Efficient Diagnosis of Leukemia using Neural Networks

V. Rajalakshmi[1]*, G.S. Anandha Mala[2], Mandava Chatana[2] and D. Saranya[1]

[1] Faculty of Computing, Sathyabama University, Chennai – 600119 India.
[2]Department of CSE, Easwari Engineering College, Chennai – 600089 India.

(Received: 03 January 2016; accepted: 08 February 2016)

Leukemia is one of the sub types of blood cancer, where if untreated it leads to death in weeks or months .In order to detect Leukemia in a person, we need to find out the symptoms, number of cells and certain chemicals present in the blood. When the values of these cells and chemicals are abnormal than the healthy range of values, the person is said to have Leukemia. There are many health service websites like webMD, doctorspring, icliniccare offering health seekers to interact anonymously with doctors at anytime from anywhere in the world and get solutions. These websites acts like a blog where health seeker can post their questions and different doctors give solutions as comments below. Since many doctors reply to single person, it conflicts the solution, where the health seeker will not know which solution to consider and misleads the health seeker to take wrong steps. So, we have created an application by generating a mathematical formula, by using the healthy range of values of cells and chemicals present in the body to diagnosis Leukemia using MATLAB. When the patient's cells and chemical values are abnormal when compared to this healthy range of values, the application detects Leukemia in the patient. The main advantage of our application is, it accurately detects the presence of Leukemia in a person.

Keywords: Leukemia, blood cancer, K-means clustering, neural network, Rule based classification.

Today blood cancer has been increasingly identified as one of the major causes of deaths. The different types of the blood cancer include Leukemia, Lymphoma and Myeloma. Researchers stated that Leukemia was the 11th most common cause of cancer-related death in the world. It commonly occurs among children and adult above 40 years of age. Leukemia affects the white blood cells. The infected white blood cells will occupy the entire bone marrow and damage the bone marrow and also these abnormal cells will spoil the normal unaffected white blood cells, red blood cells and infects the whole body. The infected body will automatically lose its strength and becomes weak which leads to anemia. Figure 1 shows the difference between normal blood cells and affected abnormal blood cells.

Further Leukemia is divided into three types[12] as

- Acute Lymphocytic Leukemia: Acute lymphocytic leukemia occurs among children of 1 – 12 years and older people. It's a fast-growing cancer called as lymphoblast. It occurs when the bone marrow produces a large number of immature lymphoblasts. The abnormal lymphoblasts will grow quickly and replace normal cells in the bone marrow.
- Chronic Myelogenous Leukemia (CML):Chronic Myelogenous Leukemia will affect the stem cells present in the bone marrow. Chronic MylogenousLeukemia develops when these stem cells does not work properly or behave in an

* To whom all correspondence should be addressed.
E-mail: rajalakshmi.it@sathyabamauniversity.ac.in

abnormal way

- Chronic Lymphocytic Leukemia (CLL): This type of leukemia affects the older people above 50 years of age who have a medical history of blood pressure and diabetics.

Evaluation of candidate control genes for diagnosis and residual disease detection in leukemic patients for all the above types using 'real-time' quantitative reverse-transcriptase polymerase chain reaction (RQ-PCR)[13] is chosen as the base method for detecting Leukemia. The method was designed in Europe using amplified control gene measurement.

Related work

There are various methods for detecting Leukimia in the litereature. They are discussed as follows: Girija[1] made an attempt to detect and prevent cancer in early stage and suggests what kind of therapy should be given using data mining techniques. It confesses that it is an effective way to reduce cancer deaths in early stage and it helps doctor to concentrate on particular therapy.Durairaj [3] attempted to predict acute myleloid leukemia cancer using data mining- A survey and their paper compares the accuracy level of the reviewed papers and classification of algorithms.Yu Ping Wang[5] developed CS based classification method for subtyping leukemia using gene expression data. This gives an 100% accuracy. It helps to save large calculations and storage data when processing large data sets

Eldosoky[6] have made an attempt to detecte leukemia by using ultra wide band pulse.Van Dongen,[7] have analyzed the of fusion gene transcripts from chromosome aberrations in acute leukemia for detection of minimal residual disease and Investigated the minimal residual disease in acute leukemia.Vinod[8] haveanalyzed the gene in patient's bone marrow prior to leukemia remission, a mast cell disease accociated with acute myeloid leukemia.Paul OP[9] et al. have made an attemot todetecte the intact prostate cancer cells in the blood of men with prostate cancer and developed a procedure identify and characterize the living prostate cancer cells in the blood of patients.Kawasaki, Ernest[10] diagnosed the chronic myeloid and acute lymphocytic leukemias by detection of mRNA sequences. Scholefield[11] have Randomised the controlled trial of faecal-occult-blood screening for colorectal cancer.Reulen, Raoul C., David L. Winter, Clare Frobisher, Emma R. Lancashire, Charles A. Stiller, Meriel E. Jenney, Roderick Skinner, Michael C. Stevens, Michael M. Hawkins, and British Childhood Cancer Survivor Study Steering Group[12] have made an attempt to detect Long term specific mortality among survivors of childhood cancer.

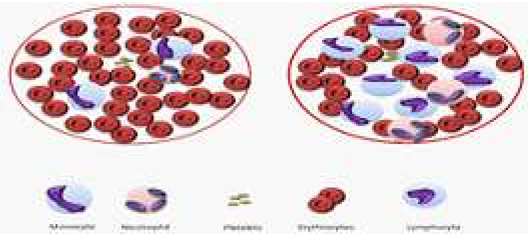


Fig. 1. Leukemia Versus Normal Blood cells

Problem Definition

Using the results of various blood tests, two methods using data mining based detection [LDCS] and Neural Network based detection [NNLD] are designed. Both these methods are based on the actual cases used as training records. Their performance is compared with respect to the number of training data used and their efficiency.

Preliminaries

•Clustering by K-Means Clustering Algorithm

Let X = $\{x_1,x_2,x_3,……..,x_n\}$ be the set of data points and V = $\{v_1,v_2,…….,v_c\}$ be the set of centers.

1) Arbitrarily choose *'c'* cluster centers.

2) Find the Euclidean distance between each data point and cluster centers.

3) Assign every data point to the cluster center whose distance from the cluster center is minimum of all the clustercenters.

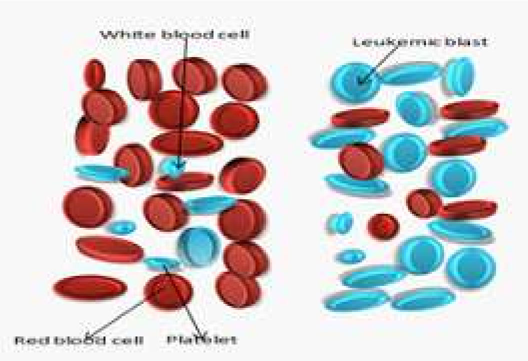


Fig. 2. Difference between Normal blood cells and Leukemic cells

4) Recalculate the new cluster center using:

$$v_i = \left(\frac{1}{c_i}\right)\sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

•Classification by sequential covering Rule based system:

Rules are learned sequentially, each for a given class Ci will cover many tuples of Ci but none (or few) of the tuples of other classes

Steps:

1) Rules are learned one at a time

2) Each time a rule is learned, the tuples covered by the rules are removed

3) Repeat the process on the remaining tuples until *termination condition*, e.g., when no more training examples or when the quality of a rule returned is below a user-specified threshold.

Neural networks are similar to biological neural networks in performing functions collectively in parallel by the units, instead of a clear delineation of subtasks to which individual units are assigned. An ANN is typically defined by three types of parameters:

1. The interconnection pattern between the different layers of neurons

2. The learning process for updating the weights of the interconnections

3. The activation function that converts a neuron's weighted input to its output activation.

In supervised learning, we are given a set of example pairs $(x, y), x \in X, y \in Y$ and the aim is to find a function $f : X \rightarrow Y$ in the allowed class of functions that matches the examples.

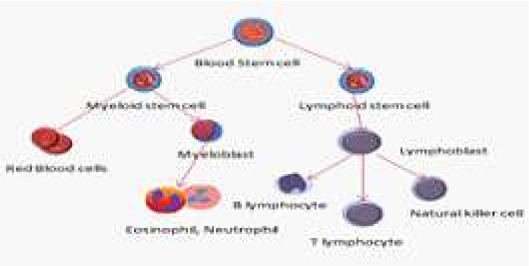


Fig. 3. Distribution of the Blood cells

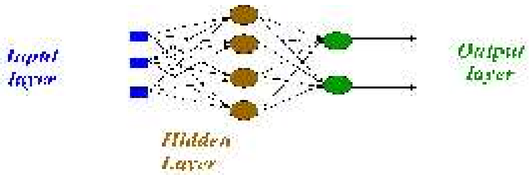


Fig. 4. Neural Network System

Implementation

Figure 5 shows the architecture of LDCS system, Database is a collection of large amount of information. Database management system helps us to access the information needed from database. So, first step is to collect the information from the database and then the information is filtered. Filtering is a process to eliminating the irrelevant data, and then the filtered data is preprocessed. Preprocessing which is similar to filtering, where the data is screened very carefully to eliminate redundant information. The preprocessed data is again stored in the database and an algorithm or formula is applied with required attributes to solve the problem and to detect the disease and sensitivity level. The cells and chemicals we use for our analysis are RBC, WBC, Uric Acid, BUN, ALT, AST, Creatinine, and Phosphorus.

Algorithm for LDCS System

Step1: Define training set of data to train the system.

Step2: The chosen data records are used to train the system by using k means clustering.

Step3: The three groups of data obtained after clustering are analysed and their boundaries are identified.

Step4: The lower and upper boundaries are calculated for every attribute for every cluster. It can be done by,

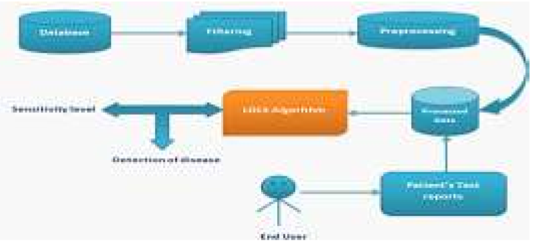


Fig. 5. Architecture of the system

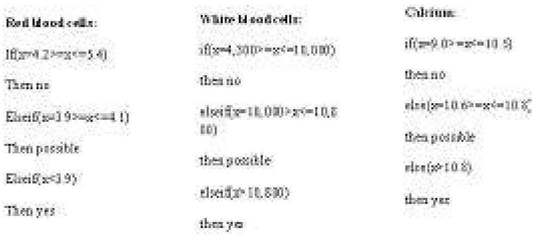
```
B=bwboundaries(BW);
B=bwboundaries(BW,conn,options);
```

Step5: Decision Rules For Classification are formed Using (If………..Else……..Then):

There are two different readings of attributes for both female and male.

Female:

Similarly we have to write if else coding for calcium, uric acid, alt, ast , bun, and phosphorus. The procedure is repeated for all blood cells counting for male. A list of all If…Then… Rules are developed.



```
Red blood cells:
If(x>=4.2>=x<=5.4)
Then no
Elseif(x>=3.9>=x<=4.1)
Then possible
Elseif(x<3.9)
Then yes

White blood cells:
if(x>=4,300>=x<=10,000)
then no
elseif(x>=10,000>x<=10,8
00)
then possible
elseif(x>10,800)
then yes

Calcium:
if(x>=9.0>=x<=10.5)
then no
else(x>=10.6>=x<=10.8)
then possible
else(x>10.8)
then yes
```

Step6: The test data is used to validate the system.

A Set of records are chosen using stratified sampling method and used for training the using clustering algorithm or neural network and a system for detecting leukimia is built. The quality and number of training records make the performance of the system .

Experimental Results

All the points with green color are people not having leukemia. All the points with red color represents the data about number of people having leukemia. All the points with yellow color defines people with a possibility of having leukemia.There is one more case, where few points lay between the boundaries. These are in either or condition. Then we should find the difference between the boundaries.

Table 1. Sample training data used in K- Means algorithm

Age	Diabetes	BP	Family History of Cancer	RBC	WBC	Cal	Uric Acid	ALT	AST	BUN	Phosphorus	Cancer Detection
9	No	No	Yes	4.5	10,111	10.6	7.3	39	36	22	4.4	Possible
25	Yes	No	Yes	5.1	5,500	9.9	6.1	9	11	12	3.2	No
66	Yes	No	No	1.2	3,000	11.2	1.1	5	4	30	1.2	Yes
88	Yes	Yes	No	3.4	2,000	18	2.6	40	45	36	7.1	Yes
13	No	No	Yes	5.8	4,600	10	6.6	20	25	15	3.1	No
19	Yes	No	Yes	4.5	9,000	9.9	2.5	21	30	10	3.1	No
6	No	No	Yes	4.1	10,222	10.6	2.2	38	35	21	4.2	Possible
77	Yes	Yes	No	3.3	11,000	11.9	2	41	46	33	5.3	Yes
78	Yes	No	No	4.6	10,500	10.7	2.3	39	37	22	4.1	Possible
66	Yes	Yes	No	2.9	25,000	13.3	1.1	48	46	64	1.6	yes

Table 2. Sample data to be predicted for the disease

Age	Diabetes	BP	Family History of Cancer	RBC	WBC	Cal	Uric Acid	ALT	AST	BUN	Phosphorus	Cancer Detection
23	No	No	No	5.1	6,500	10	5	14	14	13	2.6	?
33	Yes	Yes	Yes	4.6	10,333	10.6	3.3	40	36	22	4.3	?
56	No	No	Yes	1.1	12,754	12.4	10.1	43	48	36	5.9	?
64	Yes	Yes	No	1.5	14,653	13.2	9.9	65	65	33	5.8	?
23	No	Yes	Yes	4.8	5,777	9.3	7.1	25	18	14	3.3	?
10	No	No	Yes	5.1	7,654	9.4	6.6	30	25	11	4	?
17	No	No	No	2.2	23,880	14.1	11.3	54	47	35	4.9	?
23	Yes	Yes	Yes	5	4,791	10.1	5.4	22	26	16	3.2	?
44	Yes	Yes	No	2.6	15,000	15	1.1	49	57	43	5.1	?

Let the value of the unknown point plotted in between be A and B. Let the point lie between the boundaries X and Y. Such points are identified by their distance between the boundaries of every cluster.

If((X-A)>>(Y-A)) then Point belongs to Y else point belongs to X.

The methods LDCS and NNLD are implemented and their performances are measured by considereing 1000 records and calculated. The performance of LDCS and NNLD varies according to the number of training records taken. As the number of training records increases, more knowledge is achieved by the system and hence performs better. But as the training records increases, the cost of the system also increases with clustering time in LDCS and number of neurons in NNLD. Hence the number of records for training should be reduced

Table 3. Performance comparison of the methods

Methods	Training Records	Efficiency(%)
RQ-PCR	500	93.5
LDCS	50	63.4
	200	78.75
	500	83.6
	1000	93.45
NNLD	50	76.8
	200	85.96
	300	92.44
	500	94.88

Data mining technique is mainly used to solve the problems of large amounts of data.Large sets of data can be stored using data mining.It is used to discover relationship among large sets of data which is unknown.We can solve the data using statistical method or mathematical method. In our method we are using the clustering technique to detect the Leukemia.There are different types of algorithms in data mining like Decision making tree, Bayesian classification tree, Rule based classification.

Decision making tree is one of the best algorithms in data mining.The decision tree is used to generate patterns in the datasets. These patterns are used to detect leukemia easily.Decision tree is something which is similar to a flowchart to classify a new data. At each and every point of the flow chart the main thing we need to check is the attributes. And these attributes are used to classify the entire algorithm.In this decision making algorithm internal nodes are denoted by rectangles and leaf nodes are denoted by ovals.

Clustering is a method of dividing large amount of datasets into smaller sizes. There are different types of clustering techniques in which we are using k-means clustering technique. In k-means clustering, the number of clusters required is found out first. Then using an algorithm are code the association or deassociation of clusters will be done.

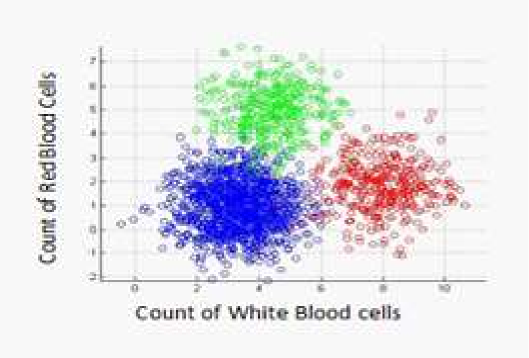


Fig. 6. Clustering of data to detect number of people having blood cancer

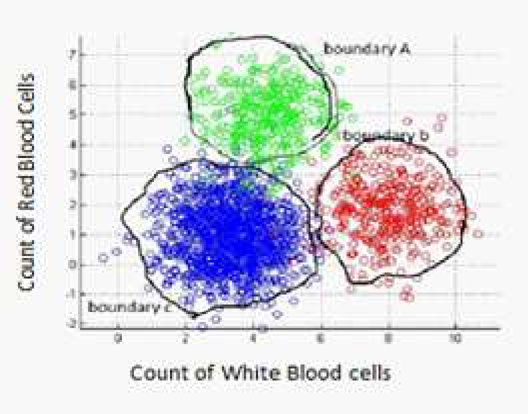


Fig. 7. Accurate detection of Leukemia

CONCLUSION

In this paper, two methods – LDCS and NNLD are developed for detecting Leukemia by using the techniques of clustering and classifications of data mining. The performance of these methods are compared with the existing method RQ-PCR. The performance show that NNLD provides higher efficiency with minimum number of training records. By using this method, it is easy to detect the Leukemia in an early stage. Since many patients ignore several tests because they charge more, this application helps people who can't afford. The procedure can also be extended for other diseases with their associated tests. The method can be used as a tool for various applications that provides medical advices based on the symptoms and tests values.

ACKNOWLEDGEMENTS

The authors are thankful to the management of Sathyabama University-India for providing us with all the resources necessary for this work.

REFERENCES

1. Ramachandran, P., N. Girija, and T. Bhuvaneswari. "Early Detection and Prevention of Cancer using Data Mining Techniques." *International Journal of Computer Applications* 2014; **97**.13.
2. Cheson, B. D., Cassileth, P. A., Head, D. R., Schiffer, C. A., Bennett, J. M., Bloomfield, C. D. & Keating, M. J. Report of the National Cancer Institute-sponsored workshop on definitions of diagnosis and response in acute myeloid leukemia. *Journal of Clinical Oncology,* 1990; **8**(5), 813-819.
3. Durairaj, M., and R. Deepika. "Prediction Of Acute Myeloid Leukemia Cancer Using Datamining - A Survey." *International Journal of Emerging Technology and Innovative Engineering,* 2015; **1**(2): 94-98.
4. Cheson, B. D., Bennett, J. M., Grever, M., Kay, N., Keating, M. J., O'Brien, S., & Rai, K. R. National Cancer Institute-sponsored Working Group guidelines for chronic lymphocytic leukemia: revised guidelines for diagnosis and treatment. *Blood,* 1996; **87**(12), 4990-4997.
5. Tang, Wenlong, Hongbao Cao, and Yu-Ping Wang. "Subtyping of Leukemia with Gene Expression Analysis Using Compressive Sensing Method."Healthcare Informatics, Imaging and Systems Biology (HISB), 2011 First IEEE International Conference on. IEEE, 2011.
6. Eldosoky, Mohamed. "K5. Diagnosis of Blood Leukemia by using Ultrawide band Pulse." Radio Science Conference (NRSC), 2013 30th National. IEEE, 2013.
7. Van Dongen, J. J. M., et al. "Standardized RT-PCR analysis of fusion gene transcripts from chromosome aberrations in acute leukemia for detection of minimal residual disease Report of the BIOMED-I Concerted Action: Investigation of minimal residual disease in acute leukemia." *Leukemia,* 1999; **13**: 1901-1928.
8. Pullarkat, Vinod A., et al. "Mast cell disease associated with acute myeloid leukemia: detection of a new c-kit mutation Asp816His." *American journal of hematology,* 2000; **65**(4): 307-309.
9. Ts'o, Paul OP, et al. "Detection of intact prostate cancer cells in the blood of men with prostate cancer." *Urology,* 1997; **49**(6): 881-885.
10. Kawasaki, Ernest S., et al. "Diagnosis of chronic myeloid and acute lymphocytic leukemias by detection of leukemia-specific mRNA sequences amplified in vitro." *Proceedings of the National Academy of Sciences,* 1988; **85**(15): 5698-5702.
11. Scholefield, J. H., and S. M. Moss. "Faecal occult blood screening for colorectal cancer.(Editorial: Colorectal Cancer)." *Journal of medical screening* 2002; **9**(2): 54-56.
12. Reulen, Raoul C., et al. "Long-term cause-specific mortality among survivors of childhood cancer." *Jama,* 2010; **304**(2): 172-179.
13. Beillard, E., et al. "Evaluation of candidate control genes for diagnosis and residual disease detection in leukemic patients using 'real-time' quantitative reverse-transcriptase polymerase chain reaction (RQ-PCR)–a Europe against cancer program." *Leukemia,* 2003; **17**(12): 2474-2486.