

Integration of SVM with Artificial Fish Swarm Algorithm for Blood Tumour Prediction

Ms. N.R. Deepa, Dr. S. Gunasekaran, R. Elakiya, U. Haritha, N. Kaleeswari, P. Purnima

Abstract— The growth of abnormal tissue is also called neoplasm which can be differentiated from the surrounding tissues by its structure. This tumour will affect the immune system which is a major leading cause of death around 13% worldwide. Blood diseases such as leukemia replaces the normal blood cells in the bone marrow and the blood. Effective modern drugs can be deployed for the blood diseases such as Chronic Lymphatic Leukemia(CLL). Data mining technique is used to categorize the blood test uniqueness(Hematology) and blood swelling to predict the disease in a early stage. Due to the increase in blood tumour diseases Support Vector Machine(SVM) is proposed for the classification of tumour and hematological data. Fish Swarm algorithm is found more efficient in optimizing the data with high accuracy.

Index Terms— Artificial fish swarm algorithm, Blood Tumour, Data mining , Hematology, Support Vector Machine..

I. INTRODUCTION

Blood tumour occurs when abnormal blood cells growing out of control through the blood and lymph system which fights off infection and produce new cells. This disease affects bone marrow, blood and lymphatic system. As the bone marrow produces new blood cells, the damaged white cells causes inability to produce normal blood cells. Hematology diseases such as leukemia, thalassemia, lymphoma,etc. Leukemia cells are the major cause for affecting immune system. The medical aspects of Hematology is concerned with the treatment of blood disorders. One of the most common Hematology test is CBC(Complete Blood Count). CBC test check the figure of blood cells circulating in the blood stream. This test is used to detect and monitor the tumour treatment. As per 2018 survey, around 16.11% of humans suffer from blood tumour. Data mining technique is to extract the information from the datasets and transform it into an understandable data which categorize CBC sample of patient as Hematology or tumour. The prediction process of this

Ms. N.R. Deepa, Department of CSE, Coimbatore Institute of Engineering and Technology, Coimbatore, India

Dr. S. Gunasekaran, Department of CSE, Coimbatore Institute of Engineering and Technology, Coimbatore, India

R. Elakiya, Department of CSE, Coimbatore Institute of Engineering and Technology, Coimbatore, India

U. Haritha, Department of CSE, Coimbatore Institute of Engineering and Technology, Coimbatore, India

N. Kaleeswari, Department of CSE, Coimbatore Institute of Engineering and Technology, Coimbatore, India

P. Purnima, Department of CSE, Coimbatore Institute of Engineering and Technology, Coimbatore, India

diseases can reduce the number of mortalities and enhance the quality of life for the infected patients. The treatment depends on the features of leukemia cells including patient's age and health. The discriminative classifier SVM is applied for predicting the Blood tumour with the identical features of Hematology. Fish Swarm algorithm is applied to optimize the probability of Hematology for the estimation of fitness value of blood cell count with high accuracy.

II. SURVEY

Data mining techniques is used to get new knowledge from the large number of data in the medical field. Some of the researchers done in the Hematological disease such as Abdullah and Al-Asmari to specify the anemia type from the patients through a predictive model. Datasets are constructed from the Complete Blood Count test results of the patients. The implemented classification algorithm such as Naïve Bayes, Multiplier Perception, J48 and SMO. J48 decision tree and SMO produce the accuracy of 93.75%. They proposed to use Artificial Neural Network, SVM and Decision tree. But no actual experiment is done. Minnie and Srinivasan used Association Rules and the clusters on the collected data. Saichanma et al used Data mining technique to predict abnormality in peripheral blood smear by using dataset of Hematological parameters gathered from Blood cell counter. The Decision tree created by the algorithm can be used as practical guideline for RBC morphological prediction by using four Hematological parameters (MCV, MCH, HCT and RBC). Finally Vijayarani and Sudha, developed Weight-based K-means algorithm for identifying Leukemia, inflammatory, bacteria or viral infection and anemia diseases from the Hemogram blood test samples from the dataset. They found that accuracy of weight-based K-means algorithm is better when compared to K-means and Fuzzy c-means algorithm.

Shuting Shen, Ziqiang Fan and Qiguo establish tumour prediction model by using logistic regression model is widely used to explore predictive morbidity of disease. Artificial Neural Networks have also been applied to the prediction of malignant tumour; they produced good results. BP neural network is a multilayer feed forward network in ANN which is trained by error back propagation algorithm. BP neural network showed a high accuracy when compared to logistic regression because the neural network cannot explain the variables, it is best to combine them with each other during practical application. The grey GM (1,1) model was widely used in system of uncertainty with the implementation of forecast method. The grey GM(1,1) has no exact condition

for the probability distribution of sample contents and was used to establish tumour prediction model.

III. EXISTING SYSTEM

The primary goal of this research work is to analyze the Blood test samples to predict the Blood tumour. Clustering algorithms namely K-means and Fuzzy C-means are used for the analysis. In addition to this analysis, a new weight-based K-means clustering algorithm is also proposed. Data mining K-means clustering algorithm is enhanced by assigning weights to the data items. The performance factor is used for finding an effective algorithm are clustering accuracy, execution time and error rate. Five diseases namely Leukemia, Inflammatory disease, Bacterial or viral infection, HIV infection and anemia are predicted from this Blood test sample analysis.

Classification is done by Association rule mining for discovering a small set of rule that forms an accurate classifier. Rule induction is used to extract the set of rules that shows the relationships between the attributes of the datasets and a class label. Deep learning is a type of neural network used to model abstraction in data through model architectures.

The classification technique such as J48 and Random tree by means of WEKA to classify Sickle Cell Disease. J48 algorithm provides high performance in classifying type of anemia. The experiment has been conducted for CBC to extract data using Intended algorithm.

IV. PROPOSED SYSTEM:

A Hemogram/Complete Blood Count is the common laboratory test. As the name suggests, a Hemogram/CBC analyses various blood components like RBC, WBC and platelets. These components are non-specific but very important indicators of general well-being of an individuals. Many of the parameters included in Hemogram do not pinpoint towards the specific diagnosis although careful analysis of the combination of several parameters included in a Hemogram may help the clinicians/Hematologists to arrive at important conclusions. Sample required is EDTA whole blood and the entire range of test can be conducted with just 2/3 ml of blood.

Existing system handles only structured data. The prediction system are broad and ambiguous. The standard SVM takes set of input data and predicts, for each given input data, which of two possible classes forms the input making the SVM and non-probabilistic binary linear classifier. SVM is a powerful method for building a classifier. It aims to create a decision boundary between two classes that enables the prediction of labels from one or more feature vectors. This decision boundary known as hyperplane is orientated in such a way that it is as far as possible from the closest data points from each of the classes. These closest points are known as Support Vectors .Artificial Fish Swarm algorithm is used for the optimization of the data that were used as the test data for blood prediction. This Swarm intelligence optimization algorithm works on Stochastic search. There were many optimization algorithms which have affinity with this method and its result will improve the

performance of this algorithm.

Table 1. Attributes Used

NO.	SYMBOL	MEANING
1	WBC	White Blood Cells
2	RBC	Red Blood Cells
3	HGB	Hemoglobin
4	HTC	Hematocrit
5	MID	Mid-range absolute count
6	MCH	Mean Cellular Hemoglobin
7	MCHC	Mean Cellular Hemoglobin Concentration
8	RDW	RBC Distribution Width
9	PLT	Platelets Count
10	MPV	Platelets Volume
11	GRAD	Percentage of WBC with granules in their cytoplasm
12	LYM	Lymphocyte percent
13	Gender	Male, Female
14	Class	Tumour, Hematology

V. IMPLEMENTATION:

The system has implemented User login. The role of user login is to import a dataset which is collected from CBC test sample for training, testing, optimizing and classifying the dataset to predict the Blood tumour.

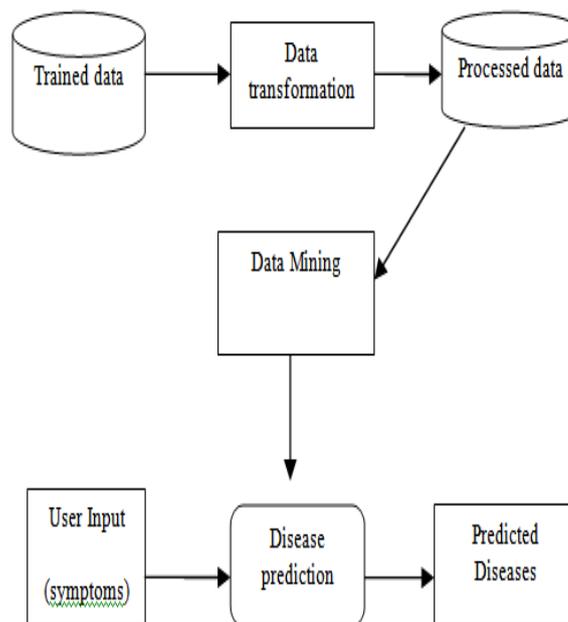


Fig 1: Block Diagram for Implementation

A. TRAINING DATA

Trained datasets are basically fit to the value of each parameters. Based on the range of each blood cells count, the patient’s data have been trained.

B. TESTING DATA

Most of the trained datasets are used for testing which are smaller in size when compared to trained data. CBC test is used to analysis the data to ensure that the data are similar. The manipulation of training and testing datasets minimizes the effect of data discrepancies for making predictions easier.

C. FEATURE EXTRACTION

Feature extraction builds derived values, that is, the raw variables have been reduced to informative features which deploys the original datasets accurately. The reduced features are likely to contain relevant information from the input dataset so that the inclined dataset can be performed instead of the complete dataset.

D. OPTIMIZATION

Optimization obtains the best possible solutions among the various datasets. This technique attempts to optimize the required features for the desired solutions.

Table 2. Rules to predict tumour from CBC Samples

ATTRIBUTES	RANGE
LYM	<1.85
RBC	<3.64
RDW	>14.700
GRAN	>5.950

Table 3. Rules to predict Hematology from CBC Samples

ATTRIBUTES	RANGE
GRAN	<5.100
LYM	>2.250
PLT	>347
GRAN	<9.900

In this journal, we use Artificial Fish Swarm Algorithm for the optimization of data in terms of the values given in the above table.

E. CLASSIFICATION

Classification is the process that segregates the outliers from the normal data. It is done by comparing the testing data with the trained data.

1. Condition for tumour

- If $LYM \leq 1.850$ and $RBC \leq 3.635$ then tumour.
- If $RDW > 14.700$ and $GRAN > 5.950$ then tumour.
- If $GRAN > 2.800$ and $LYM \leq 1.850$ and $RDW > 14.800$ then tumour.
- If $RBC \leq 4.595$ and $HTC > 33.450$ then tumour.
- If $MCH > 25.750$ and $GRAN > 4.600$ and $HTC > 36.250$ then tumour.

2. Condition for hematology

- If $GRAN \leq 5.100$ and $LYM > 2.250$ then hematology.
- If $PLT > 374$ and $GRAN \leq 9.900$ the hematology.
- If $RDW \leq 15.200$ and $RBC \leq 4.415$ then hematology.

SVM is a concept in Statistics and Computer Science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input, making the SVM a non-probabilistic binary liner classifier. SVM is a powerful method for building a

classifier it aims to create a decision boundary between two classes that enables the prediction of labels from one or more feature vectors. This decision boundary, known as hyperplane, is orientated in such a way that it is as far as possible from the closest data point from each of the classes. These closest points are called support vectors.

Here, We have proposed SVM to classify tumour and hematological diseases from CBC test samples based on the above mentioned conditions.

VI. CONCLUSION

We have proposed a model utilizing Kernel Support Vector Machine which is executed in a high arrangement PC. We have framed a disarray framework from the classifier. We have used 78% occurrences of approximately information for preparing both in Support Vector Machine and independently. The rest of the 10% cases utilized for testing both in SVM separately. The graphical portrayal of the disarray grid for every methodology. It portrays that the effectively distinguished esteem is relatively higher in SVM for preparing and testing. The execution estimates records are determined both for preparing and testing utilizing the above portrayed condition.

REFERENCES

- [1] M.Durairaj and V. Ranjani, "Data Mining application in healthcare: A study", Int.J.Sci.Technol.Res., vol.2, no. 10, pp.29-35,2013.
- [2] H.C. Koh and G. Tan , "Data Mining application in healthcare" J.Healthc.Inf.Manag., vol.19, no. 2,p.65,2011.
- [3] H. Hemonctoday. What is hematology. Available: <http://www.healio.com/hematology-oncology/news/online/{2dd178d0-7f92-46a8-add9-2c7d634d2cea}/what-is-hematology.2016>.
- [4] R. Shouval , O. Bondi, H. Mishan, a shimoni, R. Unger, and A. Nagler, "Application of Machine Learning Algorithms for Clinical predictive modeling:A data mining approach in SCT.," Bone marrow Transplant., vol.49, no.3, pp.332-7,2014.
- [5] Mayo Clinic . "Cancer blood test: lap tests used in cancer diagnosis" <http://www.mayoclinic.org/diseases-conditions/cancer/in-depth/cancer-diagnosis/art-20046459.2017>.
- [6] M. Rashid, M. Hoque, and A. Satter, "Association rule mining based on clinical observation" arXivprepr. arXiv1401.2571,2014.
- [7] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu perez, B. Lo, and G.-z.yang,"Deep learning for health informatics," IEEE.j.Biomed.Heal.Informatics,vol.21, no.1,pp.1-1,2016.
- [8] M. Abdullah and S. Al-Asmari "Anemia type prediction based on data mining classification algorithms," Communication, Management and Information Technology-Sampaio de alencar(Ed.)2017.
- [9] S. Vijayarani and S. Sudha."An Efficient Clustering Algorithm for Prediction Disease from Hemogram Blood Test Sample". Indian Journal of Science and Technology, vol.8(17),Aug 2017.
- [10] Alaa M. El-Halees, AsemH.Shurab , "Blood Tumour Prediction using Data Mining Techniques" Health informatics- An international Journal vol.6, no: 2, May 2017.
- [11] CharmiMehta "Basics of Data Mining: A Survey Paper"International Journal Trends in Research and Development, vol.4(2),ISSN: 2394-9333.
- [12] Jiawei Han , Micheline Kamber and Jian Pei,"Data Mining Concepts and Technique".
- [13] R. Alagappan "Manual of Practical Medicine".
- [14] K. Sembulingam "Essential of Medical Physiology"
- [15] J. Abinaya and A. Sumathi "Predicting Blood Tumour and Hematology using Data Mining Techniques" International journal of Computer Science Trends and Technology, vol.6 issues 3,May-June 2018.
- [16] Valder Steffen "Fish Swarm Optimization Algorithm applied to Engineering System Design" <http://www.researchgate.net/publication/26250985>