ISSN 2321-807X
Volume 13 Number1
Journal of Advances in chemistry

Detection and Classification of Leukemia using MPFCM Segmentation and Random Forest with Boosting Techniques

T.C.Kalaiselvi
Assistant Professor, Department of ECE , Kongu Engineering College,Perundurai,
Tamil Nadu, India .kalaiselvi@kongu.ac.in

ABSTRACT:

Identification of blood disorders is through visual inspection of microscopic blood cell images. From the identification of blood disorders lead to classification of certain diseases related to blood. We propose an automatic segmentation method for segmenting White blood cell images. Firstly, modified possibilistic fuzzy c-means algorithm is proposed to detect the contours in the image. The GLCM features are extracted and features are selected by MRMR. Adaptive boosting and LS Boosting has been utilized to classify blast cells from normal lymphocyte cells. Comparison performance of classification accuracy was carried out. The effectiveness of the classification system is tested with the total of 80 samples collected. The evaluated results demonstrate that our method outperformed the existing systems with an accuracy of 88 %.

Keywords—MPFCM segmentation, ADA Boosting LS Boosting, GLCM features, MRMR Feature selection

1. Introduction

The assessment of the white blood cells in the bone marrow of patients is very informative in clinical practice. Segmentation is the process of partitioning a digital image into multiple segments based on pixels[7]. Segmentation is a critical and essential component of image analysis system[8]. The result of image segmentation is a collection of segments which combine to form the entire image. Various techniques have been proposed for segmenting an image in a better way. From the segmentation result, geometrical features such as area, perimeter etc [1] were detected for the final detection of immature cells. Three different classification techniques such as Tree Bagger, LS Boosting and ADA boosting were employed for classification[4] [5] [6], in order to classify the lymphocyte (WBC) as healthy and leukemic.

2. Proposed Methodology

2.1 MPFCM Clustering Segmentation

Image segmentation is the process of partitioning a digital image into multiple segments. Image segmentation is typically used to locate objects and boundaries in images [1].MPFCM is a good clustering algorithm [2] [3] to perform classification tests because it possesses capabilities to give more importance to typicalities or membership values. In order to avoid the constraint corresponding to the sum of all typicality values of all data to a cluster must be equal to one cause problems particularly for a big data set. It produces memberships and possibilities simultaneously, with the usual point prototypes or cluster centers for each cluster. The objective function is defined by

$$J_{pfcm} = \sum_{i=c}^{c} \sum_{k-1}^{n} (a\mu_{ik}^{m} + bt_{ik}^{\eta}) \times ||z_k - v_i||^2 + \sum_{i=1}^{c} \gamma i \sum_{k=1}^{n} (1 - t_{ik})^{\eta} \quad (1)$$

Subject to the constraints $\sum_{i=1}^{c} \mu_{ik}$=1 for all k and

0≤ μ_{ik},≤1. Here a>0, b>0, m >0, and η>0. U is the partition matrix. T is the typicality matrix. V is a vector of cluster centers, X is a set of all data points, z represents a data point, n is the number of data points and c is the number of cluster centers which are described by s coordinates.

2.1.1Algorithm

Step 1: Initialize prototype

$$vi = \sum_{k=1}^{N} (a\mu_{ik}^{m} + bt_{ik}^{\eta}) Z_k / \sum_{k=1}^{N} (a\mu_{ik}^{m} + bt_{ik}^{\eta})$$

$1 \le i \le C$ (2)

Step 2: For each cluster compute penalty parameter γ_i

$$\gamma_{i=} K \sum_{k=1}^{n} \mu_{ik}^{m} ||Z - v_i||^2 \quad (3)$$

Step 3: For each prototype calculate the distance

$$D_{ikA}^2 = (z_k - v_i)^T A_i (z_k - v_i), \quad 1 \le i \le C,$$

$1 \le k \le N$ (4)

Step 4: Calculate membership and typicality values

$$\mu_{ik} = \left(\sum_{j-1}^{c} \left(\frac{D_{ikA_i}}{D_{jkA_i}} \right)^{2/(m-1)} \right)^{-1}, 1 \le i \le C,$$

$1 \le k \le N$ (5)

$$t_{ik} = \frac{1}{1+\left(\frac{b}{r_i} D^2_{ik\,A_i}\right)^{1/(q-1)}} \quad , 1 \le i \le C, 1 \le k \le N \tag{6}$$

Step 5: Update the values of prototypes.

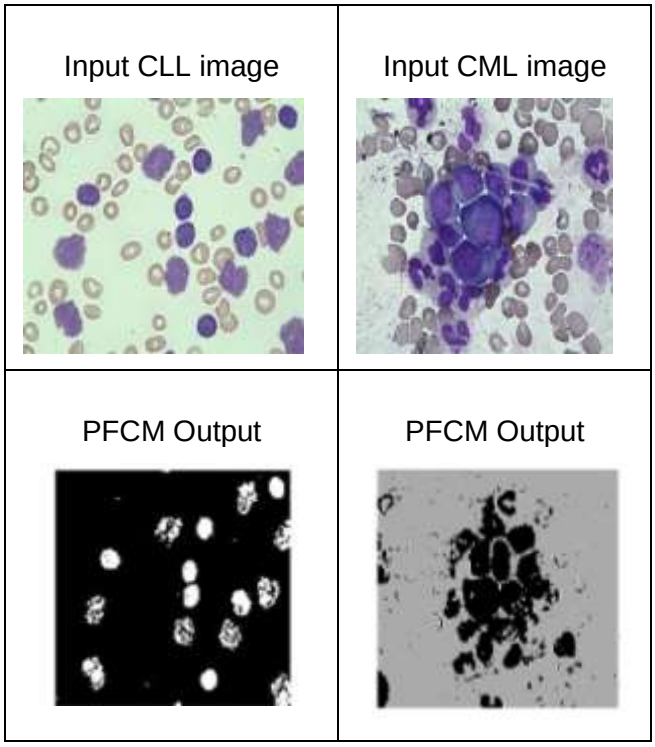

Fig.1 Input and Segmented Output of CLL & CML

Table.1 Evaluation Parameters of PFCM

EVALUATION PARAMETERS	PFCM
Rand Index	0.9922
Jaccard Index	0.9866
Dice Coefficient	2.0000

2.2 Feature Extraction

Feature extraction is a special form of dimensionality reduction [1]. When the input data to an algorithm is too large to be processed, then the input data will be transformed into a reduced representation set of features. Textural features based on the gray level co-occurence matrix (GLCM) are extracted from each image that are used to distinguish between normal and abnormal cancer cells. Co-occurrence matrices are calculated for four directions: 0º, 45º, 90º and 135º degrees.

GLCM has following features: Autocorrelation, Contrast, Correlation, Cluster Prominence, Cluster Shade, Dissimilarity, Energy, Entropy, Homogeneity, Maximum probability , Sum of squares, Sum average, Sum variance, Sum entropy, Difference variance, Difference entropy, Information measure of correlation, information measure of correlation, Inverse difference normalized are listed in table.2

Table.2 Feature Extraction Result

Features	Normal	CLL	CML
Contrast	1.1912	2.8458	1.6078
correlation	0.9994	0.9986	0.9987
cluster prominence	1610	2132	1171
Cluster shade	-2510	-3092	-1976
dissimilarity	0.2549	0.4516	0.4005
energy	0.3964	0.423	0.3414
entropy	-0.3278	-0.346	-0.2818
homogeneity	0.9428	0.9304	0.8996
Max probability	0.478	0.5347	0.5126
Sum of squares	28.2088	38.0687	21.3773
Sum of average	1.1912	2.8458	1.6078
Sum of variance	0.9994	0.9986	0.9987
Sum entropy	1610	2132	1171
Diff variance	-2510	-3092	-1976
Diff entropy	0.2549	0.4516	0.4005
Info measure of correlation 1	0.3964	0.423	0.3414
Info measure of correlation 2	-0.3278	-0.346	-0.2818
Inverse diff normalized	0.9428	0.9304	0.8996
Inverse diff moment normalized	0.478	0.5347	0.5126
autocorrelation	28.2088	38.0687	21.3773

2.3 Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model construction. Features are selected based on MRMR and tabulated in table.3.

2.3.1 Minimum-Redundancy and Maximum-Relevance (MRMR)

Feature-selection method that can use either mutual information, correlation, distance/similarity scores to select features [10]. The relevance of a feature set S for the class c is defined by the average value of all mutual information values between the individual feature x_i and the class c as follows

$$D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i; c) \qquad (7)$$

The redundancy of all features in the set S is the average value of all mutual information values between the feature x_i and the feature x_j:

$$R = \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i; x_j) \qquad (8)$$

The MRMR criterion is a combination of two measures given above and is defined as follows

$$\text{mRMR} = \max_{S} \left[\frac{1}{|S|}\sum_{x_i \in S} I(x_i; c) - \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i; x_j)\right] \quad (9)$$

Suppose that there are n full-set features. Let f_i be the set membership indicator function for feature x_i, so that f_i=1 indicates presence and f_i=0 indicates absence of the feature x_i in the globally optimal feature set. Let c_i=I(x_i;c) and a_{ij}=I(x_i;x_j).

The above may then be written as an optimization problem

$$\text{MRMR} = \max_{f \in \{0,1\}^n} \left[\frac{\sum_{i=1}^{n} c_i f_i}{\sum_{i=1}^{n} f_i} - \frac{\sum_{i,j=1}^{n} a_{ij} f_i f_j}{\left(\sum_{i=1}^{n} f_i\right)^2}\right] \quad (10)$$

Table.3 Feature Selection Result

Features	Normal	CLL	CML
Contrast	1.1912	2.8458	1.6078
Cluster prominence	1610	2132	1171
Cluster shade	-2510	-3092	-1976
Sum of squares	28.2088	38.0687	21.3773
Sum of average	6.8499	8.3153	5.2302
Sum of variance	70.1377	101.1808	47.8642
Sum entropy	0.8094	0.7115	1.1627
Diff variance	1.1912	2.8458	1.6078
Diff entropy	0.3995	0.424	0.676
Autocorrelation	27.7618	36.7772	20.6703

2.4 Classification

Classification is the task of assigning to the unknown test vector, a label from one of the known classes. Classification methods aimed to find mathematical models to recognize the membership of each object to its proper class on the basis of a set of measurements [7]. Once a classification model has been obtained [8], the membership of unknown objects to one of the defined classes can be predicted.

2.4.1 LS Boosting

- It fits regression ensembles
- At every step, the ensemble fits a new learner to the difference between the observed response and the aggregated prediction of all learners grown previously.
- The ensemble fits every new learner to $Y_n - \eta f(x_n)$, Where, Y_n - observed response, $f(x_n)$ - aggregated prediction from all grown weak learners, η - learning rate(0-1)

2.4.2 ADA Boosting

ADA Boost is a machine learning algorithm [4] that boosts the performance of other learning algorithms, known as weak learners, by weighting and combining them. The basic idea is that multiple weak learners can be combined to generate a more accurate ensemble, known as a strong learner. Various versions of the ADA Boost algorithm [6] have proven to be very competitive in terms of prediction accuracy in a variety of applications.

3. Performance Measures

The final results of the ensemble classifiers were analyzed and its performance evaluation is done based on the results obtained from the confusion matrix. Also the results are compared based on the three parameters namely accuracy, specificity and sensitivity.

3.1 Confusion Matrix

A confusion matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning .A table of confusion, is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. The table of confusion includes True Positive(Cancer identified as Cancer), True Negative(Cancer identified as Non Cancer), False Positive(Non Cancer identified as Non Cancer) ,False Negative (Non Cancer identified as Cancer)

3.2 Performance Parameters

The parameters are namely

(i) Accuracy - statistical measure of how well a binary classification test correctly identifies or excludes a condition.

$$\text{Accuracy} = \frac{\text{No of true positives + No of true negatives}}{\text{True positives + false positives + false negatives +true negatives}}$$

(i.e) Accuracy=(TP+TN)/(TP+FP+TN+FN)

where TP-True positive, TN-true negative,

TN-true negative, FN-false negative

(ii) Sensitivity -Measures the proportion of actual positives which are correctly identified as such and it is complementary to the false negative rate.

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{(Number of true positives + Number of false negatives)}}$$

(iii) Specificity - Measures the proportion of negatives which are correctly identified as such and is complementary to the false positive rate. Mathematically, this can also be written as

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Number of true negatives + Number of false positives)}}$$

(i.e.) Specificity = TN/(TN+TP)

4. Results

4.1 LS Boosting

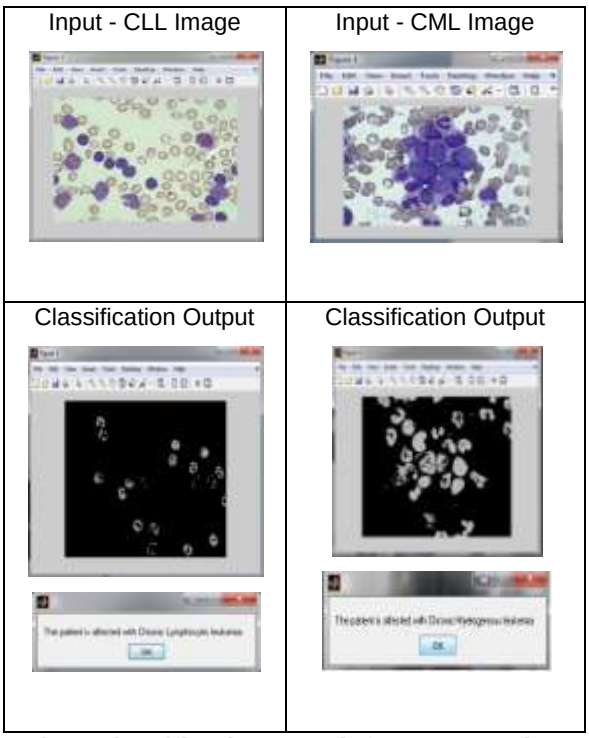

Fig.2 Classification Result for LS Boosting

4.2 Adaptive Boosting

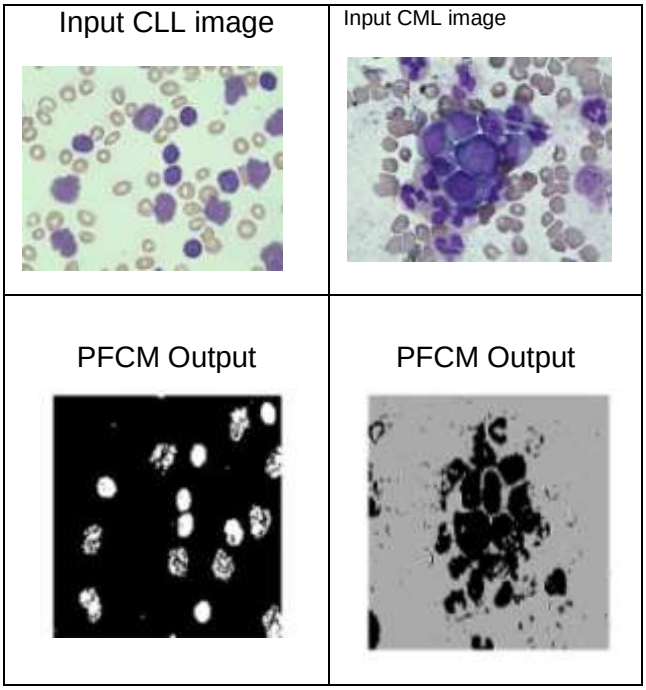

Fig.3 Classification Result for Adaptive boosting

4.3. Performance Comparison

The accuracy, specificity and sensitivity of adaptive and LS boosting is tabulated in table.4 comparison is done as shown in fig.4 where LS boosting has high performance

Table.4 Performance Comparison of ADA and LS Boosting

Algorithm/Parameters	Adaptive Boosting	LS Boosting
Accuracy	72.22%	88.88%
Specificity	0.5833	0.8333
Sensitivity	0.7917	0.9167

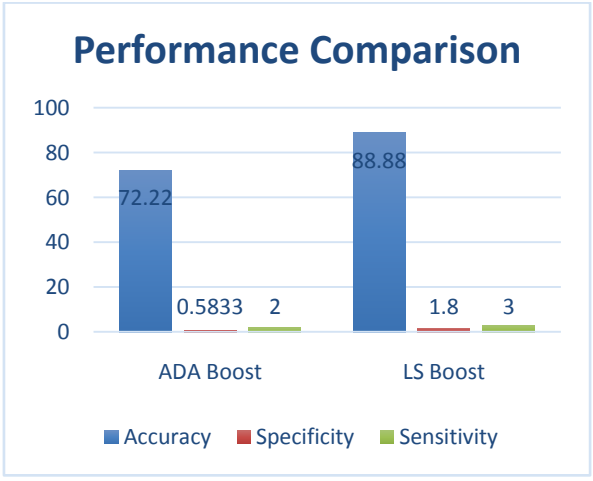

Fig.4 Performance Comparison of Classification algorithm

5.Conclusion

Thus the modified possibilistic fuzzy c-means algorithm approach was evaluated and tested for various blood cells. This algorithm gives the segmentation result of white blood cells. This algorithm begins with detecting the cells in the region. By using those regions, white blood cells alone segmented. Then classification algorithms such as adaptive boosting and LS boosting were implemented and performance is measured.

References

[1] Athira Krishnan, Sreekumar K , A Survey On Image Segmentation And Feature Extraction Methods For Acute Myelogenous Leukemia Detection in Blood Microscopic Images, International Journal Of Computer Science And Information Technologies, Vol. 5 (6) , 7877-7879,2014
[2] Arindam chaudhuri, intuitionistic fuzzy possibilistic c means clustering algorithms, advances in fuzzy systems, article id 238237, volume 2015
[3] Aruna bhat,possibility fuzzy c-means clustering for expression invariant face recognition, international journal on cybernetics & informatics vol. 3, no. 2, april 2014

[4] Claudia Henry Richard Nock, Frank Nielsen,Real Boosting a la Carte with an Application to Boosting Oblique Decision Trees, IJCAI-07

[5] Esteban Alfaro, Noelia García, MatíasGámez, and David Elizondo(2008). Bankruptcy forecasting: 'An empirical comparison of AdaBoost and neural networks'. DecisionSupport Systems, 45(1):110_122.

[6] Gall, J., Yao, A., Razavi, N., Van Gool, L., and Lempitsky,V. (2011)' Hough forests for object detection, tracking, and action recognition'. InPattern Analysis andMachine Intelligence.IEEE.

[7] Genuer, R., Poggi J.M., Tuleau-Malot C (2010)'Variable Selection Using Random Forests', Pattern Recognition Letters, vol. 31, pp. 2225-2236.

[8] Harun.N.H, A. S. Abdul Nasir, M. Y. Mashor, R. Hassan Unsupervised Segmentation Technique For Acute Leukemia Cells Using Clustering Algorithms International Journal Of Computer, Control, Quantum And Information Engineering Vol:9, No:1, 2015

[9] Hongwei hu , bo ma , yuwei wu , weizhang ma and kai xie,kernel regression based online boosting tracking*, journal of information science and engineering 31, 267-282 (2015)

[10] Hudson F. Golino1, Cristiano Mauro Assis Gomes2,Visualizing Random Forest's Prediction

Results, Published Online December 2014 in SciRes. http://www.scirp.org/journal/psych

[11] Kalaiselvi Chinnathambi, Asokan Ramasamy, Premkumar Rajendran, An Effective Clustering Technique For Wbc Image Segmentation And Its Classification, International Journal Of Digital Signal And Image Processing Vol. 2, No. 1(March 2014)
[12] Kulkarni V Y, Sinha P K, Petare M (2012), 'Analyzing Random Forest Classifier using Different Split Measures', Proceedings of International Conference on Soft Computing and Problem Solving, Jaipur, India. , Springer AISC series.

[13] Nazeeh Ghatasheh, Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study, International Journal of Advanced Science and Technology Vol.72 (2014), pp.19-30

[14] Nicolas De Jay , Simon Papillon-Cavanagh , Catharina Olsen2,an R package for parallelized mRMR ensemble feature selection, August 19, 2014

[15] Vrushali Y Kulkarni, Pradeep K Sinha ,Effective Learning and Classification using Random Forest Algorithm, International Journal of Engineering and nnovative Technology (IJEIT) Volume 3, Issue 11, May 2014.