ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 8, August 2014

Comparative Study Between Sparse Representation Classification and Classical Classifiers on Cervical Cancer Cell Images

Simi Susan Samuel[1], Anit V.Mathew[2], Subha Sreekumar [3]

P.G student, Computer Science and Engineering, Mangalam College of Engineering, Ettumanoor, Kerala, India[1]

P.G student, Computer Science and Engineering, Mangalam College of Engineering, Ettumanoor, Kerala, India [2]

Assistant Professor, Computer Science and Engineering, Mangalam College of Engineering, Ettumanoor, Kerala India [3]

Abstract: The pap-smear classification is still a challenging task for machine–aided cervix cancer diagnosis because it is tedious even for a trained cytologist to analyse and diagnose each slide obtained from different patients. Various methods have been proposed for medical image classification. In this paper, a multi-feature set sparse representation classification (mfSRC) is proposed in order to classify the pap smear cell images. Here, this representation goes through a training stage employing Genetic Algorithm guided by multi feature set dictionary learning approach. The data consists of 917 images of Pap-smear cells, classified carefully by cyto-technicians and doctors. Each cell is described by 20 numerical features, and the cells fall into 7 classes. In order to understand the relevance of sparse representation in the classification of pap smear images, the performance of other classical classifiers were also evaluated. Results show that classification accuracy of sparse representation generally outperforms other classical classifiers.

Keywords: Cervical cancer, Sparse Representation, Classical classifiers, Genetic Algorithm, Pap smear

I. INTRODUCTION

Cervical cancer is one of various types of cancer found in female, develops in the cervix. Cervical cancer can usually be found early by having regular screening with a Pap test. Pap smear [13] means human cells samples stained by the so-called Papanicolaou method. Being alert to any signs and symptoms of cervical cancer can also help avoid unnecessary delays in diagnosis. Early detection greatly improves the chances of successful treatment and prevents any early cervical cell changes from becoming cancerous. It is better to have regular screening of cervical cancer because women with early cervical cancer and pre-cancers usually have no symptoms, and symptoms will arise only after the invasive growth of cancerous cells.

Every year new approaches or hybrid artificial intelligence techniques are being proposed for such types of image classification. Aim of all techniques is to improve the classification accuracy, because the cancerous cells are very hard to distinguish. Different type of classification tasks are quite interesting and challenging in the biomedical field due to the amending results in digital image processing domains.

The pap smear database used here consist of seven categories of cervical cancer cells. The size, color, shape and the texture of the nucleus and cytoplasm is used. The density of cells in a certain area, can influence the diagnose. It takes a skilled cytotechnician, to be able to differentiate between the different cells. Every glass slide can contain up to 300000 cells. Therefore it is a time consuming job viewing the slides. Our method exploits the different appearance of nucleus and cytoplasm in different classes. Upon the multi feature set information, our method builds a sparse representation-based classifier for cervical cell cancer image classification with improved performance.

The Ant Colony Optimization (ACO) is used for the construction of a hybrid algorithmic scheme which effectively handles the Pap Smear Cell classification problem [2]. This algorithmic approach is properly combined with a number of nearest neighbour based approaches for performing the requested classification task, through the solution of the so-called optimal feature subset selection problem.

The automated segmentation technique for cytoplasm and nucleus segmentation is proposed in [3].

A hybrid intelligent scheme focussing on genetic algorithm based feature selection and nearest neigbor classification is also proposed for Pap Smear diagnosis [10] .

In recent years, the sparse representation-based classification (SRC) [6] has attracted many research interests due to the promising results in image and signal processing tasks [6], [11]. The generalization ability of SRC can be improved when SRC is learned from multimodal data[1]. Our experimental results validate the soundness of sparse representation classification followed by Genetic algorithm .The results are also compared with classical classifiers like Artificial Neural Network(ANN),Sequential minimal optimization(SMO), Adaboost, Naive Bayes Classifier.

II. METHOD AND DATA

A. Characteristics of cervical cancer cells

Specimens for diagnosis are taken from several areas of the cervix. The specimens most often contain cells from the columnar epithelium and the squamous epithelium.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 8, August 2014

The columnar epithelium is located in the upper part of the cervix, and the squamous epithelium in the lower part. Between these two is the metaplastic epithelium, also called the transformation zone or the squamo-columnar junction [9].

In the squamous epithelium there are 4 layers of cells. The cells form at the basal layer and while maturing they move up through the parabasal layer, the intermediate layer, and finally the superficial layer. The cells in the basal layer divide and deliver cells to the layers above it. While the cells mature and move through the layers, they change shape, color and other characteristics. When the cells reach the superficial layer they are rejected and replaced by the cells coming from below. The basal layer has small round cells with a relatively big nucleus and small cytoplasma. When maturing, the nucleus becomes smaller and the cytoplasma becomes larger. The shape of the cells becomes less round the more mature they are [9].

The columnar epithelium only contains a single layer of cells containing columnar cells and reserve cells. The reserve cells divide into new reserve cells and new columnar cells. In normal columnar epithelial cells, the nucleus is located at the bottom of the cytoplasm. When viewed from the top, the nucleus seems larger. When viewed from the side, the cytoplasm seems larger.

The metaplastic epithelium consists of reserve cells from the columnar epithelium. When the cells have matured fully in the metaplastic epithelium, they look like the cells found in the squamous epithelium. In dysplastic cells, the genetic information is somehow changed, and the cell will not divide as it should. This is a precancerous cell. Depending on which kind of cell that divides incorrectly, it is given diagnoses like dysplasia and carcinoma in situ. The term 'plasia' means growth, and 'dysplasia' means disordered growth. The dysplastic cells are divided into mild, moderate and severe dysplastic. The grading is determined from the likelihood of the cells later on turning into malignant cancer cells. A high amount of the mild dysplastic cells will disappear without becoming malignant, whereas severe dysplastic cells likely will turn into malignant cells. Squamous dysplastic cells generally have larger and darker nuclei and tend to cling together in clusters. In severe dysplasia, nuclei are large, with dark granules and usually deformed. The cytoplasm is small and dark compared to nuclei[9]. Cervical carcinoma is the most common malignancy of the female genital system. Carcinogenesis is a long-lasting process, which begins from normal epithelium, that becomes dysplastic, evolves to carcinoma in situ, and then to cancer. The long time interval between the stages, allows the possibility of an early diagnosis, with complete cure [9].

B. Database

A database of single cells has been collected at the Herlev University Hospital, Denmark, by means of a digital camera and a microscope. Skilled cyto-technicians and doctors manually classified each cell into one of 7 classes. Each cell was examined by two cyto-technicians, and difficult samples also by a doctor. In case of disagreement the sample was discarded. The database thus holds diagnoses that are as certain as possible, given the practical and economical constraints at the hospital. The pap-smear benchmark dysplasia, Moderate squamous non-keratinizing dysplasia, Severe squamous non-keratinizing dysplasia, Squamous cell carcinoma in situ intermediate, Superficial squamous epithelial, Intermediate squamous epithelial, Columnar epithelial. The distributions of cells are also specified in the pap smear benchmark data [9].

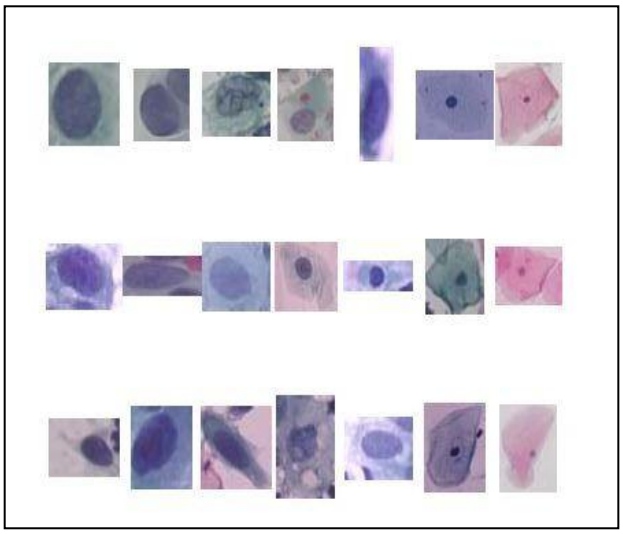

Fig 1. The seven classes of cervical cells categorized by skilled cytotechnicians and doctors [9].

Cells in the first three columns from the right are categorized as abnormal cells while the others are normal cells.

C. Feature Extraction

Feature extraction deals with converting information to a format that is usable for the classifier algorithms. For ex. pictures cannot easily be feed directly into the classifier algorithms. Instead special characteristics are extracted from the pictures. This phase is quite crucial for the success of the classification algorithm. Therefore when possible, experts in the actual field are used to identify the features. Feature extraction is used to find features that possibly can help in the classification. The 20 features are extracted from the segmented nucleus and cytoplasm and formed 3 feature sets instead of single feature set, because collected features consist of color, texture and geometric features and so keeping whole in one dictionary and implementing learning approach on whole dictionary will consider the relationship among these. This will increase the generalization error. The 20 features are shown in Table 1.

TABLE I
SUMMARY OF 20 FEATURES IN THE DATABASE

Nucleus area	Nucleus elongation
Nucleus perimeter	Maxima in cytoplasm
Cytoplasm area	Nucleus roundness
Cytoplasm perimeter	Minima in cytoplasm
N/C ratio	Cytoplasm shortest diameter
Cytoplasm perimeter	Cytoplasm longest diameter
Nucleus brightness	Nucleus position
Nucleus shortest diameter	Cytoplasm elongation
Maxima in nucleus	Cytoplasm roundness
Nucleus longest diameter	Minima in nucleus

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 8, August 2014

D. Method Framework

Initially the features specified above are extracted and made to occupy in three sub dictionaries. The sub dictionaries obtained in this way usually contain several similar samples coming from different classes, which might be harmful for classification. To learn discriminative sub dictionaries, we propose a genetic algorithm-based multi feature set dictionary learning algorithm, which selects the topmost discriminative training cell nuclei, and encourages large disagreement among different sub dictionaries. The resultant of this learning approach is optimised sub dictionaries.

When a new testing image comes, the 3 sub-dictionaries of features are created and SRC algorithm is employed on each sub dictionaries and the final label is predicted by taking a majority vote of sub dictionary labels.

III.PROBLEM FORMULATION

When considering a single dictionary for learning, it will consider the relationship among the features in them. This will increase the generalization error. The conventional SRC for single feature set introduced in [6] and it is reviewed in [1].

IV.TRAINING PHASE:GENETIC ALGORITHM

The idea of proposed method is being motivated from the paper[1]. Here multi feature based optimization on dictionaries using GA is presented. This optimization is done in order to promote the diversity among various dictionaries. So, the goal of this step is to select the samples from the three dictionariesF^1,F^2,F^3 and to form optimized O^1,O^2,O^3 respectively. For, this a binary sample selector S=[S^1 S^2S^3] where S^1 ,S^2,S^3 corresponds to the sample selectors of sub dictionaries F^1,F^2,F^3.The sample selectors are combination of 0's and 1's and the columns corresponding to 1 in sub dictionaries are selected to form optimised sub dictionaries and columns corresponding to 0 are not selected. The output of Genetic Algorithm is the best binary sample selector, which can able to select the best samples in original dictionaries. So in terms of GA algorithms these binary sample selectors are chromosomes. As usual, the steps of Genetic algorithm [8] includes chromosomes crossover , mutation and here we validated the chromosome each time.

The steps of GA are as defined:

Input: original sub-dictionaries

Output: Learned sub dictionaries

1. Initialize N chromosomes, $S^{1,1}$, $S^{1,1}$ …… $S^{N,1}$
2. For generations 1 to G
3. Evaluate the fitness of each individual.
4. While population < N
5. Select individuals from the population to be parents
6. Call crossover operator to produce offspring
7. Call mutation operator
8. End while
9. End for
10. Return the best individual (the best solution).
11. Selecting the samples from original dictionaries using the best individual to form O^1,O^2,O^3

V. SPARSE REPRESENTATION CLASSIFICATION

Sparse representations very much help in classification [12] [4][7]. When SRC is used with single feature set, there is only one dictionary D with any no:of classes. And a testing sample x of any class. The sparse representation problem can be formulated as

$$\hat{\alpha} = \{arg_x \min \|\alpha\|_1 \quad \text{s.t } \|x - D\alpha\|_2 \leq \in \quad (1)$$

$\in$ is the parameter to control the tolerance of the reconstruction error; α is the sparse linear combination coefficient we want to learn. (1) can be effectively solved by many algorithms, e.g., Basis Pursuit and Orthogonal Matching Pursuit .

The reconstruction error must be considered here to classify the testing sample x into one of the classes. The reconstruction error for each class is calculated and the label corresponding to class of least reconstruction error is assigned to the testing sample. In this paper this sparse representation classification is performed on each sub dictionaries corresponding to the sub dictionaries of testing sample so this classification can be called as multi feature set sparse representation classification. Then a majority voting is conducted among the labels predicted on sub dictionaries of testing sample in order to find the label predicted by classifier.

VI. OTHER CLASSICAL CLASSIFIERS

Sequential Minimal Optimization(SMO)

Training a Support Vector Machine (SVM) requires the solution of a very large quadratic programming (QP) optimization problem. SMO breaks this large QP problem into a series of smallest possible QP problems. These small QP problems are solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop. The amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets. SMO's computation time is dominated by SVM evaluation, hence SMO is fastest for linear SVMs and sparse data sets. It is a new SVM learning algorithm that is conceptually simple, easy to implement, is often faster, and has better scaling properties. Unlike previous SVM learning algorithms, which use numerical quadratic programming (QP) as an inner loop, SMO uses an analytic QP step. Because SMO spends most of its time evaluating the decision function, rather than performing QP, it can exploit data sets which contain a substantial number of zero elements[14].

A. Artificial Neural Networks (ANNs)

In machine learning and related fields, artificial neural networks (ANNs) [5] are computational models inspired by an animal's central nervous systems (in particular the brain) which is capable of machine learning as well as pattern recognition. Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs. Artificial neural network refers to the inter–connections between the neurons in the different layers.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 8, August 2014

An ANN is typically defined by three types of parameters:

- The interconnection pattern between the different layers of neurons
- The learning process for updating the weights of the interconnections
- The activation function that converts a neuron's weighted input to its output activation.

B. Naive Bayes Classifier

A naive Bayes classifier [5] assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix[5].

C. Adaboost

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. AdaBoost is an algorithm for constructing a "strong" classifier as linear combination of "simple" "weak" classifier.

VII. PERFORMANCE EVALUATION

A. Evaluation Metrics:

The TP rate, FP rate, precision, RMSE are employed for evaluating the multi classification results. The accuracy is calculated as the number of correctly classified images. The TP rate is the rate of correctly identified classes. The Precision rate is equal to TP rate divided by sum of TP and FP rate. The Root Mean Square Error (RMSE) (also called the root mean square deviation, RMSD) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power. The performance evaluation is as shown in fig.2

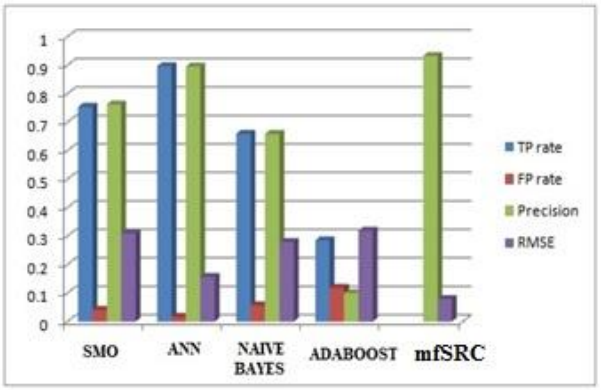

Fig. 2. Performance Evaluation of different algorithms on Pap smear images

B. Comparative Analysis

From graph itself, it is understood that mfSRC performs significantly better than classical algorithms. Here mfSRC outperforms others (93.3 % precision) even when it is doing a 7-class classification. SMO shows TP rate of 75.4 % which has a precision rate of 76.3%. Also the FP rate of this algorithm is .041 .The performance of Naive Bayes shows a TP rate value of 66%.The Adaboost algorithm has 28.6% TP rate ,showing a poor performance. The RMSE value for sparse representation is very much less when compared to other algorithms. The RMSE value of Adaboost is higher in this case.

VIII. CONCLUSION AND FUTURE SCOPE

In this study, a comprehensive survey between different algorithm for classifying cervical cancer cells is conducted. The aim is to study the relevance of sparse representation algorithm which is a different algorithm for classification on the cervical cancer cell images. For this the performance of some other classifiers like ANN, SMO, Bayesian, and Adaboost are evaluated along with Genetic algorithm plus SRC. And the SRC algorithm shows a good performance with an accuracy of 93.3%.The classical algorithm shows poor performance when compared to SRC in this field. Future work is to improve the execution of mfSRC by implementing a hybrid technique along with Genetic Algorithm. Also by replacing GA with any other optimization algorithm may result in amending performance of mfSRC.

ACKNOWLEDGMENT

The authors would like to thank Mr. Vinodh P. Vijayan for valuable discussions and comments that improved the quality and clarity of the manuscript.

REFERENCES

[1] Y.Shi, Y.Gao, ''Multimodal sparse representation based classification for lung needle biopsy images'' IEEE Trans on Biomedical Eng.,Vol.60,No.10,pp.2675-2685.
[2] Y.Marinakis and G.Dounias,''Nature inspired intelligence in medicine:Ant colony optimization for pap-smear diagnosis'', *Int. J. Artif. Intell. Tools* Vol.17, Issue 02, April 2008
[3] Zhi Lu, G.Carneiro, and Andrew P. Bradley, "Automated nucleus and cytoplasm segmentation of overlapping cervical cells" Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, Volume 8149, 2013, pp 452-460.
[4] W.Dai, B.Mailhé, & W.Wang ''Dictionary learning for sparse represenations:algorthms and applications'', ,May 2013.
[5] H,Bhavsar,.GAnatra,''A comparative study of training algorithms for supervised machine learning''.
[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y.Ma "Robust face via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009
[7] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50,no.10. M. Mitchell, ''An Introduction to Genetic Algorithm'', Cambridge, MA, USA:MITPress,1998.
[7] J. Jantzen, J. Norup, G.Dounias, B.Bjerregaard, '' Pap-smear Benchmark Data For Pattern Classification''
[8] Y Marinakis,G Dounias, '' Pap Smear diagnosis using a hybrid intelligent scheme focussing on genetic algorithm based feature selection and nearest neigbor classification''.
[9] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf.Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1794–1801.
[10] Roberto Rigamonti, Matthew A. Brown, Vincent Lepetit,'' Are Sparse Representations Really Relevant for Image Classification ?. Jan Jantzen and George Dounias ''Analysis of pap smear image data''John C. Platt,''Fast Training of Support Vector Machines''.