32-5.72 trillion cells/liter |
| Hemoglobin | | grams/deciliter | 13.5-17.5 grams/deciliter |
| Hematocrit | | percent | 38.8-50.0 percent |
| White blood cell count | | billion cells/liter | 3.5-10.5 billion cells/liter |
| Platelet Count | | billion/liter | 150-450 billion/liter |
| Blood protein - Total | | gm/deciliter | 6.0 – 8.4 gm/deciliter |
| Blood protein - Albumin | | gm/deciliter | 3.5 - 5.0 gm/deciliter |
| Blood protein - Globulin | | gm/deciliter | 2.3 - 3.5 gm/deciliter |

Final Result: None

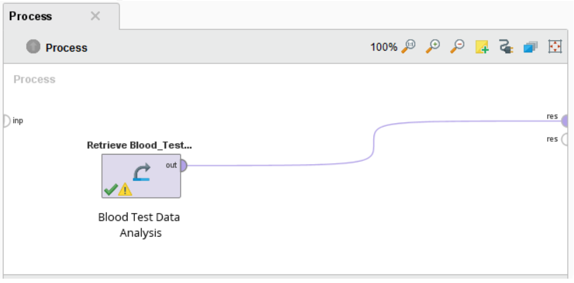
Data View User Interface: Using this view user can view all old data along with expert opinion. This data can be download using excel format for further analysis in rapid miner tool.



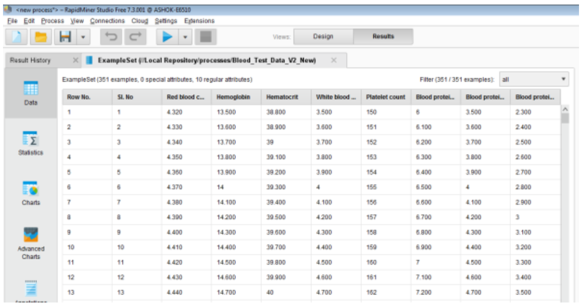
DATA MINING - BLOOD CANCER

| Sl No | Red blood cell count | Hemoglobin | Hematocrit | White blood cell count | Platelet count | Blood protein - Total | Blood protein - Albumin | Blood protein - Globulin | Final Result |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 4.32 | 13.5 | 38.8 | 3.5 | 150 | 6 | 3.5 | 2.3 | Negative |
| 2 | 4.33 | 13.6 | 38.9 | 3.6 | 151 | 6.1 | 3.6 | 2.4 | Negative |
| 3 | 4.34 | 13.7 | 39 | 3.7 | 152 | 6.2 | 3.7 | 2.5 | Negative |
| 4 | 4.35 | 13.8 | 39.1 | 3.8 | 153 | 6.3 | 3.8 | 2.6 | Negative |
| 5 | 4.36 | 13.9 | 39.2 | 3.9 | 154 | 6.4 | 3.9 | 2.7 | Negative |
| 6 | 4.37 | 14 | 39.3 | 4 | 155 | 6.5 | 4 | 2.8 | Negative |
| 7 | 4.38 | 14.1 | 39.4 | 4.1 | 156 | 6.6 | 4.1 | 2.9 | Negative |
| 8 | 4.39 | 14.2 | 39.5 | 4.2 | 157 | 6.7 | 4.2 | 3 | Negative |
| 9 | 4.4 | 14.3 | 39.6 | 4.3 | 158 | 6.8 | 4.3 | 3.1 | Negative |
| 10 | 4.41 | 14.4 | 39.7 | 4.4 | 159 | 6.9 | 4.4 | 3.2 | Negative |
| 11 | 4.42 | 14.5 | 39.8 | 4.5 | 160 | 7 | 4.5 | 3.3 | Negative |
| 12 | 4.43 | 14.6 | 39.9 | 4.6 | 161 | 7.1 | 4.6 | 3.4 | Negative |
| 13 | 4.44 | 14.7 | 40 | 4.7 | 162 | 7.2 | 4.7 | 3.5 | Negative |

Data Model using Rapid Miner tool: K-means clustering is performed on data set using Rapid miner and analysis done on different stages/data. This helps to predict the different set of blood test data on corpus data collected.



Results View:


<new process*> – RapidMiner Studio Free 7.3.001 @ ASHOK-E6510
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Result History
ExampleSet (//Local Repository/processes/Blood_Test_Data_V2_New)
Data
Statistics
Charts
Advanced Charts

ExampleSet (351 examples, 0 special attributes, 10 regular attributes)

Filter (351 / 351 examples): all

| Row No. | Sl. No | Red blood c... | Hemoglobin | Hematocrit | White blood ... | Platelet count | Blood protei... | Blood protei... | Blood protei... |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 4.320 | 13.500 | 38.800 | 3.500 | 150 | 6 | 3.500 | 2.300 |
| 2 | 2 | 4.330 | 13.600 | 38.900 | 3.600 | 151 | 6.100 | 3.600 | 2.400 |
| 3 | 3 | 4.340 | 13.700 | 39 | 3.700 | 152 | 6.200 | 3.700 | 2.500 |
| 4 | 4 | 4.350 | 13.800 | 39.100 | 3.800 | 153 | 6.300 | 3.800 | 2.600 |
| 5 | 5 | 4.360 | 13.900 | 39.200 | 3.900 | 154 | 6.400 | 3.900 | 2.700 |
| 6 | 6 | 4.370 | 14 | 39.300 | 4 | 155 | 6.500 | 4 | 2.800 |
| 7 | 7 | 4.380 | 14.100 | 39.400 | 4.100 | 156 | 6.600 | 4.100 | 2.900 |
| 8 | 8 | 4.390 | 14.200 | 39.500 | 4.200 | 157 | 6.700 | 4.200 | 3 |
| 9 | 9 | 4.400 | 14.300 | 39.600 | 4.300 | 158 | 6.800 | 4.300 | 3.100 |
| 10 | 10 | 4.410 | 14.400 | 39.700 | 4.400 | 159 | 6.900 | 4.400 | 3.200 |
| 11 | 11 | 4.420 | 14.500 | 39.800 | 4.500 | 160 | 7 | 4.500 | 3.300 |
| 12 | 12 | 4.430 | 14.600 | 39.900 | 4.600 | 161 | 7.100 | 4.600 | 3.400 |
| 13 | 13 | 4.440 | 14.700 | 40 | 4.700 | 162 | 7.200 | 4.700 | 3.500 |

Chart Plot: Final result Vs Number of test cases

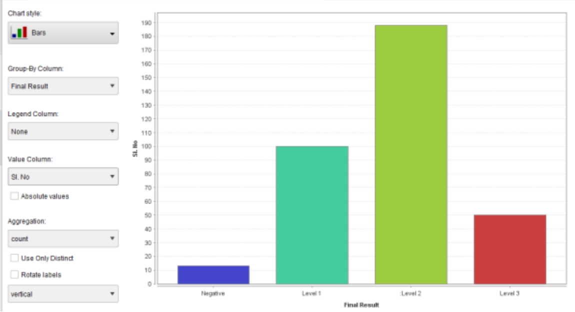


Chart style: Bars

Group-By Column: Final Result

Legend Column: None

Value Column: Sl. No

□ Absolute values

Aggregation: count

□ Use Only Distinct

□ Rotate labels

vertical

Based on the above mentioned design and implementation, it can be concluded that doctor can get the final result that is based on level of cancer disease by providing pre-defined clinical tests and its value.

Also this gives more option for doctor what should be the next step for each level of result. This will be achieved by taking doctor suggestion and mine this input using text mining algorithm.

## ACKNOWLEDGMENTS

Authors express sincere thanks to A. Ananda Shankar and Prof. Shilpa Chaudhari for their great support and guidance. Authors appreciate the reviewer's useful comments.

## REFERENCES

[1] Abu Khousa, E.; Campbell, P., "Predictive data mining to support clinical decisions: An overview of heart disease prediction systems," Innovations in Information Technology (IIT), 2015 International Conference on , vol., no., pp.267,272, 2012.

[2] Types of cancer types and different stages http://www.indushealthplus.com/blood-cancer-types-stages.html

[3] Giudici, P.: "Applied Data Mining: Statistical Methods for Business and Industry", New York: John Wiley, 2012.

[4] R. Andrews, J. Diederich, A. B. Tickle," A survey and critique of techniques for extracting rules from trained artificial neural networks", Knowledge-Based Systems,vol.- 8,no.-6, pp.-378-389,1995.

[5] Clinical data test and normal range reference
http://www.mayoclinic.org/diseases-conditions/cancer/in-depth/cancer-diagnosis/art-20046459st
http://www.mayoclinic.org/tests-procedures/complete-blood-count/details/results/rsc-20257186
http://www.bloodbook.com/ranges.html.