The 2010 IEEE/ICME International Conference on Complex Medical Engineering
July 13-15, 2010, Gold Coast, Australia

Clustering Nuclei Using Machine Learning Techniques

Yu Peng, Mira Park, Min Xu, Suhuai Luo, Jesse S.Jin, Yue Cui, W. S. Felix Wong
Leonardo D. Santos

Abstract — Cervical cancer is the second most common cancer among women. Meanwhile, cervical cancer could be largely preventable and curable with regular Pap tests. Nuclei changes in the cervix could be found by this test. Accurate nuclei detection is extremely critical as it is the previous step of analysing nuclei changes and diagnosis afterwards. Recently, computer-aided nuclei segmentation has increased dramatically. Although such algorithms could be utilised in the situation for sparse nuclei since they are intuitively detected, the segmentation for the complicated nuclei clusters is still challenging task. This paper presents a new methodology for the detection of cervical nuclei clusters. We first detect all the nuclei from the cervical microscopic image by an ellipse fitting algorithm. Second, we chose some high-relevant features from all the features we obtained in last step via F-score, which is based on to what extent one feature attributes to results. All the ellipses are then classified into single ones and cluster ones by C4.5 decision tree with selected features. We evaluated the performance of this method by the classification accuracy, sensitivity, and cluster predictive value. With the 9 selected features from the original 13 features, we came by the promising classification accuracy (97.8%) .

I.INTRODUCTION

Cancer is a group of diseases in which cells in the body grow, change, and multiply out of control. Cervical cancer refers to the erratic growth of cells that originate in the tissues of a cervix. It is usually a slow-growing cancer that may not have symptoms but can be found with regular Pap tests. According to U.S. National Cancer Institute, cervical cancer is the second most common cancer in women, and the third most frequent cause of cancer death, accounting for nearly 300,000 deaths annually worldwide, especially in middle and low income countries. Fortunately, cervical cancer could be largely preventable and curable with regular Pap tests, which is used to find cell changes in the cervix [1].

Recently, computer assisted screening and applications of digital image are widely reached for cervical cancer diagnosis and treatment [2].The use of image segmentation in Pap tests is increasing gradually. There is no doubt that the more cervical cells can be detected, the more analysis of cells change can be done. Abnormal cells could be treated before they turn into cervical cancer or in an early stage.

Images segmentation is the first step towards image understanding and image analysis [3].To increase the accuracy in computer-assisted diagnosis, accurate nuclei segmentation is crucial. After nuclei detection, the features of the individual nucleus could be obtained and analysed. Cytological features of a tissue image including nuclei count, nuclei size distribution, and nuclei shape distribution are significant features for decision making in pathology [4].

The features can be acquired easily by image segmentation when the nuclei are separated in images. However, in pathological conditions, nuclei in tissues are mostly clustered. Overlooking clustered nuclei and analysing only isolated nuclei can dramatically increase analysis time or affect the statistical validation of the result [5]. Therefore, the solution is accurately detecting clustered regions and isolated nuclei before applying segmentation algorithms such as a watershed algorithm, interactive region growing to segment clustered nuclei.

Currently, many techniques of the discriminating isolated nuclei and clustered nuclei have been employed based on a certain features of objects, such as object area, perimeter and circularity [6]. The convex hull based method is one of the most widely used techniques. Clustered nuclei could be identified when a ratio between the smallest convex polygon of each object and each real vector space beyond a certain threshold [3]. Hereby, the smallest convex polygon for a set of points(S) in a real vector space of S is the minimal convex set containing S. it is common to use the term 'convex hull' for this kind convex polygon. However, the inaccuracy could be brought about by the incomprehensiveness of cluster detection only based on few features. It is easy to omit some other features more related with result. And, this feature selection method itself is mainly based on experience. To solve this problem, we proposed a decision tree based method to detect clustered nuclei by using as many features of each object as possible.

Manuscript received February 5, 2010. This work was supported in part by the CSC-Newcastle Scholarship ,ARC LP0669645 and IntelliRAD .Yu Peng, Jesse S. Jin, Mira Park, Suhuai Luo, Min Xu and Yue Cui are with the School of Design, Communication and IT, University of Newcastle, Callaghan 2308 Australia (e-mail: Yu.Peng@uon.edu.au). W. S. Felix Wong is with the Department of Obstetrics and Gynaecology, School of Medicine, University of New South Wales, Sydney, NSW 2052, Australia. Leonardo D. Santos is with Department of Anatomical Pathology, Sydney South West Pathology Service, Liverpool Hospital, Liverpool NSW 2170, Australia (e-mail: Leonardo.Santos@sswahs.nsw.gov.au).

In this paper, we first utilise ellipse detection to obtain the potential nuclei including clustered and isolated ones. Secondly, relative features are extracted for all the ellipses. Thirdly the features are selected with high result relevance by applying F-score. Finally C4.5 decision tree is generated to discriminate the clustered nuclei. Fig.1 illustrates the system overview.

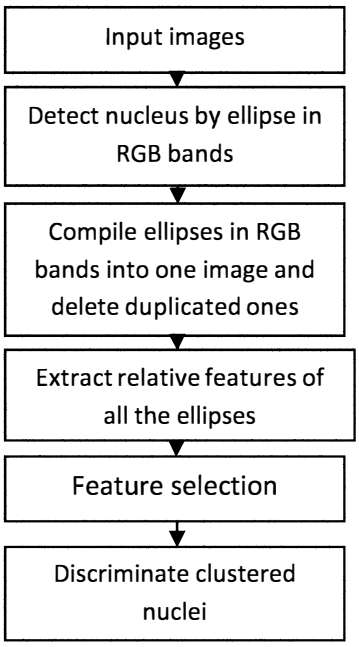

Fig.1.the system overview

II. METHODOLOGY AND EXPERIMENTS

A. Cervical Images

Cell images were acquired from the cervix uteri. Those cells were dyed with Ki 67 to observe with fluorescent microscopes [7]. The analogue image minified by 400 times in the microscope is digitized in the images grabber to 24 bits RGB images with a resolution of 1200×1600 pixels. Ten cell images are used in experiments. The reason we only choose 10 images is there are a big amount of cells in each image. Actually, we obtained 1674 candidates from these 10 images in all, which is an acceptable number for our experiment.

B. Preprocessing

We apply an adaptive nonlinear diffusion algorithm to remove noise from the image. The nonlinear anisotropic diffusive process has shown the good property of eliminating noise while preserving the accuracy of edges [8]. Moreover, we choose an adaptive nonlinear diffusion algorithm applied the central limit theorem to select the threshold [3]. This method could avoid the filtering threshold varies in a different image situation.

C. Creation of Ellipse

Colour is perceived by humans as a combination of tristimulus R (red), G (green) and B (blue), which are usually called three primary colours [9]. We use these three colour spaces separately to increase the accuracy of ellipses. Because if we simply extract the intensity layer of the images first and then apply gray scale methods directly on them, we will ignores the chromaticity information in the image [10].

We utilised the edge-based ellipse detection algorithm to the three colour bands of all the images. We then combined all the ellipses of three bands of each image into one in order to delete the repeated ellipses by setting the threshold of distance and angle and selecting the ellipses for the best candidates of nuclei .AsFig.2shown, all the single and cluster ellipses could be obtained in (d) after combining the results in (a), (b) and (c). The problem of cell clusters detection is transferred to discriminate single and cluster ellipses.

D. Extraction of Relative Ellipse Features

After ellipses fitting, potential single nucleus and nucleus clusters have been modelled as ellipses. Each ellipse has basic parameters such as a long axis, a short axis, centre of an ellipse, centre of a nucleus covered by one ellipse. With help of these parameters, we can obtain various meaningful features of ellipses, such as ratio of ellipse area and image area, ratio of ellipse perimeter and image perimeter, convex hull mentioned above and so on. F-score is calculated for these features to indicate their importance. After that, we choose the features whose F-score is in an acceptable scope to train decision tree.

E. Feature Selection

When using C4.5 decision tree, it is advantageous to limit the number of input features in the procedure of training tree in order to have a good predictive and small computationally intensive model. Because some of the features we acquire are little relevant with the final result of the tree. With a small feature set, the explanation of a rationale for the classification decision can be readily realized [11].

F-score [12] is a simple but effect technique which measures the discrimination of two sets of numbers. Given training vectors $x_k, k = 1,2,\dots,m$,if the number of positive and negative instances are $n_+$and n_-respectively, then F-score of the ith feature is defined as

$$F_{(i)} = \frac{(\overline{x_i}^{(+)} - \overline{x_i})^2 + (\overline{x_i}^{(-)} - \overline{x_i})^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(\overline{x_{k,i}}^{(+)} - \overline{x_i}^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(\overline{x_{k,i}}^{(-)} - \overline{x_i}^{-)})^2} \quad (1)$$

Where $\overline{x_i}, \overline{x_i}^{(+)}, \overline{x_i}^{(-)}$are the averages of the ith feature of the whole, positive, and negative datasets, respectively;

$\overline{x_{k,i}}^{(+)}$ is the ith feature of the kth positive instance, and $\overline{x_{k,i}}^{(-)}$ is the ith feature of the kth negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative [12].

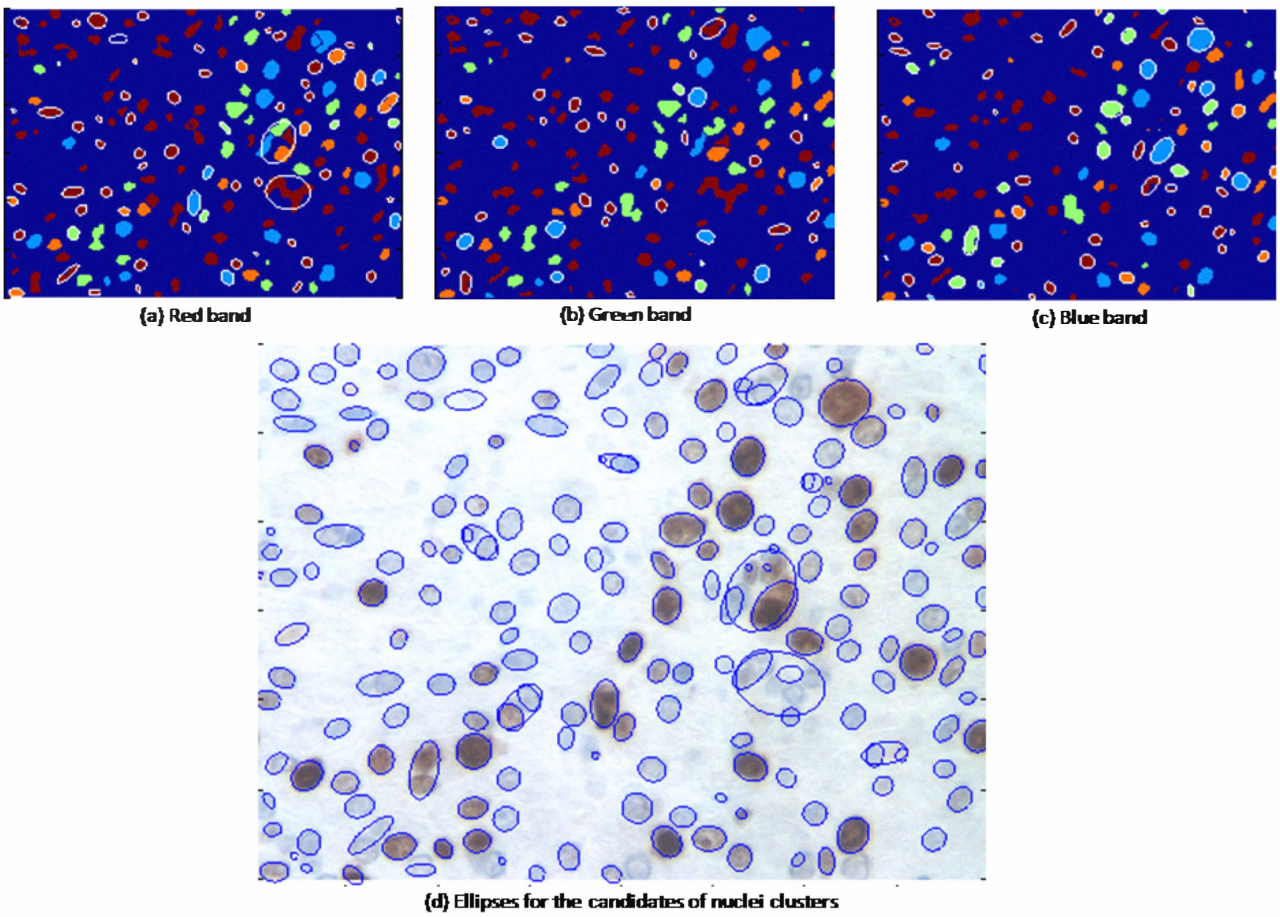

Fig. 2. (a) Fitted ellipses (in white lines) for segments of nuclei in a red band image. (b) Fitted ellipses (in white lines) for segments of nuclei in green a band image. (c) Fitted ellipses (in white lines) for segments of nuclei in a blue band image. (d) The original image is superimposed with the candidates of nuclei clusters.

F. C4.5 Decision Tree with Features Selected

We first grow the tree, using a set of training data, quite often to its largest size. Secondly, we prune the tree to a smaller one by the excepted errors when testing this tree on the unseen cases. Finally, the classification rules would be generated from the tree with pruning. The three steps are as follow:

1) Constructing Decision Tree: If any algorithm can be said to as the foundation of his program, it is the process of generating an initial decision tree from a set of training cases [13].

Moreover, as Fig.3 shows the core in the procedure of building tree is *gain ratio* criterion, express the proportion of information generated by the split that is useful, then

$$gain\ ratio = \frac{gain}{split\ \ info} \qquad (2)$$

Where $gain$ measures the information that is gained by partitioning input data according to the split while $split\ \ info$ represents the potential information generated by dividing input data via the split.

Initially, we calculated $gain\ ratio$ for each possible split since all the features had continuous values, by which we decided the best split with the highest $gain\ ratio$. Once the split of each feature is generated, a comparison among all the features is carried out to find one with maximum $gain\ ratio$ as a node. Finally the training data was parted into two subsets by this node. After then, we iterate the above steps on the two subsets until all the cases were be classified.

2)Pruning Decision Tree: The initial one we obtained as above is a complicated tree that over fits the data by inferring more structure than is justified by the training

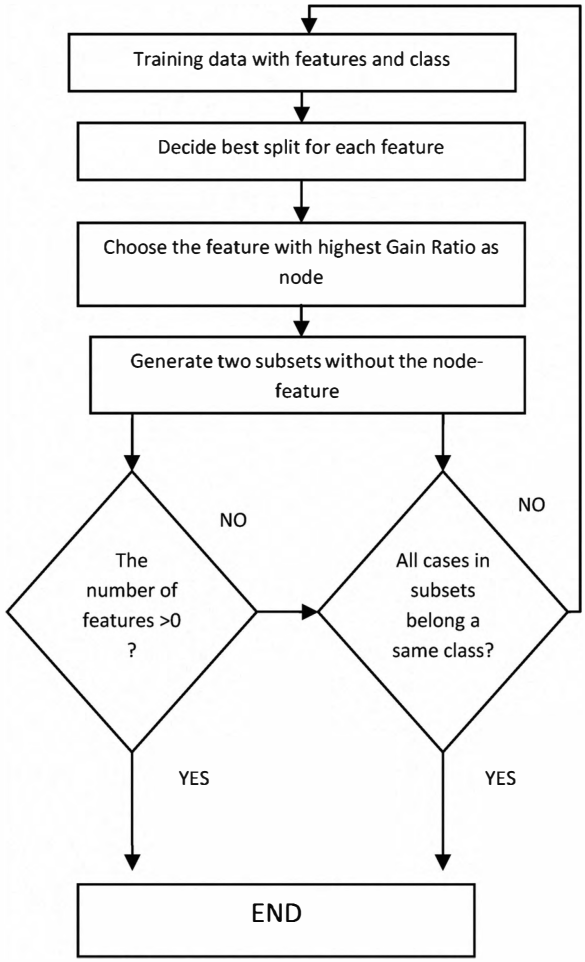

Fig.3.the procedure of constructing the initial tree

cases. As many authors mentioned, most of all decision trees can benefit from simplification [14].

In this work, we chose error-based pruning. When N training cases are covered by a leaf, E of them incorrectly, the resubstitution error rate of this leaf is E/N. For a given confidence level CF, the upper limit on the probability of expected errors on unseen cases can be found from the confidence limits for the binomial distribution, then

$$\text{Expected erros} = U_{CF}(E, N) \times N \qquad (3)$$

In most of simulation conducted with C4.5 decision tree classifier, confidence level is chosen to be equal 25% [15]. The main ideal: starts from the bottom of the tree and examines each nonleaf subtree. If replacement of this subtree with a leaf, or with its most frequently used branch, would lead to lower predicted errors, we prune the tree accordingly.

G. Measures for Performance Evaluation

We have used confusion matrix and the effectiveness index produced from it including classification accuracy, sensitivity and cluster (positive) predictive value to evaluate our proposed method. A confusion matrix contains information about actual and predicted classifications done by a classification system [16]. Table I shows the confusion matrix for a two class classifier.

Table I confusion matrix

Actual	predicted	
	positive	negative
positive	True positive(TP)	False negative(FN)
negative	False positive(FP)	True negative(TN)

$$\text{Classification accuracy } (\%) = \frac{TP+TN}{TP+FP+FN+TN},$$

$$\text{Sensitivity } (\%) = \frac{TP}{TP+FN},$$

$$\text{Positive predictive value} = \frac{TP}{TP+FP} \times 100$$

III.RESULT

To evaluate the effectiveness of our approach, we conducted experiments on the cervical images. All the ellipses in RGB bands were complied into one image as deleting duplicated ones after using ellipse detection in the three bands for each image. Table II shows all the ellipses after compiling of 10 images. All clustered nuclei and isolated nucleus are enclosed by the ellipses. Therefore, the problem of detecting the clustered cells was transferred to classify the two kinds of ellipses: single and cluster, by whose features. Table III shows the 13 relative features of all the ellipses.

Table II all the valid ellipses in 10 images after compiling

images	cluster	single	all
1	16	149	165
2	22	166	188
3	3	130	133
4	5	214	219
5	16	167	183
6	21	99	120
7	6	155	161
8	7	169	176
9	6	188	194
10	7	128	135

Table III list of 13 relative features of each the ellipses

No.	features	description
1	av	Average intensity of ellipse
2	std	standard intensity of ellipse
3	sim	Similarity of image and ellipse
4	dist	Distance between image and ellipse
5	areaxy	Area of image
6	areaEll	Area of ellipse
7	perimxy	Perimeter of image
8	perimEll	Perimeter of ellipse
9	Bwec	Eccentricity of ellipse
10	Bweq	the diameter of possible circle of ellipse

11	solidity	ratio image area and convex area
12	ratioa	Area ratio of image and ellipse
13	ratiop	Perimeter ratio of image and ellipse

The importance of each feature is evaluated by F-score. Table IV shows the importance of the relative features by F-scores on the training set with 50-50% training-test partition. The degree of cluster with features from high to low, are $F_7, F_{11}, F_{10}, F_8, F_4, F_6, F_5, F_2, F_9, F_{12}, F_3, F_{13}$ and F_1.

Table IV the 13 features with F-score

No.	features	F-score(50-50%training-test partition)
1	av	0.0025
2	std	0.9856
3	sim	0.3042
4	dist	1.6929
5	areaxy	1.4451
6	areaEll	1.6827
7	perimxy	2.2729
8	perimEll	1.9099
9	Bwec	0.4811
10	Bweq	1.9994
11	solidity	2.0213
12	ratioa	0.3641
13	ratiop	0.0786

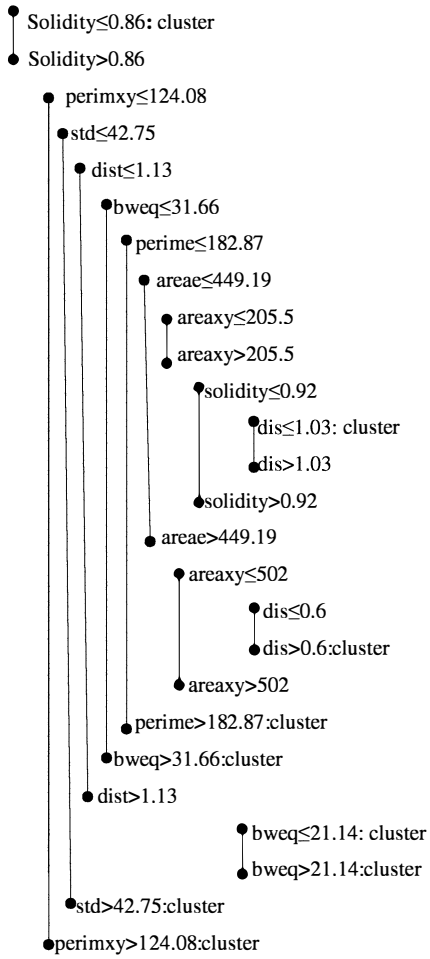

Fig.3. the initial tree from the training data

Therefore, we chose the first 9 features, which are with relative high F-scores (all above 0.98), as the inputs for training the decision tree. Fig.3 shows the decision tree without pruning from the training data covering image 1,2,3,4 and 5. However, this tree is quite complex that need some improvement. As explained above, Fig.4 shows the decision tree after pruning based the judgement of expected error numbers.

We present values of classification accuracy, sensitivity, positive predictive value in table V.As we can see from Table V, our method could discriminated single and cluster ones excellently with a promising classification accuracy (97.8%), moreover, the acceptable outcome of sensitivity (85.1%) demonstrated most of the ellipses covering clustered cells can be indicated. Nevertheless, the not that satisfied positive predictive value (80%) indicated tiny amount of single ones are be considered as clustered by our approach.

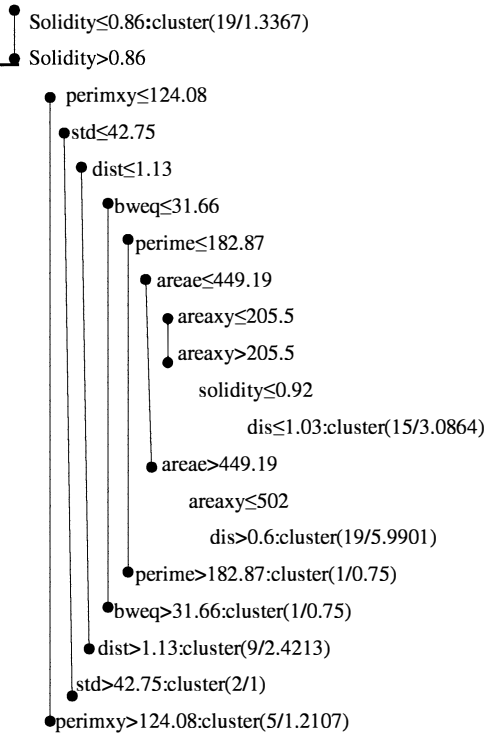

Fig.4. the decision tree after pruning with (N/E), N is the number of unseen cases with E is expected errors when the rule is used on the unseen cases

Table V classification accuracy, sensitivity and positive predictive value for this method

Actual	predicted	
	positive	negative
positive	40	7
negative	**10**	**729**

Classification accuracy	Sensitivity	Positive predictive value
97.8%	85.1%	80%

IV.CONCLUSION

In this study, we proposed an approach based on C4.5 decision tree with feature selection for nuclei cluster detection. This method could detection nuclei clusters efficiently applying features as many as possible without high computational cost. As an extension of this work, we plan to extract features from a larger image dataset. It is observed that our method yields the promising performance of cluster cells decision. This work can be extremely helpful for accuracy and avoiding information lost in image segmentation.

REFERENCE

[1] Available: http://www.cancer.gov/ cancertopics/types/cervical.

[2] D.Parry,E.Parry,"Medical Informatics in Obstetrics and Gynecology",Ideal Group Inc,2009,ch.11.

[3] M. Park, et al. "Microscopic Image Segmentation Based on Color Pixels Classification," The First International Conference on Internet Multimedia Computing and Service, 2009.

[4] S.Kothari and Q. Chaudry, et al. "Extraction of Informative Cell Features by Segmentation of Densely Clustered Tissue Images." Conf Proc IEEE Eng Med Biol Soc 1: 6706-9,2009.

[5] N.Malpica and C. O. de Solorzano, et al. "Applying Watershed Algorithms to The Segmentation of Clustered Nuclei." Cytometry 28(4): 289-97,2009.

[6] R.M.Haralick,"Computer and Robot Vision",Vol.I.,Addison-Wesley,1992.

[7] T.Nakano and K.Oka,"Differential Values of Ki-67 Index and Mitotic Index of Proliferating Cell Population",Cancer,7280:2401-2408,1993.

[8] J.S.Jin, Y.Wang and J. Hiller, "An Adpative Nonlinear Diffusion Algorithm for Filtering Medical Images",IEEE transaction on Information Technology in Biomedicine,4(4):298-305,2000.

[9] H.D.Cheng, X.H.Jiang,Y.Sunand and J.Wang,"Color Image Segmetation:advances and prospects",Pattern Recongnition,34:2259-2281,2001.

[10] Y. Mei, D. Androutsos, "Wavelet-based Color Texture Retrieval Using The Independent Component Color Space." Conf IEEE International Conference in Image processing,,978-1-4244-1764-3,2008

[11] M. F.Akay,"Support Vector Machine Combined with Feature Selection for Breast Cancer Diagnosis",Expert Systerms with application 36:3240-3247,2009.

[12] Y.W.Chen and C.J.Lin, " Studies in Fuzziness and Soft Computing", Berlin :Springer, 2008,ch.4.

[13] J.R.Quinlan,"C4.5 Programms for Machine Learning",Mrogan Kaufmann,1992,ch.2.

[14] J.R.Quinlan,"C4.5 Programms for Machine Learning",Mrogan Kaufmann,1992,ch.4.

[15] J.R. Beck, and M.Garcia ,et al."A Backward Adjusting Strategy and Optimization of The C4.5 Parameters to Improve C4.5's Performance",Proceedings of 21st in ternational FLAIRS Conference,2008.

[16] R.KoHAVI and F.Provost ,"Glossary of Terms",Editorial for the speical issue on application of Machone learning and the Knowledge Discovery Process,30(2-3),1998.