MOMENTS KULLANARAK GÖĞÜS KANSERİ SINIFLANDIRMA
BREAST CANCER CLASSIFICATION USING MOMENTS

Alaa Eleyan

Electrical & Electronic Engineering, Mevlana University
Yeni Istanbul Cad. No 235 42003, Selçuklu, Konya, TURKEY
(Video, Image & Speech Processing Laboratory)
aeleyan@mevlana.edu.tr

ÖZETÇE

Bu makale göğüs kanserinin belirlenmesi için, moment tabanlı bir sistem önermektedir. Bu çalışmada, problemin çözümü için uygulanmış olan sınıflandırıcıyı geliştirmek yerine, giriş özniteliklerinin iyileştirilmesi amaçlanmıştır. Veri setindeki örnekler üzerinden, ortalama, değişinti, çarpıklık ve kurtosis gibi dört moment kullanılarak, yeni öznitelikler elde edildi. 10-fold yöntemi ile farklı denemeler yapılarak, Wisconsin göğüs kanser veri seti üzerinde sınıflandırıcı performansı, önerilen öznitelikler kullanılarak ölçüldü. Sonuçlar, kullanılan tüm sınıflandırıcılar için önerilen özniteliklerin, sınıflandırma başarımlarını artırmada etkin olduğun göstermiştir.

ABSTRACT

This paper presents a system for detecting breast cancer based on moments. Instead of trying to improve the applied classifier we focused on improving the input attributes. We extracted new features from database samples using the first four moments namely, mean, variance, skewness and kurtosis. Through simulations, 10-fold cross validation method was applied to the Wisconsin breast cancer database to evaluate the classification performances. Various classifiers were used for evaluating the proposed approach. Results indicate advantage of such features in improving classification performance for all of the applied classifiers.

Keywords: breast cancer, moments, neural networks, support vector machines, Bayes classifier, classification.

1. INTRODUCTION

Tumors are anomalous growths in human body. Typically, old cells die, and new ones take their place. Yet, sometimes, this process goes wrong. New cells form when we don't need them, and old cells don't die when they should. These extra cells form a mass called a tumor. Tumors fail into two categories, namely, benign or malignant. Malignant tumors are cancer and benign tumors are not.

The accurate classification of breast cancer is an essential and fatal problem in the medical community. The problem of breast cancer detection attracted many researchers in the last decay. Researchers managed to achieve higher classifications results [1][2][8]. In [3] they used artificial neural networks (ANNs) with C4.5 decision tree method. In Abonyi and Szeifert [4] they applied supervised fuzzy clustering approach. Ubeyli [5] used various classifier for detection of breast cancer namely, support vector machine, probabilistic neural networks, recurrent neural networks, combined neural networks and multilayer perceptron neural networks. In [6] least square SVM was used. Association Rule (AR) was used in [7] as automatic diagnostic system. Recently artificial metaplasticity neural network was used by Cedeno et al. [8] for the WBCD database classification.

In this paper, at first stage, we extract a new feature vector from the original data attributes. At second stage, classification performance of these new vectors (attributes) using various classifiers was performed and compared with classification performance using the original vectors (attributes). The new attributes helped to improve the classification performance regardless of the used classifier.

This paper is organized as follows: section 2 gives an overview about the Wisconsin breast cancer database. Section 3 lists and explains the classifiers used in this paper. Section 4 introduces the proposed feature extraction approach. Simulation results and discussions are reported in Section 5. Finally, conclusions are provided in Section 6.

2. WISCONSIN BREAST CANCER DATABASE

A malignant tumor is a tumor that grows from the breast cells forming a breast cancer. Many factors are increasing the chance of having a breast cancer such as aging, gender, race, family history, genetics or personal behaviors (smoking, drinking, and diet). But risk factors don't tell us everything. Having a risk factor, or even several, does not mean that the person will get the disease. Scientists still don't know how some of these factors cause cells to grow cancerous.

In this paper, the Wisconsin breast cancer database (WBCD) was used [9]. The database consists of 699 samples taken from Fine Needle Aspirates (FNA) of human breast tissues. Each sample consists of nine attributes shown in table 1. Each attribute assigned an integer value between 1 and 10 with 10 being the closet

978-1-4673-0056-8/12/$26.00 ©2012 IEEE

to malignant. Database contains 16 samples with missing attributes. Many research papers discarded these samples during evaluation of their classification algorithms. For sake of fairness, these 16 samples are discarded and just 683 samples were used during this study.

Table 1. Wisconsin breast cancer database description (683 samples; 239 malignant and 444 benign)

#	Attribute	Range	Mean	Standard Deviation
1	Clump Thickness	1-10	4.4422	2.8208
2	Uniformity of Cell Size	1-10	3.1508	3.0651
3	Uniformity of Cell Shape	1-10	3.2152	2.9886
4	Marginal Adhesion	1-10	2.8302	2.8646
5	Single Epithelial Cell Size	1-10	3.2343	2.2231
6	Bare Nuclei	1-10	3.5447	3.6439
7	Bland Chromatin	1-10	3.4451	2.4497
8	Normal Nucleoli	1-10	2.8697	3.0527
9	Mitoses	1-10	1.6032	1.7327

3. CLASSIFIERS

3.1. Nearest Neighbor

Classifying query samples based on their distance to samples in a training set can be a simple yet effective way of categorizing new points. A variety of metrics can be used to find the distance. Following is some the distance metrics that can be used

- Euclidean distance

$$D_1 = d(\mathbf{p}, \mathbf{q}) = \sqrt{(\mathbf{p} - \mathbf{q}).(\mathbf{p} - \mathbf{q})} \quad (1)$$

where $\mathbf{p} = (p_1, p_2, \dots, p_n)$ and $\mathbf{q} = (q_1, q_2, \dots, q_n)$.

- Mahalanobis distance

$$D_2 = d(\mathbf{p}, \mathbf{q}) = \sqrt{(\mathbf{p} - \mathbf{q})^T \mathbf{C}^{-1} (\mathbf{p} - \mathbf{q})} \quad (2)$$

where **C** is the covariance matrix. If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance.

- City block metric

$$D_3 = d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i| \quad (3)$$

- Minkowski metric

$$D_4 = d(\mathbf{p}, \mathbf{q}) = \left(\sum_{i=1}^{n} |p_i - q_i|^p\right)^{1/p} \quad (4)$$

Notice that for $p = 1$, the Minkowski metric gives the city block metric, for $p = 2$, the Minkowski metric gives the Euclidean distance and $p = \infty$, the Minkowski metric gives the Chebychev distance.

- Chebychev distance

$$D_5 = d(\mathbf{p}, \mathbf{q}) = \max_i (|p_i - q_i|) \quad (5)$$

- Cosine distance

$$D_6 = d(\mathbf{p}, \mathbf{q}) = -\frac{\mathbf{p}.\mathbf{q}}{\|\mathbf{p}\|\|\mathbf{q}\|} \quad (6)$$

3.2. Naïve Bayes Classifier

For pattern recognition, the Bayes classifier is the best classifier, the Bayes error the best criterion to evaluate feature sets, and a posteriori probability functions are thus optimal features [10,11]. Let $\omega_1, \omega_2 \dots \omega_L$ denote the object classes, and X a vector contain sample attributes. The *a posteriori* probability function of ω_i given X is defined as

$$P(\omega_i/X) = \frac{p(X/\omega_i)P(\omega_i)}{p(X)} \quad (7)$$

where $P(\omega_i)$ is *a priori* probability, $p(X/\omega_i)$ the conditional probability density function of ω_i, and $p(X)$is the mixture density. The Maximum *A Posteriori* (MAP) decision rule for the Bayes classifier is

$$p(X/\omega_i)P(\omega_i) = \max_j \left(p(X/\omega_j)P(\omega_j)\right), X \in \omega_i \quad (8)$$

For WBCD database the sample X is classified either to ω_i (benign or malignant) of whom the *A Posteriori* probability given X is the largest between the two classes.

3.3. Neural Networks

Neural networks are biologically inspired and imitate human brain. Feed forward neural networks are a basic type of NN capable of approximating generic classes of functions [8]. They are the most commonly used NN architectures [12, 13]. Multilayer Perceptron MLP is a famous class of FFNN which used Backpropagation training algorithm. In the neurons are organized in the form of layers. The neurons in a layer get input from the previous layer and feed their output to the next layer. In this type of networks connections to the neurons in the same or previous layers are not permitted. Neurons in different layers are linked. These connection links have weights which need to be adjusted iteratively depending on the input signals till the desired output, MSE error value or epoch number is reached. NN are trained by examples, when unknown signal applied, it generalize the past examples and produce an output (14].

3.4. Support Vector Machines

The foundations of Support Vector Machines (SVM) have been developed by Vapnik [15] and since then it are gained popularity due to many attractive features, and promising empirical performance. Training SVM is a quadratic optimization problem. SVM construct the decision surface in the higher dimensional space mapping the input signals into that space using nonlinear mapping.

For two-class problem, assuming optimal hyperplane in higher dimensional space is generated, the classification decision of an unknown signal X will be made based on kernel function. The kernel function forces the operations to be carried out in the input space rather than in the higher dimensional space. Choosing proper kernel function is dependent on the type of the problem and the given data. According to [5] optimal results for SVM were achieved using RBF kernel function.

4. PROPOSED APPROACH

WBCD Database contains 9-attributes for each one of the 683 samples. The idea here is to extract further information out of these attributes. We proposed in this work to extract four extra attributes for each sample, namely, mean, variance, skewness and kurtosis. These new attributes will be used as a feature vector instead of the original sample attributes and/or concatenated with the original attributes to form new feature vectors.

Mean is the first row moment which is the arithmetic average of a set of values. The mean of a set of values X=[x_1, x_2, ..., x_n] is typically denoted by μ and is given as:

$$\mu = E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{9}$$

where E is the expectation operator.

The second central moment about the mean is the variance and it is given as:

$$\sigma^2 = E[(X-\mu)^2] \tag{10}$$

The third central moment, called skewness, is a measure of the lopsidedness of the distribution. The skewness of a random variable X denoted γ_1 is defined as

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} \tag{11}$$

where μ_3 is the third moment about the mean μ, σ is the standard deviation.

The fourth central moment, called kurtosis, is a measure of whether the distribution is tall and skinny or short and squat, compared to the normal distribution of the same variance. Since it is the expectation of a fourth power, the fourth central moment, where defined, is always non-negative; and except for a point distribution, it is always strictly positive.

$$\gamma_2 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} \tag{12}$$

Using these moments, 4 attributes will be generated for each of the 9-attribute samples (μ, σ^2, γ_1 and γ_2). These attributes will be used as input for the various classifiers mentioned in previous section. Also these attributes can be concatenated to the other 9 attributes to form another feature vector which can be applied to the classifiers.

5. SIMULATION RESULTS & DISCUSSION

The classifiers proposed in section 3 were implemented using MATLAB ver. 7.11 with neural networks, statistics and bioinformatics toolboxes. In WBCD database there are 683 samples, of which 239 are benign samples and 444 are malignant samples. For training phase, 273 of 683 samples were used. The rest were used for testing. The training set consists of 178 benign samples and 95 malignant samples. For reliability of the results, 10-fold cross validation approach was utilized and results were averaged. The performance comparison and correct classification rates are tabulated in Table 2.

For BPNN, the number of input neurons equal to the number of attributes used as input (4, 9 or 13). Number of hidden layers is 1 with 8 hidden neurons. Output neurons are 1 and activation function is sigmoid. Training will stop if 1000 epochs or mean square error, MSE=0.001 reached.

As shown in table 2, using the extracted statistical features from each of the 9-attribute samples to form a new feature vector is a good choice. The performance results shows that, independent of the applied classification algorithm, the new feature vectors with 4 attributes (mean, variance, skewness, kurtosis) performs even better than the original 9-attribute vectors. For example in BPNN the performance increases by nearly 2% when using just the moment attributes as input. Another example is the 5% increase with Cos distance. However, concatenating the new attributes with the original ones didn't help to improve the performance.

Table 2. Performance classification for WBCD breast cancer database using various algorithms and number of attributes.

Algorithm		Original 9 Attributes	Moments 4 Attributes	Combined 13 Attributes
Nearest neighbor	D1	95.7805	95.9268	93.8537
	D2	93.3504	96.3668	96.1324
	D3	95.8780	96.0488	94.6341
	D4	95.0244	96.0000	93.1707
	D5	93.4390	95.8293	92.9512
	D6	88.8049	93.9024	93.4146
Naïve Bayes		95.4146	95.6098	95.5366
BPNN		93.9268	96.0976	95.8537
SVM		96.2927	96.5610	96.5610

6. CONCLUSIONS

In this paper we proposed a new approach for improving the classification performance for WBCD breast cancer database. We extracted new attributes from the available 9 attributes for each sample. Four attributes namely, mean, variance, skewness and kurtosis were extracted for each sample and a new feature vectors were generated. These vectors were used as new descriptors for the samples and applied to various classifiers. Also, these vectors were combined with their original respective vectors to form 13-attribute vectors. Simulation results indicate that this approach helped to improve the performance of the classification for most of the classifiers with improvement reached up to 5% for some classifiers.

7. REFERENCES

[1] Choua, S.-M., Leeb, T.-S., Shaoc, Y. E., and Chenb, I.-F. (2004). Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications, 27, 133–142.

[2] Ryua, Y. U., Chandrasekaranb, R., and Jacobc, V. S., (2007). Breast cancer prediction using the isotonic separation technique. European Journal of Operational Research, 181, 842–854.

[3] Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research, 4, 77909.

[4] Abonyi, J. Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers, Pattern Recognition Letters 14 (24) 2195–2207.

[5] Übeyli, E. D. (2007). Implementing automated diagnostic systems for breast cancer detection. Expert Systems with Applications, 33(4), 10541062.

[6] Polat, K., and Günes, S. (2007). Breast cancer diagnosis using least square support vector machine. Digital Signal Processing, 17(4), 694–701.

[7] Karabatak, M. Cevdet-Ince, M. (2009), An expert system for detection of breast cancer based on association rules and neural network, Expert Systems with Applications 36, 3465–3469..

[8] Marcano-Cedeño A., Quintanilla-Domínguez J., Andina D., (2011), WBCD breast cancer database classification applying artificial metaplasticity neural network, Expert Systems with Applications, 38(8) , 9573-9579

[9] Mangasarian, O. L. and Wolberg, W. H.: (1990), Cancer diagnosis via linear programming, SIAM News, 23(5), 1-18.

[10] Fukunaga, K. Introduction to Statistical Pattern Recognition. Academic Press, second edition, 1991.

[11] Chengjun, L; Wechsler, H.(1998) Probabilistic reasoning models for face recognition," IEEE Conference on Computer Vision and Pattern Recognition Proceedings, 827-832

[12] Haykin, S. (2008). Neural networks and learning machines, New York: Prentice Hall, 3rd ed., ISBN: 978-0131471399

[13] Basheer, I. A., and Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods, 43(1), 3–31.

[14] Bishop, C. M. (1996). Neural networks for pattern recognition. Oxford: Clarendon Press.

[15] Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer, NY. ISBN 0-387-94559-8.