A Comparison of Neural Network and Fuzzy c-Means Methods in Bladder Cancer Cell Classification

Y. Hu and K. Ashenayi
Electrical Engineering Department, University of Tulsa, Tulsa, OK 74135
R. Veltri, G. O'Dowd and G. Miller
CytoDiagnostics Inc. Oklahoma City
R. Hurst and R. Bonner
Urology Department, Oklahoma University Health Sciences Center, Oklahoma City

Abstract

We report the performances of cancer cell classification by using supervised and unsupervised learning techniques. A single hidden layer feed-forward NN with error back-propagation training is adopted for supervised learning, and c-means clustering methods, fuzzy and non-fuzzy, are used for unsupervised learning. Network configurations with various activation functions, namely sigmoid, sinusoid and gaussian, are studied. A set of features, including cell size, average intensity, texture, shape factor and pgDNA are selected as the input for the network. These features, in particular the texture information, are shown to be very effective in capturing the discriminate information in cancer cells. It is found, based on the data from 467 cell images from six cases, the neural network approach achieves a classification rate of 96.9% while fuzzy c-means scores 76.5%.

1. Introduction

Until recently, bladder cancer was diagnosed almost exclusively by either cystoscopy, wherein a fiber optic device is inserted into the bladder and lesions are detected visually by a urologist, or by conventional Papanicolaou staining of bladder cells obtained from urine or from a bladder wash(hereafter called "conventional cytology")[6]. The major use of cystoscopy is to detect tumors in patients expressing the symptom complex characteristic of bladder cancer, which do not occur until the tumor has progressed to a more dangerous grade or stage. The difficulty with conventional cytology is that its recognition rate to low grade lesions is highly sensitive to the training of the cytopathlogist. Human can learn to recognize bladder cancer cells visually, but the process of screening samples generally requires a high level of skill and knowledge [6], and the work is generally fatiguing and boring due to its repetitive nature, therefore results are sometime inconsistent. Automatic cell classification has been studied for decades by using conventional pattern recognition techniques [7][10], here we would like to look into the application of neural network(NN) and fuzzy c-means for bladder cancer cell classification, and compare their performances.

Neural network has been utilized in many areas due to its potential high speed inherent in its parallel architecture, learning ability and non-linear classification nature. Among various successful applications, pattern recognition is one in which NN has shown results comparable or superior to the conventional approaches.

From pattern recognition view point, neural network, to the essence, constructs a non-parametrical discriminate surface—boundary—in its often multidimensional input vector space. This surface is built up progressively by exploring the discriminate information from labelled patterns in the training process. The trained network is then used to classify future patterns by extrapolating the information learned. The discriminate surface is virtually coded into weights and activation function threshold values of the networks during training process.

In this work a single hidden layer feed forward neural network with error back propagation training algorithm is used. Back propagation algorithm is a gradient based learning procedure. Although it has the drawback of getting stuck into local optima, back propagation is by far the most popular method used and does perform well.

Different from neural network, clustering classification takes an unsupervised learning approach. By giving the number of clusters in the data, the clustering algorithm explores the underlying structure in the data and automatically partitions them into groups. This unsupervised learning process typically produces a set of centroids representing the prototypes of data groups, and they are used for future classification[5]. Conventional clustering techniques tend to make a 'hard' partition on the data. This is suitable for well-separated-compact data. In cases when there are no clearly separated clusters in the data, the introduction of fuzzy concept will help to reach a better partition[9][3]. Among various fuzzy version of clustering techniques, fuzzy c-means is the one used most frequently, which is also adopted for our cancer cell classification.

0-7803-1901-X/94 $4.00 ©1994 IEEE

In section 2, we present neural networks of various configurations and describe the fuzzy c-means technique. Section 3 considers the feature extraction. Experiment results are presented in section 4. Section 5 is discussions and conclusion.

2. Principles and Configurations

In a multilayer neural network, hidden layers are of particular importance. How well and how quick the network converges to an approximation of the discriminate surface, to a large degree, depends on the number of hidden neurons and the type of the activation function used for each neuron. Too many hidden neurons will degrade the generalization capability of the network. The issue can be directly analogized to polynomial curve fitting. Allowing too few or too many parameters to be used in the polynomials will lead to under or over fitting. Therefore, there is an optimal number of hidden neural number for each individual recognition task. Works have been done in trying to quantify the generalization quality by using the concept of entropy or other complexity measurements.

Neural network can be perceived to have an underlying function decomposition mechanism [4]. An arbitrary function, e.g. the discriminate surface in classification application, can be represented by a collection of simple primitive functions, which corresponds to the activation function associated with each neuron. It has been proven, meeting certain conditions, neural networks with many types of activation functions are convergent [4]. Studies show the often used sigmoid activation function is not necessarily the optimal choice. It has been suggested in certain class of problems the use of sinusoid or gaussian activation functions reduces the training time substantially [1][2]. In this work, sigmoid, gaussian and sinusoid activation functions, denoted by $f_1(x), f_2(x)$ and $f_3(x)$ respectively, are used in the network:

$$f_1(x) = \frac{1}{1 + \exp(-\frac{x}{a})} \tag{1}$$

$$f_2(x) = \exp(-\frac{x^2}{2\pi\sigma^2}) \tag{2}$$

$$f_3(x) = \frac{1+sin(fx)}{2} \tag{3}$$

The c-means clustering is an iterative process which partitions the given data set $X = \{x_i: 1<i<n\}$ into "c" clusters by minimizing the within cluster sum of distances J_m, defined as

$$J_m(U,v) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \|x_k - v_i\|^2. \tag{4}$$

U is a matrix composed of $u_{ik} \in [0,1]$, u_{ik} describes the belongness of x_i to cluster k,

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{\|x_k - v_i\|^2}{\|x_k - v_j\|^2})^{1/(m-1)}}, \quad 1<i<c; \ 1<k<n, \tag{5}$$

$v = \{v_i: 1<i<c\}$ and v_i is the centroid of ith cluster,

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}, \quad 1<i<c. \tag{6}$$

In above equations, m controls the degree of fuzziness. As can be seen from Eq. (5), with larger value m, the membership u_{ik} tends to be closer to $1/c$, therefore fuzzier. When u_{ik} is clipped to 0 or 1, the clustering becomes a conventional 'hard' partition one. If one imagines U as a multi-dimensional vector, in case of conventional c-means, as u_{ik} in U is chosen from $\{0,1\}$ the convergence of U will effectively follow a staircase like path. However, for a fuzzy c-means approach, as $u_{ik} \in [0,1]$ the path will be a continuous and smooth one, therefore the clustering is less likely trapped into local optima.

3. Feature Extraction

The cell images are obtained through microscope and individually separated. Ideally, raw images (gray scale values) should be used directly as network input as they contain all original information. If the network can explore the discriminate information coming with the raw images itself, the hidden features will be revealed. However, the use of raw image as the input leads to a large input data size, and consequently substantially increases the complexity of the network as well as training time. Perhaps the more serious problem is the lacking of the invariant property in a trained network.

In pattern recognition, only the discriminate information contributes to correct identification of objects, while the rest does not or even degrades the performance. Feature extraction is to map the raw data into feature domain, while at the same time preserving the discriminate information of the original data. The direct benefit of feature extraction is the substantial reduction of input data size. For object recognition, if the features are chosen to be invariant to geometrical transformations, the classification performance will be significantly improved.

By carefully observing the cell images it is revealed that the abnormal cells have either a larger size, irregular shape, rougher surface or darker appearance. Based on above observations and with hardware implementation in mind, a set of four simple visual features, including area, average intensity, shape factor(roundness) and texture, are identified. All of them can be realized in hardware without much difficulty. In addition, pgDNA value is also used, although this does not represent a visual property.

Shape Factor

As most non-cancer cells have a close to round shape, and the cancer cells look more irregular, roundness factor seems to be a simple and effective discriminate feature. This is calculated as

$$\text{Shape Factor} = \frac{\text{perimeter}^2}{\text{area}} \qquad (7)$$

An ideal circle will give a shape factor of 4π, while any shape other than a circle will produce a value greater than 4π.

Texture

Texture is an important visual feature for many pattern recognition tasks. Texture describes the interdependent characteristics of pixels within a neighboring area. Regular texture has more or less periodical patterns, while random texture is best described by its 'coarseness'. Various statistical models can be used for feature extraction from a random texture image [8]. To minimize the computation, only simple convolution by using a 3×3 mask is considered. This effectively extracts the high frequency information from an image.

Fig. 1 shows pixels in a small area of cell image and the convolution mask applied. From the configuration of kernel it can be found this mask has the effect of high pass filtering. In fact the mask resembles a Laplacian kernel, commonly used for edge sharpening.

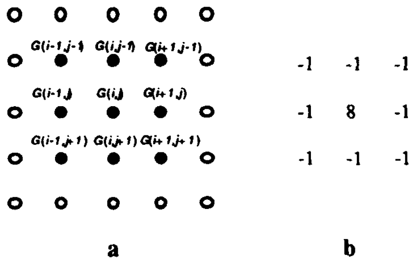

Fig. 1 a. pixel positions, **b.** the convolution mask applied

The texture T_x for cell x can be obtained as follows,

$$T_x = \frac{\sum\limits_{(i,j)\in \text{cell } x} (O(i,j))^2}{Area}$$

$$= \frac{\sum\limits_{(i,j)\in \text{cell } x} (8G(i,j) - \sum\limits_{(k,l)\in \eta} G(k,l))^2}{Area} \qquad (8)$$

$O(i,j)$ is the convolution output at location (i,j) on a cell image. $G(i,j)$ corresponds to the intensity value at location (i,j). η represents the 8 neighboring locations of (i,j).

As only texture information of the cell surface is of the interests, the high frequency information on the boundary between cell and background (zero pixel value) should be avoided. Above equation can be modified to

$$T_x = \frac{\sum\limits_{(i,j)\in \text{cell } x} O(i,j)^2}{Area}$$

$$= \frac{\sum\limits_{(i,j)\in \text{cell } x} (S\times G(i,j) - \sum\limits_{(k,l)\in \eta'} G(k,l))^2}{Area} \qquad (9)$$

η' represents the non-zero neighborhood region of (i,j), and S is the number of non-zero pixels. This effectively avoids the boundary problem by modifying the mask in those regions.

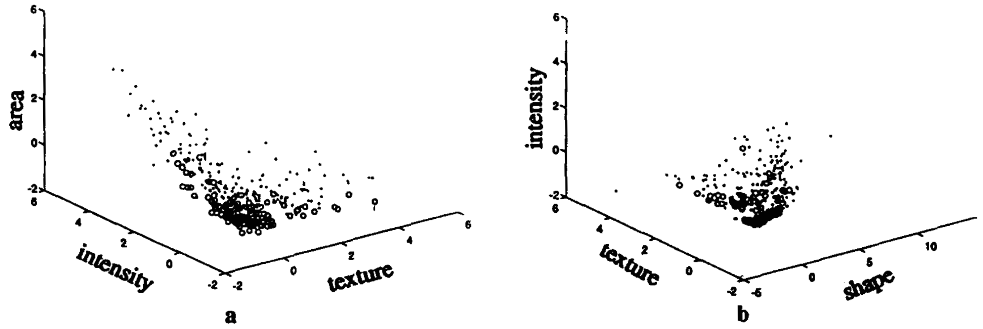

Fig. 2 Training data vectors plot in feature space
a. area-intensity-texture, b. intensity-texture-shape.

4. Experiments

All together 467 cell images from 6 cases are obtained. All cells have been labelled manually by experts as cancer or non-cancer. Among them 263 from two cases are abnormal (cancer) cells and 204 from the rest of four cases are normal cells. In order to restore the aspect ratio to 1:1, the original images are expanded horizontally by a factor of 1.78. After aspect ratio correction, the cell images are tailored into 60×60 pixel images. The cell image is centrally aligned in this area and used as input of network.

Four features, namely area, average intensity, texture and shape factor defined in previous section are extracted from images cell by cell. Based on these a feature vector list is formed. Apart from the raw images, also provided is additional non-perceptual feature information pgDNA. This supplementary information will be shown to be useful in enhancing the performance of classification.

As is often the case in pattern recognition certain features have substantially larger numerical values than others. To prevent those features from dominating the training process, all features are normalized by their corresponding standard deviations. Some cells, even after normalization, still have feature values out of transition region [-5,5], beyond which is a saturated region in a sigmoid activation function. To keep the feature value within the transition region, a scaling factor of 8 is identified and used in sigmoid activation function of Eq. (1). Similarly, a value 2.5 is also chosen as σ for gaussian activation function.

Shown in Fig. 2 is the input training data vectors in feature space. Non-cancer cells are symbolized by 'o' and cancer cells by '•'. One can see that non-cancer cells cluster tightly around origin of the feature space while cancer cells spread out along each axis. This spreading out manifests the irregular nature of cancer cells in shape, cell surface smoothness, intensity and size.

All input data are split into two groups for training and testing. Roughly 80% of the cells (normal and abnormal) are used for training. The rest is retained for testing.

Experiments with all four perceptual features were conducted under various conditions (i.e. hidden neuron number, random seeds). For the optimal number of hidden neurons 4, the classification score reaches 94.6% for sigmoid network, and 86.7% and 91.8% for sinusoid and gaussian networks respectively. Here, it is worth to point out that by using texture feature alone, with sigmoid network, a classification rate of 88% can be achieved. This demonstrates the rich information contained in cell surface texture. Table 1 reports the test results for three networks with four feature input on a 98 test cell set. Values listed in Table 1 for each experiment are 1) classification rate, 2) iterations taken for the network to converge, and 3) total square error at the output layer for all test patterns.

When pgDNA is used in combination with the four perceptual features to train the network, there is a noticeable improvement for sigmoid network, from previous 94.6% to 96.9%. Sinusoid and gaussian networks archive 90.8 and 95.9 respectively. This indicates pgDNA does provide additional discriminate information. Table 1 shows the experiment results.

In the fuzzy c-means approach, the fuzziness parameter m in Eqs.(4), (5) and (6) is chosen to be 2. The same partition of data as in NN approach is used. The initial centroids are randomly chosen among training patterns. It takes typically fewer than 20 iterations for either fuzzy or non-fuzzy clustering to converge. For non-fuzzy approach, the classification rates are 65.5% and 75.5% for four feature input and

Table 1. Performance of networks for various input, activation functions.

Input Data	Networks	Classify Rate	Learning Speed	Output Error
4 Features	Sigmoid	94.6%	8,500	5.58
	Sinusoid	86.7%	200	10.00
	Gaussian	91.8%	1,000	7.09
5 Features	Sigmoid	96.9%	10,000	2.98
	Sinusoid	90.8%	200	9.41
	Gaussian	95.9%	3,000	3.58

Table 2. Performances of classification through unsupervised learning

Input Data	Clustering	Classify Rate	Iterations
4 Features	non-fuzzy	65.5%	4
	fuzzy	70.4%	5
5 Features	non-fuzzy	75.5%	9
	fuzzy	76.5%	11

five feature input respectively. When fuzzy approach is adopted, the classification rates are improved to 70.4% for four feature input and 76.5% for five feature input. Table 2 lists all the experimental results.

To study the issue of local optima, different random seeds are used to assigned initial centroids. It is found, for the same configuration, clustering converges to the same partition. Therefore the local optima may not exist with our data and clustering configuration. Experience suggests, with few centroids, e.g. two, the clustering is unlikely to converge to a local optima.

It is clear clustering techniques do not perform as well as NN approach. This can be explained by the fact that NN has an inherent supervised learning mechanism, which is capable of forming an arbitrary discriminate hyper-surface. In case of clustering approach, it is an unsupervised learning process. The use of optimization function Eq. (4) assumes the data pattern scatters around each centroid in a hyper-spherical shape, which is not true for the bladder cancer cells.

5. Discussions and Conclusion

Above described investigations present the results obtained by using different artificial neural networks and c-means clustering techniques for bladder cancer cell classification. Comparison between networks with sigmoid, sinusoid and gaussian activation functions shows sigmoid network is able to achieve higher classification rate, but sinusoid and gaussian networks converge faster. It is suggested this quicker convergence and lower classification rate are possibly due to the fact that sinusoid and gaussian network have higher degree of non-linearity. This allows the network to quickly fit the *given training data*. But also because of this non-linearity, the generalization ability suffers. Results indicate that our feature extraction does generate a set of simple, nevertheless, effective input features.

Compared with NN approach, clustering techniques reach a much lower classification rate. This is mainly due to the fact that c-means clustering assumes a hyper-spherical data distribution, therefore the discriminate surface is constrained by this geometrical shape. Although fuzzy version of c-means clustering does improve the classification to some degree, neural network is still by far a much better choice for our data.

References

[1] K. Ashenayi, J. Vogh and M. R. Sayeh, 1992, "Gaussian Perceptron Capable of Classifying "2N+1" Distinct Classes of Input Patterns," *IASTED Journal of Control and Computers*, Vol. 20, No. 2, pp. 54-60.

[2] K. Ashenayi, J. Vogh, M. R. Sayeh, B. Karimi and T. Baradaaran, 1992, "Multiple Threshold Perceptron Using Sinusoidal Function," *IASTED Journal of Modelling and Simulation*, Vol. 12, No.1, pp.22-26.

[3] J. C. Dunn, 1973, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *J. Cybernetics*, Vol. 3, pp.32-57.

[4] K. Hornick, M. Stinchcombe and H. White, 1989, "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks*, Vol. 2, No. 5, pp.359-366.

[5] Jain A K, 1986, 'Cluster analysis', in Handbook of Pattern Recognition and Image Processing, Ed. by Young Z Y and Fu K S.

[6] L. G. Koss (Ed), 1979, "Diagnostic cytology and its histologic bases," Vol.2, pp.767.

[7] Y. Noguchi, Y. Tenjin and T. Sugishita, 1983, "Cancer-cell detection system based on multispectral images," Anal. Quant. Cytol. Vol.5, pp.143-151.

[8] T. Reed and J. Hans Du Buf, 1993, "A Review of Recent Texture Segmentation and Feature Extraction Techniques," *CVGIP:Image Understanding*, Vol.57, pp.359-372.

[9] L. A. Zadeh, 1965, "Fuzzy Set," *Inform. Control*, Vol. 8, pp.338-353.

[10] G. Zajicek, M. Shohat, Y. Melnick and A. Yegeuz, 1983, "Image analysis of nucleated red blood cells," Comp. Biomed. Res. Vol.16, pp.347-356.