Classification of Bio Optical signals using K- Means Clustering for Detection of Skin Pathology

G.Subramanya Nayak
Department of E &C Engineering,
Manipal Institute of Technology
Manipal University
Manipal -576104

Ottolina Davide
Biomedical Engineering Department
Politechnico Di Milano
Piazza Leonardo Davinci
Milano, ITALY

Puttamadappa C
Department of E &C Engineering
SJB Institute of Technology
Uttarahalli Road, Kengeri
Bangalore -60

ABSTRACT

Early diagnosis of precancerous and malignant lesions is critical for the improving of the current poor survival rate of patients with a variety of tumors. The development of new high-specificity and high-sensitivity imaging technologies can play an important role in the early diagnosis, accurate staging, and treatment of cancer. Bio-optical signals are the result of the optical functions of the biological systems, occurring naturally or induced by the measurement. The identification of the state of human skin tissues is discussed here. The Bio-optical signals recorded in vitro have been analyzed by extracting various statistical features. Using MATLAB programs, various statistical features are extracted from both normal and pathology spectra. Different features like mean, summation, skewness, etc were extracted. The values of the feature vector reveal information regarding tissue state. These parameters have been analyzed for the discrimination between normal and pathology conditions. For analysis, a specific data set has been considered. Using K-means clustering, signal classification was done in MATLAB. Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters. This is done such that patterns in the same cluster are alike and patterns belonging to two different clusters are different. Clustering has been a widely studied problem in a variety of application domains including neural networks, statistics etc.

Keywords

Statistical Analysis, K-Means Clustering

1. INTRODUCTION

The cancer is one of the leading causes of death all over the world, but, if detected early, can be curable. In the current study, the data analysis and classification of pathological conditions of the optical spectra of skin cancer are performed using MATLAB programs. The proposed study is done on MITI-FOPTO database. A large amount of data needs to be analyzed for the classification of normal and pathology conditions. Using MATLAB functions, statistical features (median, variance, summation, skewness, etc.) are extracted for different spectra. Then, the best ones are employed for the classification of spectra in normal/pathological clusters, using k-means algorithm.

Before feature extraction, the signals are normalized and then filtered to eliminate undesirable spikes due to noise and other disturbances.

The main schematic model consists of four modules as shown in Figure 1.

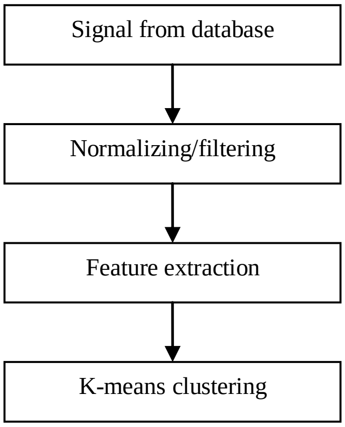

Figure 1. Schematic model

This paper deals with the data analysis and classification of pathological conditions of the optical spectra of skin cancer performed using MATLAB@ 7.1.

2. FEATURE EXTRACTION

The statistical analysis and classification for discrimination among normal and malignant conditions were performed using MATLAB on the set of spectral data. Using MATLAB, the following features were extracted: arithmetic mean, median, variance, standard deviation, RMS (Root Mean Square), Summation, Skewness, Kurtosis. A brief description of each of the above features is given below.

3.1 Arithmetic Mean

In simple terms, the arithmetic mean of a list of numbers is the sum of all the members of the list divided by the number of items in the list. If a set of data is denoted by $X = (x_1, x_2, ..., x_n)$, then the sample mean is typically denoted with a horizontal bar over the variable

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n)$$

3.2 Median

A median is described as the number separating the higher half of a sample, a population, or a probability distribution, from the lower half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one. If there is an even number of observations, the median is not unique, so one often takes the mean of the two middle values. At most half the population has values less than the median and at most half have values greater than the median.

3.3 Variance

The variance of a sample is one measure of statistical dispersion, averaging the squared distance of its possible values from the expected value (mean). Whereas the mean is a way to describe the location of a distribution, the variance is a way to capture its scale or degree of being spread out. The unit of variance is the square of the unit of the original variable. The positive square root of the variance, called the standard deviation, has the same units as the original variable and can be easier to interpret for this reason. If random variable X has expected value (mean) μ = E(X), then the variance Var(X) of X is given by:

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mu)^2]$$

3.4 Standard deviation

The standard deviation of a multiple set of values is a measure of statistical dispersion of its values. The standard deviation is usually denoted with the letter σ. It is defined as the square root of the variance.

$$\sigma = \sqrt{\mathrm{E}((X-\mathrm{E}(X))^2)} = \sqrt{\mathrm{E}(X^2) - (\mathrm{E}(X))^2}$$

where E(*X*) is the expected value of *X*. Standard deviation, being the square root of variance, measures the spread of data about the mean, measured in the same units as the data.

3.5 RMS

The root mean square is a statistical measure of the magnitude of a varying quantity. It is especially useful when variants are positive and negative, e.g. sinusoids. It can be calculated for a series of discrete values or for a continuously varying function. The RMS of a collection of n values is

$$x_{\mathrm{rms}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}}$$

The RMS of a periodic function is equal to the RMS of one period of the function. The RMS value of a continuous function or signal can be approximated by taking the RMS of a series of equally spaced samples.

3.6 Summation

Summation is the addition of a set of numbers; the result is their sum or total. The "numbers" to be summed may be natural numbers, complex numbers, matrices, or signals. An infinite sum is a subtle procedure known as a series. Summation can also be represented as

$$\sum_{i=m}^{n} x_i = x_m + x_{m+1} + x_{m+2} + \ldots + x_{n-1} + x_n.$$

3.7 Skewness

It is a measure of the asymmetry of the probability distribution ofa real-valued random variable.

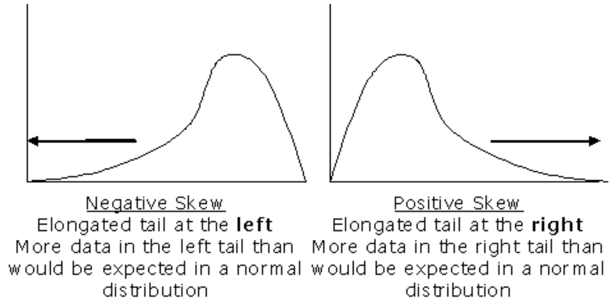

Fig. 2.1 - Positive and Negative skews

Considering the distribution in the above figure, the right side of the distribution tapers differently than the left one. These tapering sides are called tails, and they provide a visual means for determining which of the two kinds of skewness a distribution has:

1. **Negative skew** (he left tail is longer): the mass of the distribution is concentrated on the right of the figure. The distribution is said to be left-skewed.
2. **Positive skew** (the right tail is longer): the mass of the distribution is concentrated on the left of the figure. The distribution is said to be right-skewed.

Skewness, the third standard moment is written as γ_1 and defined as

$$\gamma_1 = \frac{\mu_3}{\sigma^3},$$

where μ_3 is the third moment about the mean and σ is the standard deviation. Equivalently, skewness can be defined as the ratio of the third cumulant κ_3 and the third power of the square root of the second cumulant κ_2:

$$\gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}}.$$

This is analogous to the definition of kurtosis, which is expressed as the fourth cumulant divided by the fourth power of the square root of the second cumulant.

For a sample of n values the skewness is

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}},$$

where x_i is the i^{th} value, $\overline{x}$is the sample mean, m_3 is the sample third central moment, and m_2 is the sample variance..

3.8 Kurtosis

The fourth standardized moment is defined as

$$\frac{\mu_4}{\sigma^4},$$

where μ_4 is the fourth moment about the mean and σ is the standard deviation. This is sometimes used as the definition of kurtosis in older works, but is not the definition used here.

Kurtosis is more commonly defined as the fourth cumulant divided by the square of the second cumulant, which is equal to the fourth moment around the mean divided by the square of the variance of the probability distribution minus 3,

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3,$$

which is known as “excess kurtosis”. The "minus 3" at the end of this formula is often explained as a correction to make the kurtosis of the normal distribution equal to zero.

For a sample of n values the sample kurtosis is

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3$$

where m_4 is the fourth sample moment about the mean, m_2 is the second sample moment about the mean (that is, the sample variance), xi is the i^{th} value, and $\overline{x}$ is the sample mean.

3. CLASSIFICATION

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) shares some common trait - proximity according to some defined distance measure.

We can distinguish non-fuzzy or hard clustering, where data is divided into crisp clusters, and each data point belongs to exactly one cluster, from fuzzy clustering, where the data points can belong to more than one cluster, and associated with each of the points are membership grades which indicate the degree to which the data points belong to the different clusters.

Clustering has been a widely studied problem in a variety of application domains including neural networks, AI, and statistics. The k-means method (a non-fuzzy clustering technique) has been shown to be effective in producing good clustering results for many practical applications. However, a direct algorithm of k-means method requires time proportional to the product of number of patterns and number of clusters per iteration. This can be computationally very expensive, especially for large data sets.

4.1 K-means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early classification is done. At this point it is needed to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. Then, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop it is possible to notice that the k centroids change their location step by step until no more changes are done. In other words they do not move any more. Finally, this algorithm aims at minimizing an objective function, like total intra-cluster variance, or, the squared error function

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

where there are k clusters S_i, $i = 1, 2, ..., k$, and μ_i is the centroid or mean point of all the points $x_j \in S_i$.

More schematically, the algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

In MATLAB language, there is an apposite function, called [IDX,C] = K Means (X,k), which is used to calculate the results of k-means clustering. It partitions the points in the N-by-P data matrix X into k clusters and returns the k cluster centroid locations in the k-by-p matrix C.

This partition minimizes the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances. Rows of X correspond to points, columns correspond to variables. K-means returns an N-by-1 vector IDX containing the cluster indices of each point. By default, MATLAB K-means uses squared Euclidean distances. Although it can be proved that the procedure will always terminate, the K-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-means algorithm can be run multiple times to reduce this effect.

The numerical values of the features extracted are used for classification of normal and pathological signals. The classification process is carried out in MATLAB, too.Clustering is done such that patterns in the same cluster are alike and patterns belonging to different clusters are different. The values obtained from the feature extraction are loaded in the MATLAB code and k-means clustering is performed.

4. RESULTS AND DISCUSSION

The numerical values of the features extracted from 26 spectra and results are tabulated in Table 5.1 and 5.2 respectively.

Table 5.1 Features of the Normal spectra

	N1	N2	N3	N4	N5	N26
Mean	0,4436	0,4635	0,4561	0,4629	0,4845	0,4613
Median	0,4431	0,4755	0,4168	0,4331	0,4143	0,4306
Variance	0,0825	0,0777	0,0689	0,0745	0,0722	0,0792
Std	0,2872	0,2788	0,2624	0,273	0,2688	0,2887
RMS	0,5284	0,5408	0,5261	0,5373	0,554	0,5689
Summation	451,144	471,338	463,824	470,761	492,734	464,341
Skewness	0,2509	0,1513	0,5098	0,3259	0,4118	0,4799
Kurtosis	1,8568	1,8007	2,1022	1,9033	1,7336	1,9602

Table 5.2 Features of the pathology spectra

	P1	P2	P3	P4	P5	P26
Mean	0,4057	0,4264	0,4132	0,3975	0,4093	0,4075
Median	0,338	0,328	0,3534	0,3137	0,3418	0,327
Variance	0,0868	0,1038	0,0857	0,0909	0,0843	0,0979
Std	0,2945	0,3222	0,2927	0,3014	0,2904	0,3065
RMS	0,5013	0,5343	0,5063	0,4988	0,5017	0,5017
Summation	412,6075	433,6014	420,2719	404,2201	416,2344	405,5428
Skewness	0,445	0,4586	0,4145	0,4489	0,4142	0,4597
Kurtosis	1,803	1,746	1,7965	1,7576	1,7816	1,7228

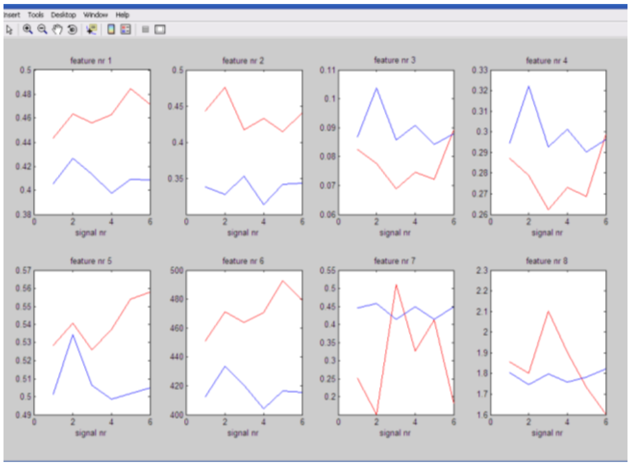

Figure 5.1 – Features comparison: mean, median, variance, Standard deviation, RMS, skewness, kurtosis (normal in red; pathological in blue).

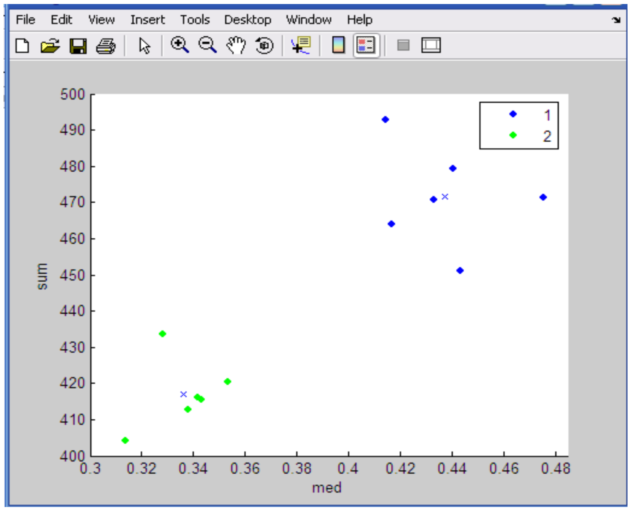

Figure 5.2 - Classification scatter plot: median and summation are given respectively in the x and y axis. Crosses represent cluster centroids positions.

After having obtained the numeric values about every statistical feature, it was possible to find those features that best enhance the differences between normal and pathological spectra. Figure 5.1 graphically shows that median and summation are two of the best ones.

Therefore, only two of the eight features extracted were used in K-Means classification, in order to plot a 2D image that shows the distribution of the data points in the different clusters as shown in Figure 5.2.

6. CONCLUSION

Although, because of the lack of a great number of available signals, performance parameters should be evaluated more accurately carrying out tests on a larger amount of tissue samples, the method reported in the current study achieves the discrimination between normal and pathological tissues. Moreover, a small time needed to acquire and analyze the optical spectra together, with high rates of success, proves the method attractive for real time applications. The results of feature extraction and classification have been verified. The value of the feature vector revealed considerable information regarding the tissue state. Hence newer methods like the one described can be implemented for the discrimination of the normal and pathological tissues.

7. ACKNOWLEDGMENTS

Our thanks to Prof.T.G.S.Chandrashekarappa for his valuable suggestions and guidance.

8. REFERENCES

[1] G. Subramanya Nayak, C. Puttamadappa, Akshata Kamath, B. Raja Sudeep, K. Kavitha, 2008"Classification of Bio-Optical Signals using Soft Computing Tools," snpd,pp.661-663.

[2] Cancer Research UK ,January 2007. "UK cancer incidence statistics by age". Retrieved on 2007-06-25.

[3] WHO February 2006. "Cancer". World Health Organization. Retrieved on 2007-06-25.

[4] American Cancer Society ,December 2007. "Report sees 7.6 million global 2007 cancer deaths". Reuters. Retrieved on 2008-08-07.

[5] SV Deo, Sidhartha Hazarika, Nootan K Shukla, Sunil Kumar, Madhabananda Kar, Atul Samaiya , “Surgical management of skin cancers: Experience from a regional cancer centre in North India”, Indian Journal of Cancer 2005, vol.42, issue 3

[6] WHO World health statistics. GLOBOCAN 2000: Cancer Incidence, Mortality and Prevalence Worldwide. Version 1.0. IARC CancerBase No. 5. Lyon, IARC Press; 2001.

[7] Howe HL, Wingo PA, Thun MJ, Ries LAG, Rosenberg HM, Feigal EG, et al . “Annual report to the nation on the status of cancer (1973 through 1998), featuring cancers with recent increasing trends”, J Natl Cancer Inst 2001;93:824-42.

[8] Godbole VK, Toprani HT, Shah HH. “Skin cancer in Saurashtra”. Ind J Pathol Bacteriol 1968;11:183-9. [PUBMED]

[9] National Cancer Registry Programme, Indian Council of Medical Research. Consolidated report of the population based cancer registries1990-96.

[10] Richard Goering, "Matlab edges closer to electronic design automation world," EE Times, 10/04/2004

[11] Cleve Moler, the creator of MATLAB (December 2004). "The Origins of MATLAB". Retrieved on April 15, 2007.

[12]http://www.mathworks.com/access/helpdesk/help/techdoc/matlab.shtml

[13]http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

[14] http://fconyx.ncifcrf.gov/lukeb/kmeans.html