



International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 25 • 2017

Adaboost and Fuzzy Rule Based Ensemble Classifier in Predicting Cervical Cancer

Peter Kulandai Raj^a and S. Sheeja^b

^aResearch Scholar, Department of Computer Science, Karpagam University, Coimbatore – 641 021. Email: faffyroypeter@gmail.com

^bAssociate Professor and Head, Department of Computer Applications, Karpagam University, Coimbatore – 641 021. Email: sheejaajize@gmail.com

Abstract: Imaging development has constructed the clinical discoveries and technological innovations in the different fields, such as materials science, fundamental physics and digital computing. Advances in the medical imaging and their kinds have produced an array of contemporary methods for figuring out the disease and deriving uniquely valuable information from the testimony and class feature. Continuous development in the imaging era has improved the applications of the image processing and analysing is essential in new areas for high impact. Many researchers have focused to apply new techniques to identify the disease. Due to that in the present time the data accrued by the means of applying image processing strategies often allow not the most effective approach in detecting the disease, but additionally estimating its severity, there aren't many effective strategies centred inside the disease detection. Partial classification of the significant problem in the image processing when a disease must be diagnosed amidst numerous possible pathologies, it can be handy to perform a partial category, in which candidate areas are categorized as being the result of the disease of interest or not, instead of making use of a whole class into any of the viable sicknesses. To deal with this trouble, good sized observe of different strategies for constructing ensemble classifier is used. The proposed approach achieves the improvement in **classification of accuracy and possibly to extract the least number of features** which show the way to simplification of gaining knowledge learning to the task in predicting cervical cancer.

1. INTRODUCTION

Image processing and analysis are the most significant topics for the modern Society. Algorithms of the image processing and analysis are often used in many fields, for example, in Medical, Biology, Natural Sciences, Enterprise and Engineering. The improvements of the clinical technology over the recent years and increase of the connection among the structure and characteristic have made the imaging more essential field. The ever present of the digital generation has made the images as crucial part of wide variety of studies regions from Nano technology to Astronomy [1].

Single Models are naturally difficult to clear up the unbounded data due to the fact endless volume of data is continuously evolving. So, the new data is included into the model without repeating the whole learning system model and version are not followed even as the facts is conceptually up to date. Ensemble Methods that have lately obtained a great deal of interest inside the machine learning models. Ensemble strategies such as Boosting and Bagging have established to be quite green in a system studying framework.

Significant progress in image processing is led to a good number of real world applications. The improvement of the certain disciplines of computer science like artificial intelligence, pattern recognition, image processing, neural networks etc., promise the technological aid to address the various issues in machine vision. One of the applications considered in the present study is concerned with processing of images of the human body cell affected by the cervical cancer or prone to be affected by the cervical cancer in the anvil.

In the real world, medical specialists visually accomplished the inspection of human body cells and which is affected by different disease for recognition and classification. This manual evaluation process is however, tedious, time consuming and more ever very much subjective with human errors. The decisions making capability of the human inspector also depends on his repute.

An expert system is designed by W. Mitlehner, it's via meta model based on the priori information squamous lesions are decided and severity is concluded [2]. A composite classification scheme implemented by Ram Balasubramanian for combining several classifiers is a distinctly different design methodologies. The classifiers are selected from several state-of-the-art pattern classification schemes with a view to obtain superior performance. In the scheme, no priori information except a set of pre-classified data is assumed to be available and pap smear images being tested for biological cell classification [3]. The use of connectionist methods such as multi-layered perceptron, radial basis function (RBF) networks, and ensembles of such networks are investigated. [4] RBF ensemble algorithms based on the fluorescence spectra potentially provide automated and near real-time implementation of precancer detection in the hands of non-experts [5][6].

The literature survey has revealed that there is a fair amount of scope for cell ailment identification inside the area of scientific filed. There is a need to predict the disease that automatically apprehend, classify and quantitatively defects the cell disease signs. Many diseases exhibit trendy symptoms that because of the human body cells. For that, the proposed fuzzy based totally Ada boost ensemble multiclass classifier method is proposed to deal with the difficulty in discovering the ailment and to improve the category accuracy.

2. METHODOLOGY

The proposed fuzzy is based totally on the '*Ada boost ensemble multiclass classifier*' method. This proposed method is to deal with the difficulty in discovering the ailment and improve the category accuracy. Because, the boosting algorithms are originally designed to turn the weak classifiers into strong ones. The performance of such boosted classifiers turned out to be so good that they became an own discipline in the field of pattern recognition. Adaboost M2 [7] changes the goal of the weak classifiers to predict a set of plausible groups and evaluates the weak classifiers. Which penalizes the weak classifiers for failing to include the correct group in the predicted plausible group set and for each incorrect label in the predicted plausible group set.

2.1. Boosting Algorithm

"Boosting is a machine learning ensemble meta-algorithm for reducing bias primarily and variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones" as described in Wikipedia. Thus AdaBoost[8] attempts to overcome the difficulty by extending the communication between the boosting algorithm and the weak learner. First, it allows the weak learner to generate more expressive hypotheses,

which, rather than identifying a single label in Y instead choose a set of “plausible” labels. This may often be easier than choosing just one label.

It also allows the weak learner to indicate a “degree of plausibility.” Thus, each weak hypothesis outputs a vector $[0,1]^*$, where the components with the values close to the 1 or 0 correspond to those labels considered to be plausible or implausible, respectively. Note that the vector of the values is not a probability vector, i.e., the components need not to sum to one. While it gives the weak learning algorithm more expressive power, it also places a more complex requirement on the performance of the weak hypotheses. Rather than using the usual prediction error, it asks that the weak hypotheses do well with respect to a more sophisticated error measure that it calls the pseudo-loss. Unlike the ordinary error which is computed with respect to a distribution over examples, pseudo-loss is computed with respect to a distribution over the set of all pairs of examples and incorrect labels. By manipulating this distribution, the boosting algorithm focuses the weak learner not only hard to classify examples, but more specifically on the incorrect labels that are the hardest to discriminate which achieves boosting if each weak hypothesis has the pseudo-loss slightly better than the random guessing by using the fuzzy k -means algorithm.

2.2. Fuzzy-K means

The fuzzy k -means [7] clustering algorithm partitions data points into k clusters S_l ($l = 1, 2, \dots, k$) and clusters S_l are associated with the representatives (cluster centre) C_l . The relationship between a data point and cluster representative is fuzzy. That is, a membership $u_{i,j} \in [0, 1]$ is used to represent the degree of belongingness of the data point X_i and cluster centre C_j . Denote the set of data points as $S = \{X_i\}$. The FKM algorithm is based on the minimizing the following distortion:

$$J = \sum_{j=1}^k \sum_{i=1}^N \mu_{i,j}^m d_{i,j}$$

with respect to the cluster representatives C_j and memberships $u_{i,j}$, where N is the number of data points; m is the fuzzifier parameter; k is the number of clusters; and d_{ij} is the squared Euclidean distance between the data point X_i and cluster representative C_j . It is noted that $u_{i,j}$ should satisfy the following constraint:

$$\sum_{j=1}^k \mu_{i,j} = 1, \text{ for } i = 1 \text{ to } N$$

The major process of FKM [9] is mapping a given set of representative vectors into an improved one through the partitioning data points. It begins with a set of initial cluster centres and repeats this mapping process until a stopping criterion is satisfied. It is supposed to be that no two clusters have the same cluster representative. In the case that two cluster centres coincide, a cluster centre should be perturbed to avoid the coincidence in the iterative process. If $d_{ij} < \eta$, then $u_{i,j} = 1$ and $u_{i,l} = 0$ for $l \neq j$, where η is a very small positive number. The fuzzy k -means clustering algorithm is now presented as follows

- 01 Initial guesses for the means c_1, c_2, \dots, c_k
- 02 Until there are no changes in any mean:
 - 02.1 Use the estimated means to find the degree of membership $u(j, i)$ of x_j in Cluster i ;
for example, if $d(j,i) = \exp(-\|x_j - c_i\|^2)$, one might use

$$u(j, i) = d(j, i) / \sum_j d(j, i)$$

2.2 For i from 1 to k

2.2.1 Replace c_i with the fuzzy mean of all of the examples for Cluster i

$$C_i = \frac{\sum_j u(j, i)^2 x_j}{\sum_j u(j, i)^2}$$

2.2.2 end_for

3 end_until

2.3. Proposed Method of fuzzy based Ada Boost Ensemble Multiclass Classifier

The Ada BoostedM2 k -NN algorithm follows the basic structure of the Adaboost by iterating through the training set to produce an ensemble of classifiers as the proposed hypothesis. However, during each iteration, instead of testing the current hypothesis with the whole set of training instances, boosted fuzzy k -NN holds out a training instance and classifies the held-out instance using the rest of the training set. By using this “leave one out” method, each training instance will not help to classify itself. It is important because by not leaving the instance out, the fuzzy k -NN classifier [10] (distance weighted) will always achieve 100% training set accuracy, and boosting is not possible. In traditional boosting, an instance that is misclassified would have its weight changed. However, in the case of the fuzzy k -NN classifier, changing the weights of the misclassified instance will not help to classify itself. Therefore, during each iteration and for each training instance that is classified incorrectly, the algorithm will determine and modify the influence of its k -nearest neighbours. In the start of the Boosted k -NN algorithm, a weight term w_0i is associated with each training instance, and the weight terms are initialized to zero. Boosted fuzzy k -NN uses a k -NN with the distance weighting of $1/d$ to classify each instance using the rest of the training instances. [11] Then the Boosted fuzzy k -NN will modify the influence of the ‘ k ’ nearest neighbours in the following manner during each iteration. If a query instance is classified incorrectly, Boosted k -NN will examine each of its ‘ k ’ nearest neighbours, and modify their weights such that they are more likely to classify the instance correctly with the next iteration. Thus, the modified weight term will increase the value of the vote for the correct class and decrease the value of the vote for the incorrect class. Each iteration through the training set, boosted k -NN produces a model with the modified weight terms. Boosted fuzzy k -NN loops through the training set multiple times and returns an ensemble of models as the final hypothesis.

3. RESULT AND DISCUSSION






















In this section, we report our experimental results. It deals with the image features of the pap smears images observed through the microscope are categorized into Seven class data such as Normal Columnar, Normal Superficial, Normal Intermediate, Light dysplastic, moderate dysplastic, severe dysplastic, ‘carcinoma in situ’. Depending on the various features spelled out in the below table, an expert can identify if a woman can be identified for cervical cancer.

We Determine the cumulative reconstitution losses (i.e., the cumulative misclassification error of the labels. cumulative misclassification cost is defined for evaluating the performance of our fuzzy based totally Ada boost ensemble multiclass classifier

$$CMC = c \times FN_{rate} + FP_{rate}$$

where, FN_{rate} is the false negative rate, FP_{rate} is the false positive rate and c is the cost factor. C is the performance evolution while introducing the cost sensitive of the boosting algorithm. It becomes cumulative misclassification error when $c = 1$ which is presented in the Figure 1. The ensemble achieves a classification error of under 24% using 100 or more trees. It achieves the lowest error for 400 or more trees.

Table 1
Cervical cancer test and training datasets with multiclass

<i>N= Nucleus</i> <i>C= Cytoplasm</i>	<i>Normal</i> <i>Columnar</i>	<i>Normal</i> <i>Superficial</i>	<i>Normal</i> <i>intermediate</i>	<i>Light</i> <i>dysplastic</i>	<i>Moderate</i> <i>dysplastic</i>	<i>Severe</i> <i>dysplastic</i>	<i>Carcinoma</i> <i>'in Citu'</i>
Cell Image							
Cell Mask							
Binary image							
N_Area	804	1475	2396	3091	4527	5908	3271
C_Area	27804	65192	5390	6166	4328	14109	1816
N/C ratio	0	0	0	0	1	0	1
N_YCol	86	118	144	90	69	71	141
C_YCol	193	217	204	146	95	130	170
N_Short	30	33	43	60	72	72	57
N_Long	35	55	71	77	80	107	79
N_Elong	1	1	1	1	1	1	1
N_Round	1	1	1	1	1	1	1
C_Short	182	302	93	120	103	136	70
C_Long	242	340	127	123	132	195	96
C_Elonged	1	1	1	1	1	1	1
C_Rounded	1	1	0	1	0	0	0
N_Perimeter	101	141	188	257	245	285	228
C_Perimeter	674	1062	332	433	359	666	257
N_Position	0	0	0	0	0	0	0
N_Maximum	44	65	86	95	213	215	108
N_Minimum	37	49	55	60	213	210	79
C_Maximum	649	1728	114	172	126	357	56
C_Minimum	655	1674	147	158	189	343	91

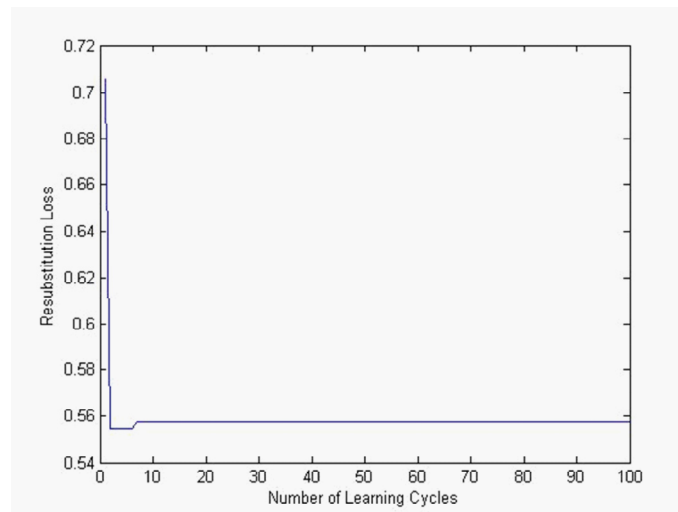


Figure 1: Resubstitution Loss for the number of learning cycles

Test classification error against the number of members in the ensemble. After the construction ensemble classifier, the study has evaluated the model by using the training dataset. Based on the performance evaluation, the following Figure 2 presented classification error against the training ensemble models.

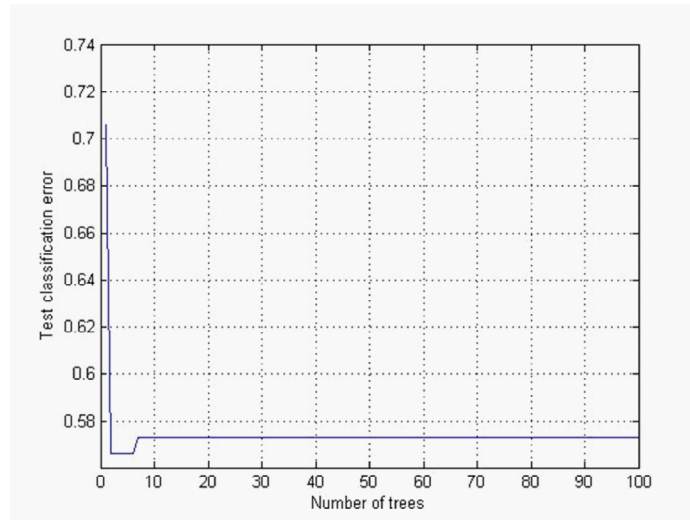


Figure 2: Test classification error

K-fold cross as the validation is one way to improve over the holdout method. The data set is divided into ‘ k ’ subsets, and the holdout method is repeated ‘ k ’ times. Each time, one of the ‘ k ’ subsets is used as the test set and the other $k - 1$ subsets are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is, how it matters less in the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set $k - 1$ times. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that the training algorithm is to rerun from scratch ‘ k ’ times, which means it takes ‘ k ’ times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set in k different times. Here 10-fold cross validation is used to cross-validated in the proposed ensemble classifier. The cross-validation error evaluated by using k -fold methods and the results is shown in Figure 3.

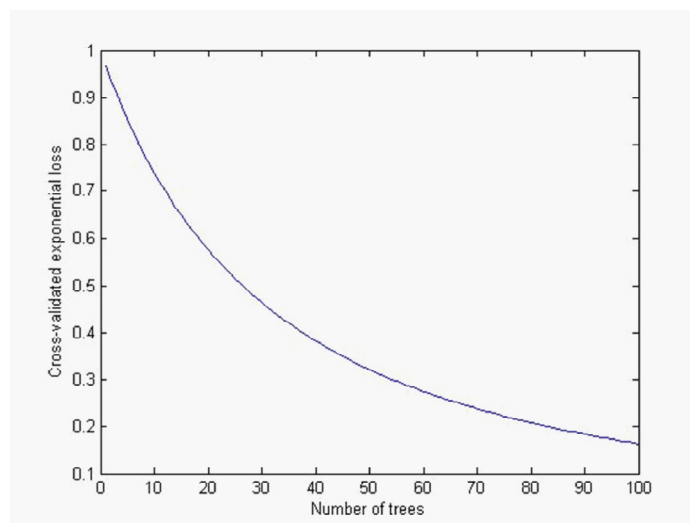


Figure 3: All classes of pap smear cells except the class 3 are having the 88% classification accuracy, and the classes 4 to 7 are having the 97% accuracy

The proposed framework facilitates anytime learning in the problems of arbitrary size and this method is also efficient, and as accurate as a single classifier, and adjusts quickly to bring changes in the target concept. It provides a natural mechanism for estimating the generalization accuracy of the model at any given stage of its construction. Moreover, the method is easy to implement and independent of the underlying classifier. The straightforward way of and comparing results presents the terms of accuracy. When comparing, the results based on accuracy appears in the single classifiers, our use of the ensemble classification suggests [12] significantly for better performance. One direction is to examine the diversity mechanism. We use different blocks of data for the various classifiers to create the classification diversity, which is necessary for the effective ensembles. We will experiment with other approaches to promote the diversity, such as the use of different classification methods, limited data re-sampling, and different methods of combining predictors. Another simple variation is to keep a working set of high-quality classifiers is slightly larger than the ensemble size, and test the all combinations based on the evaluation set. This “best n -of- m ” strategy moves us somewhat closer to a globally optimal set, while limiting the additional computation. To properly “optimize” an ensemble, it is important to have a clear understanding of the mechanisms through which the ensembles achieve their effectiveness. In our related work in the direct optimization of ensembles via evolutionary search, the study is in the early stages of a data mining task over the space of possible ensembles, varying ensemble characteristics and allowing a group of ensembles to compete for the limited resources based on accuracy [13].

The predictive accuracy on the new data might differ because of the ensemble accuracy might be biased. The bias arises because the same data is used for assessing the ensemble which is used for reducing the ensemble size. An unbiased estimate of requisite ensemble size can be obtained by using the cross validation. However, that procedure is time consuming.

4. CONCLUSION

This paper presents a proposed approach Adaboost fuzzy K-means clustering for the multiclass classifier. The results presented in this paper suggest that the creation of the ensembles of classifier can perform reasonably well. The proposed approach achieves improvement in the classification accuracy and simplify learning task. The method of multi class learning looks suitable for the ensembles building. Filtering gave surprisingly very good results as the method of diversity enhancement for this cervical cancer data set.

REFERENCES

- [1] Mao, Shasha, et al. Greedy optimization classifiers ensemble based on diversity. *Pattern Recognition* 44.6 (2011): 1245-1261.
- [2] W. Mitlehner, E. Cronin, J. Bronzino, R. Veranes, *Cytopath: An Expert System for The Classification and Diagnosis of Squamous Lesions in The Pap Smears of Pre-Menopausal Women*. Department of Engineering”, Hartford, CT, 1990.
- [3] Ram Balasubramanian, Sreeraman Rajan, RajamaniDoraiswami, and Maryhelen Stevenson, *A Reliable Composite Classification Strategy*, *The University of New Brunswick, Fredericton, Canada*, 1998
- [4] KaganTumer, Member, IEEE, Nirmala Ramanujam, Joydeep Ghosh, and Rebecca Richards-Kortum, *Ensembles of Radial Basis Function Networks for Spectroscopic Detection of Cervical Precancer*, *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 45, NO. 8, AUGUST 1998*
- [5] Haixun Wang, Wei Fan, Philip S. Yu and Jiawei Han. “Mining concept-drifting data streams using Ensemble Classifier.” *Proceedings of the ninth ADM SIGKDD international conference on knowledge discovery and data mining2003*: P(226-235).
- [6] Muezzinoglu, Mehmet K., and J. M. Zuracla. *A recurrent RBF network model for nearest neighbour classification*. *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*. Vol. 1. IEEE.

- [7] Zhang, Jianchun, and Daoqiang Zhang. A novel ensemble construction method for multi-view data using random cross-view correlation between within-class examples. *Pattern Recognition* 44.6 (2011): 1162-1171.
- [8] Wyner, Abraham J., et al. Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. *arXiv preprint arXiv:1504.07676* (2015).
- [9] Choi, Jae Young, et al. Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography. *Expert Systems with Applications* 46 (2016): 106-121.
- [10] Nguyen, DzungDinh, et al. Towards hybrid clustering approach to data classification: Multiple kernels based interval-valued Fuzzy C-Means algorithms. *Fuzzy Sets and Systems* 279 (2015): 17-39.
- [11] Huang, Ching-Wen, et al. Intuitionistic fuzzy C-means clustering algorithm with neighbourhood attraction in segmenting medical image. *Soft Computing* 19.2 (2015): 459-470.
- [12] Subudhi, Asit Kumar, Subhranshu Sekhar Jena, and Sukant Kumar Sabut. Delineation of infarct lesions by Multi-Dimensional Fuzzy C-Means of acute ischemic stroke patients. *Electrical Electronics, Signals, Communication and Optimization (EESCO), 2015 International Conference on. IEEE, 2015.*
- [13] Senthilarasu, S., and M. Hemalatha. Ensemble Classifier for Concept Drift Data Stream. *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)(2013): 1-4*