CLASSIFICATION TECHNIQUES OF LEUKEMIA USING IMAGE PROCESSING: AN OVERVIEW

A. P.AISWARIYA
Assistant Professor, Department of Computer Science,
Rajeswari College of Arts and Science for women,
Bommayapalayam.

ABSTRACT

Image classification is an important tool for extracting information from digital images .Leukemia is the result of the fast overproduction of abnormal white blood cells. Iteffects on the brink of 45,000 people within us annually. Leukemia is that the commonest sort of blood cancer in people older than 55. But, it's also the foremost common cancer in children under 15.Leukemia occurs when abnormal white blood cells within the bone marrow quickly increase and destroy normal blood cells. This leaves a person disposed to infection. Treatment approaches for leukemia depend upon the sort of white blood corpuscle affected and whether the disease is acute or chronic. Acute leukemia forms quickly and becomes severe rapidly. Chronic leukemia grows more gradually and takes longer to advance.The aim of this paper is to summarize information about few image classification techniques. The paper also elaborates different categories of image classification techniques. The image classification techniques considered in this paper are Support Vector Machines (SVM), K-nearest neighbors (K-NN), self-organizing map (SOM), and Naïve Bayes Classifier techniques.

Key Words: Leukemia, Support Vector Machines (SVM), K-nearest neighbors (K-NN),self-organizing map **(**SOM**),**Naïve Bayes Classifier techniques.

I.INTRODUCTION

Leukemia is a cancer of the blood cells. Blood is made of three types of cells: 1) White blood cells that fight infection.2) Red blood cells that carry oxygen from the lungs to the rest of the body.3) Platelets that help the blood clot and stop bleeding. These cells are made in the bone marrow, a soft area in the centre of the bones. Once these cells developed, they leave the marrow and start circulating through the bloodstream.With leukemia, the bone marrow produces an accumulation of one kind of cell usually white blood cells that no longer

function normally. Instead these abnormal cells gathering the bone marrow so it can no longer produce enough healthy cells. Leukemias are classified by how quickly they progress and by the type of cell involved. In acute leukemia, immature white blood cells, known as blasts, grow too rapidly. This can cause severe symptoms right away .In chronic leukemia, abnormal cells grow more slowly and symptoms may not appear for years. Leukemia affects two types of cells.1)Myeloid leukemias develop in white blood cells known as neutrophils and monocytes that help fight bacterial or fungal infections.2)Lymphocytic leukemias develop in B lymphocytes, T lymphocytes, or natural killer (NK) cells that produce antibodies or other substances to fight infections. There are four main types of leukemia. 1) Acute myeloid leukemia (AML).2)Acute lymphoblastic leukemia (ALL). 3) Chronic myeloid leukemia (CML). 4)Chronic lymphocytic leukemia (CLL).

TABLE I: FOUR MAIN TYPES OF LEUKEMIA

LEUKEMIA	LYMPHOCYTIC LEUKEMIA	MYELOGENOUS LEUKEMIA
Acute	Acute lymphoblastic leukemia (ALL)	Acute myeloid leukemia (AML)
Chronic	Chronic lymphocytic leukemia (CLL)	Chronic myeloid leukemia (CML)

II LITERATURE REVIEW

Image classification is an important step in the object detection and image analysis. The final output or the intermediate output can be the output of the image classification step. A lot of image classification techniques have been planned till date. Various studies have been conducted in order to conclude about the best satellite image classification technique. Because the results and its accuracy depend on a number of factors it is very hard to decide any one technique as the best technique [2].

Over the last few decades, there is a constant modification in the conventional methods as well as invention of new image classification techniques in order to get maximum accurate results. Each of the classification technique has its own advantages and disadvantages. The research now concentrates on combining the desired features of these techniques in order to increase the efficiency. As the hard classifiers cannot handle the problem of mixed pixels, the soft classifiers are used. But, soft classifiers have their own disadvantages.

Another study was accompanied on PoISAR data shared the Markov random field smoothness restriction with supervised Softmax regression model [6]. The experiments conducted during this study proved that the combination of supervised and unsupervised algorithms produced better results as compared to the ones produced by either of these techniques separately. [10] Presents the combination of fuzzy logic and neural networks in order to design a system that can detect the face and fingerprints of the person. This is done in order to determine the authenticity of the person. This system can be used for various security purposes.

III CLASSIFCATION TECHNIQUES OF LEUKEMIA

A classifier is a machine learning model that is used to distinguish different objects based on certain features.

1. NAIVE BAYES CLASSIFIER

A Naive Bayes classifier is a machine learning model that's used for classification task. The root of the classifier is based on the Bayes theorem.

Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

We can find the probability of **A** happening, given that **B** has occurred, using Bayes Theorem. Here, **B** is the proof and **A** is the hypothesis. The hypothesis made here is that the analysts/features are independent. That is occurrence of one particular feature does not affect the other. Hence it is called naive.

Bayes theorem can be rewritten as:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Variable **X** represents the parameters/features .The variable **y** is the class variable(play golf), which represents if it is suitable to play golf or not given the conditions. **X** is given as,

$$X = (x_1, x_2, x_3, ..., x_n)$$

Here x_1,x_2….x_n represent the features, i.e they can be mapped to position, temperature, humidity and windy. By exchanging for **X** and expanding using the chain rule we get,

$$P(y|x_1,...,x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

Now, you can obtain the values for each by considering at the dataset and substitute them into the equation. For all entries in the dataset, the denominator does not change, it continue static. Therefore, the denominator can be removed and proportionality can be introduced.

$$P(y|x_1,...,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

In our case, the class variable(**y**) has only two consequences, yes or no. There could be cases where the classification could be multivariate. Therefore, we must to find the class **y** with maximum probability.

$$y = argmax_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

Using the above function, we can get the class, given the predictors.

Types of Naive Bayes Classifier:

Multinomial Naive Bayes:

This is for the most part utilized for archive order issue, i.e. whether a report has a place with the classification of sports, governmental issues, innovation and so on. The highlights/indicators utilized by the classifier are the recurrence of the words present in the record.

Bernoulli Naive Bayes:

This is similar to the multinomial naive bayes yet the indicators are Boolean factors. The parameters that we use to foresee the class variable take up just qualities yes or no, for instance if a word happens in the content or not.

Gaussian Naive Bayes:

At the point when the indicators take up a constant worth and are not particular, we accept that these qualities are tested from a gaussian conveyance. Since the manner in which the qualities are available in the dataset changes, the recipe for contingent likelihood changes to,

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

ADVANTAGE:

Naive Bayes algorithms are mostly used in sentimentality analysis, spam filtering and recommendation systems etc. They are easy to implement and fast.

DISADVANTAGE:

The biggest disadvantage is that the requirement of predictors to be self-governing. In most of the real life cases, the predictors are dependent; this checks the performance of the classifier.

2. SUPPORT VECTOR MACHINES (SVM):

The support vector machine (SVM) is a classification algorithm that makes available state of the skill performance in a wide variety of application domains, including handwriting recognition, object recognition, speaker identification, face detection and text categorization. Two main inspirations suggest the use of SVMs in computational biology. In clinical bioinformatics they have allowed manufacture of powerful experimental cancer diagnostic models based on gene expression data with thousands of variables and as little as few dozen samples. Furthermore, several efficient and high-quality implementations of SVM algorithms facilitate application of these techniques in exercise. The first group of SVMs could only be applied to binary classification tasks. The preliminary experimental evidence currently available suggests that some multi-category SVMs (MCSVMs) perform well in remote gene expression-based cancer diagnostic experiments. In the explanation of methods below, k is the number of classes or distinct analytic categories, and n is the number of samples or patients in the training dataset.

ADVANTAGES:

- It has excellent simplification capacity.
- It does not face the problem of over fitting.

DISADVANTAGES:

- The structure of the algorithm is difficult and therefore challenging to understand.
- Optimal parameters cannot be distinct easily.

3. ARTIFICIAL NEURAL NETWORK (ANN)

Artificial neural networks are non-parametric classifiers. The structure of the artificial neural networks is inspired from the human nervous system. The basic unit of this type of network is unified processing rudiment known as neuron. Each neuron has two steps- training and using phase. In the training phase, the neuron learn to perform an operation while in the testing phase, they use the training information to predict the output. Generally, these neural networks are used in order to detect specific trends or patterns in the given data.

ADVANTAGES:

• It has high computation rate.

• It deals with the noisy inputs powerfully.

• This technique is data driven as it learns from the training data.

DISADVANTAGES:

• As it requires prior training, it is time-consuming.

• It is considered to be semantically poor.

• It encounters the problem of overfitting.

4.SELF-ORGANIZING MAP (SOM)

A self-organizing map **(**SOM**)** is a type of artificial neural network (ANN) that is skilled using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized demonstration of the input space of the training examples, called a map**,** and is therefore a method to do dimensionality reduction. Self-organizing maps differ from other artificial neural networks as they apply inexpensive learning as opposed to error-correction learning (such as backpropagation with gradient descent), and in the sense that they use a district function to preserve the topological properties of the input space. It is sometimes called a Kohonen map.

Each data point in the data set distinguishesthem by competing for representation. SOM mapping steps starts from adjusting the weight vectors. From there a sample vector is selected accidentally and the map of weight vectors is searched to find which weight best represents that sample. Each weight vector has nearest weights that are close to it. The weight that is chosen is satisfied by being able to become more like that randomly selected sample vector. The neighbors of that weight are also satisfied by being able to become more like the chosen sample vector. This allows the map to produce and form different shapes. Most commonly, they form square/rectangular/hexagonal/L shapes in 2D feature space.

ADVANTAGES:

1. Data mapping is easily understood.
2. Capable of organizing large, difficult data sets.

DISADVANTAGES:

1. Difficult to define what input weights to use
2. Mapping can result in splitscollections.
3.Requires that nearby points be have similarly.

5. K-NEAREST NEIGHBOURHOOD

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well,

- **Lazy learning algorithm** − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- **Non-parametric learning algorithm** − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

ADVANTAGES:

1. Non-parametric form of classification based on distance.
2. It gives accuracy about the distance and can be used to train a large number of training set.
3. Simplest among all machine learning algorithm.

DISADVANTAGES:

1. To choose the value of k is very difficult.
2. The choice of k affects the efficiency of the algorithm.
3. Memory sensitive.
4. The performance of K-NN severely degraded in the presence of noise or if the feature scale are not Consistent.

IV CONCLUSIONS

In the paper, categorize cancer with the Leukemia cancer of medical diagnostic data. Information gain has been modified for feature selections. A Leukemia cancer model that utilizes some of algorithms such as Support Vector Machines (SVM), K-nearest neighbors (K-NN),self-organizing map **(**SOM**),**Naïve Bayes Classifier techniques and enhancements to evaluate, interpret the cancer classification. The experimental results indicate this model illustrates the highest accuracy of classifications for Leukemia cancer. In the future work, would like to use some algorithm tuning in terms of weight initialization, activation function to improve performance and accuracy.

V REFERENCES

[1] https://www.wcrf.org/int/cancer-facts- figures / world widedata.

[2] Kalra K., Goswami A.K., Gupta R., "A Comparative Study of Supervised Image Classification Algorithms for Satellite Images", International Journal of Electrical, Electronics and Data Communication, ISSN: 2320-2084, Volume-1, Issue-10, Dec-2013.

[3]Niranjan Chatap, Sini Shibu, "Analysis of blood samples for countingleukemia cells using Support vector machine and nearest neighbour",IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. III (Sep – Oct.2014), PP 79-87

[4] Viswanathan PA, "Fuzzy C means Detection of Leukemia based on Morphological Contour Segmentation", Second International Symposium on Computer Vision and the Internet (VisionNet'15), Procedia Computer Science 58 (2015) 84 – 90

[5] Tomasz Markiewicz, Stanislaw Osowski, "Automatic Recognition of the Blood Cells of Myelogenous Leukemia Using SVM", Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August4, 2005

[6] Haixia Bi, Jian Sun, Zongben Xu, "Unsupervised PolSAR Image Classification Using Discriminative Clustering", IEEE Transactions on Geoscience and Remote Sensing, Volume 55 Issue 6.

[7] Jokkrich Laosai, kosin Chamnongthai"Acute Leukemia Classification by Using SVM and KMeans Clustering," Proceedings of the International Electrical Engineering Congress 2014.

[8] Lorenzo putzu,Giovannicaocci, "Leucocyte classification for Leukemia detection using image processing techniques",Elsevier volume 62,issue 3,Nov 2014, Pages 179-191.

[9] G. PrabhakarReddya, Y. Deepika, K. Sai Prasad, Dr. G. Kiran Kumar, "Fuzzy Logics associated with Neural Networks in the Real Time for Better World", International Conference on Advancements in Aeromechanical Materials for Manufacturing (ICAAMM-2016).