Cancer Detection Technique in Data Mining

Shanmitha S, Rathina Moorthy N, Suruthi S

Abstract *- Cancer is a group of diseases begins in cells that are the basic building blocks of a body. There are different types of cancers but all starts with the cells growing out of control. Segmentation is an essential step in image systems for the accurate lung disease diagnosis, it measure lung structures in Computerized Tomography (CT) images. Indeed, image processing techniques can help computer diagnosis if lung region is accurately obtained. A conventional fuzzy C-means clustering algorithm that has been implemented for segmentation of lung images still suffers with low convergence rate, getting stuck in the local minima and vulnerable to initialization sensitivity. The proposed system presents an intelligent and dynamic approach called Intelligent Fuzzy C-Means (IFCM) to segment the lung nodules automatically and classify the lung nodules effectively using support vector machine classifier. It leads to the capability of firefly search to find optimal initial cluster centers for the (FCM) and thus improve the segmentation accuracy. The features are extracted using fused tamura and haralick features after segmentation.*

The features can be trained by using trained using different kernels of support

vector machine for automatic detection of lung nodules as benign or malignant.

Index:

I. INTRODUCTION

Data mining software analyses relationship and patterns are stored transaction data based on open ended user queries. Several types of analytical software are available and the most commonly used techniques are: To extract, transform and load transaction data into the data warehouse system. And then Store and manage the data in multi-dimensional database system for provided data access to business analysts and information technology professionals. By analyses the data of application software to present the data in a useful format such as a graph or table.

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

1. The centroids of the *K* clusters, which can be .
2. used to label new data

2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.

Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

2. Related work:

Fuzzy c-means algorithm is popular algorithm for robust lung image segmentation. The fuzzy c-means algorithm attempts to partition a finite collection of n elements into a collection of c fuzzy clusters with respect to some given criterion. Many metaheuristic search algorithms have been hybridized with the fuzzy c-means algorithm to find optimal cluster centres. These algorithms explore the entire search space in the problem to determine possible solutions. These algorithms include bee optimization, harmony search, the ant colony algorithm, simulated annealing, the genetic algorithm, tabu search, the firefly algorithm and particle swarm. In this research work the data is in the form of Computer Tomography (CT) images of the lung and hence image analysis is done to perform diagnosis. The segmentation of such image is done by using an intelligent and dynamic approach called Intelligent Fuzzy C Means with firefly search (IFCM) and classified effectively using support vector machine.

Dataset Acquisition

In this module, upload the datasets. The dataset may be microarray dataset. Gather the data from hospitals, data centers and cancer research centers. The collected data is pre-processed and stored in the knowledge base to build the model.

Preprocessing

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine projects. Data-gathering methods are often insecurely controlled, resulting in out-of-range values, impossible data combination, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce ambiguous results.

Feature Selection

In this module is used to select the features of the given dataset. Attribute selection was performed to determine the subset of features that were highly correlated with the class while having low inter correlation.

Disease Diagnosis

Classification of reduced set using neural network. Based on the values acquired from training phase, the performance of the NN network is analyzed to obtain appropriate values for testing phase. In order to find the optimum structure, the NN network performance has been analyzed for the optimum number of hidden nodes and epochs. For this situation, the epochs will be set to a definite preset value. Then, the NN network was trained at the appropriate range of hidden nodes. The number of hidden nodes that have given the best performance is then selected as the optimum hidden nodes. After that, by fixing the optimum number of hidden nodes, the epochs will be analyzed in a similar way to obtain the optimum number of epochs that can give the highest or best accuracy

.

Evaluation Criteria

In this module, the performance of the proposed Genetic algorithm is extensively compared with that of some existing supervised and unsupervised gene clustering and gene selection algorithms. To analyze the performance of different algorithms, the experimentation is done on Cancer data sets. The major metrics for evaluating the performance of different algorithms are the class separability index and classification accuracy of Neural Network rule. The proposed system provide improved accuracy rate in gene classification.

Experimental Study

Data set

Leukemia is a type of cancer of the blood or bone marrow categorize by an irregular augment of undeveloped white blood cells called "blasts." It is a thick term covering a compilation of diseases. According to American Cancer Society it is approximated that 48,610 persons (27,880 men and 20,730 women) will be detect with and 23,720 men and women will terminate of leukemia in 2013 only. In turn, it is part of the even broader set of diseases disturbing the blood, bone marrow, and lymphoid system, which are all known as hematological neoplasm. Over time, leukemia cells can crowd out the normal blood cells. This can lead to serious problems such as anemia, bleeding, and infections. Leukemia cells can also spread to the lymph nodes or other organs and cause swelling or pain. There are several different types of leukemia.

- Acute lymphoblastic leukemia, or ALL.
- Acute myelogenous leukemia, or AML.
- Chronic lymphocytic leukemia, or CLL.
- Chronic myelogenous leukemia, or CML.

In general, leukemia is grouped by how fast it gets worse and what kind of white blood cell it affects. Acute Lymphoblastic Leukemia (ALL) is the most all-purpose type of leukemia in young children and Acute Myelogenous Leukemia (AML) occurs more usually in adults than in children, and more usually in men than women [12]. The young WBC can also build up in a variety of extreme dullard sites, especially the mining's, gonads, thymus, liver, spleen, and lymph nodes. Hence due to extreme lye-phobic blast or myeloid blast in the marrow they also low into the peripheral blood stream. Acute myeloid leukemia (AML) is also recognized by other names, which include acute myelocytic leukemia, acute Myelogenous leukemia, acute granulocytic leukemia, and acute non-lymphocytic leukemia. “Acute” means that this leukemia can develop rapidly if not treated, and would approximately certainly be lethal in a few months. “Myeloid” refers to the type of cell from where this leukemia begins. In most cases AML build up from cells that would wind into white blood cells (other than lymphocytes), but in some cases of AML expand in other types of blood forming cells.

IV. CONCLUSION

Cancer is potentially fatal disease. Even now the actual reason and complete cure of cancer is not invented. Detection of cancer in earlier stage is curable. In this work we have developed a system called data mining based cancer prediction system. The main aim of this model is to provide the earlier warning to the users and it is also cost and time saving benefit to the user. It predicts three specific cancer risks. Specifically, Cancer prediction system estimates the risk of the breast, skin, and lung cancers by examining a number of users provided genetic and non-genetic factors. It is validated by comparing its predicted results with the patient's prior medical record .This prediction system is available in online, people can easily check their risk and take appropriate action based on their risk status. The performance of the system is better than the existing system.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my guide Dr.Sakthivel for the continous support of our UG project and related research, for his patience,motivation,and immense knowledge.His guidance helped us in all time on doing project.I could not have imagined having a better mentor for our UG project

REFERENCES

[1] K Sakthivel, C Kavitha and A Jayanthiladevi, (2016)" Biomedical Research.

[2] Sruthi I and Robin J,"Computer aided lung cancer detectionsystem." Proc Glob Conf Commun Technol 2015,pp. 555-558.

[3] Hu S, Hoffman EA and Reinhardt JM," Automatic lungsegmentation for accurate quantitation of volumetric X-rayCT images", IEEE Trans Med Imaging 2001,pp. 20: 490-498.

[4] Sluimer I, Prokop M and van Ginneken B," Toward automatedsegmentation of the pathological lung in CT", IEEE TransMed Imaging 2005, 24,pp, 1025-1038.

[5] Punithavathy KR and Sumathi PMM," Analysis of statisticaltexture features for automatic lung cancer detection inPET/CT images",Robo Autom Contr Embed Sys 2015, pp. 1-5.

[6] Chonglun L. A new automatic seeded region growingalgorithm," Proceedings of 6th International Congress onImage and Signal Processing (CISP)", 2013,pp, 543-549.

[7] Gomathi M and Thangaraj P," A computer aided diagnosissystem for lung cancer detection using machine learningtechnique", Europ J Sci Res 2011,pp, 260-275.

[8] Jiangdian S and Caiyun Y, et.al,"Lung lesion extraction using a toboggan based growingautomatic segmentation approach." IEEE Trans MedImaging 2016,pp, 337-353.

[9] Venkatasalam K and Rajendran P,"Effective RBIR Fuzzy CMeanssegmentation HAAR wavelet with user interactivemulti threshold robust features vector", Asian J InformTechnol 2016, pp , 223-231.

[10] Gong M and Liang Y, et al,"Fuzzy C-meansclustering with local information and kernel metric forimage segmentation", IEEE Trans Image Process 2013,pp ,573-584.