Analysis of the ECG data based on a Gaussian mixture model

Xiaoyu Zhan *

Department of Management Science and Engineering, University of Jinan, Shandong, China

* Corresponding Author Email: 577967599@qq.com

Abstract. With the development of The Times, heart disease has become the first killer of human health, and the initial manifestation of heart disease usually is arrhythmia, so it is particularly important to conduct real-time monitoring and analysis of heart rhythm. In this paper, the data analysis of electrocardiogram. First, we preprocessed the heart rate data to improve the accuracy of the subsequent analysis results. Then, the features of the data values were quantified and clustered using a Gaussian mixture model to determine the plausibility of the model. Next, the various characteristic values of the arrhythmia were compared with the normal data, and the arrhythmia was classified. Finally, the entropy weight method is used to empower the risk degree of different indicators, and to evaluate the model to determine the risk level.

Keywords: GMM model, entropy weight method, Topsis, wavelet packet transformation, bivariate threshold function.

1. Introduction

With the development of social economy and the change of human life style, about 17.5 million people worldwide die from heart disease every year, accounting for 30 percent of the total deaths. China has become the largest death of heart disease in the world. The initial symptoms of heart patients are usually manifested as arrhythmia, and each heart rate abnormality category performs differently in the ECG and ECG frequency maps. Studies of ECG signals can effectively judge the heart rhythm. Therefore, more and more researchers are studying the monitoring of ECG signals. For people with heart disease, the real-time monitoring of ECG is an important means of judging the arrhythmia.

2. data preprocessing

The ECG signal has two characteristics of low amplitude and low frequency, so the signal is prone to a variety of noise, including environmental noise, power frequency interference, myelectric interference, baseline drift, and electrode contact noise [1]. This paper mainly denoizes power frequency interference, frequency interference baseline drift and myoelectric interference.

ECG signal: 1-dimensional signal with noise is indicated as follows:

$$x(m) = s(m) + \sigma e(m) \;\; m = 0,1,2,\cdots,N-1 \tag{1}$$

The x (m) is used to represent the ECG signal containing interference, s (m) is used to represent the useful signal and e (m) is used to represent the noise

The eight-scale decomposition of the ECG signals is shown in Table 1, followed by the detail coefficients cD1~cD7 and the approximation coefficient cA8. As can be seen from Figure 1, the baseline drift is mainly distributed in the cA8 scale, the power frequency interference is distributed in cD1, cD2, cD3., and the myoelectric interference is mainly distributed in the cD1~cD5 scale. Considering the distribution of interference, the ECG signal can be clean.

Table 1. Signal decomposition scale and coefficient range

Decomposition the number of layers	Scale coefficient
cD1	90~180
cD2	45~90
cD3	22.5~45
cD4	11.25~22.5
cD5	5.625~11.25
cD6	2.8125~5.625
cD7	1.40626~2.8125
cD8	0.703125~1.40625
cA8	0~0.703125

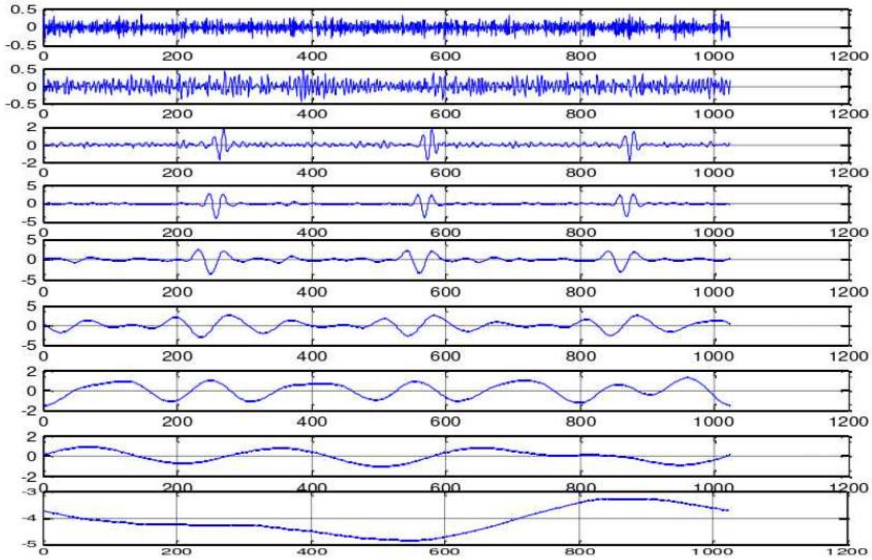
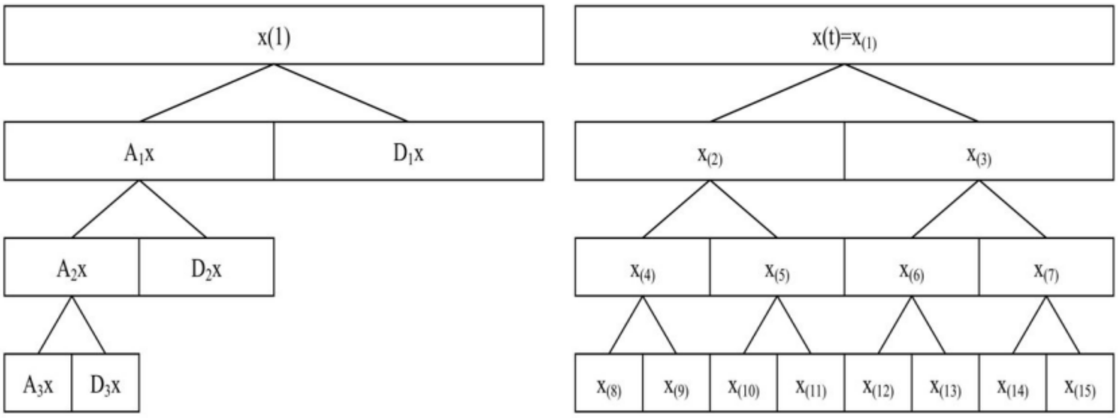


Fig 1. Wavelet coefficient and approximate coefficient of each layer of the ECG signal

Wavelet transform decomposes only the low-frequency band of the signal. The wavelet packet transform inherits the time-frequency analysis characteristics of the wavelet transform, further decomposes the undecomposed high-frequency band signal in the wavelet transform, makes different resolution selections for various frequencies at different levels, and provides a series of sub-band time domain waveforms in the full frequency band range at various scales. Wavelet packet analysis is to further subdivide the wavelet subspace in binary mode to achieve the purpose of improving frequency resolution. The relationship between the wavelet transform and the wavelet packet transform is shown in Figure 2.



a) Wavelet transformation process b) The wavelet packet transformation process

Fig 2. Comparison diagram of wavelet change and wavelet packet transformation

3. Model building and solution

There are many common clustering methods, and after comparative analysis, the Gaussian mixture model can better fit the data. Therefore, the clustering method used in this paper is the GMM algorithm.

3.1. The EM algorithm estimates the GMM parameters

Define the number of components k, set the πk (Gaussian mixture coefficient) μk (mean) for each component k, ∑k (covariance matrix), The log-likelihood function is then calculated according to Equation

$$p(x|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{2}$$

The log-likelihood function is given as follows

$$\ln p(x|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{N} \pi_k N(x_k|\mu_k, \Sigma_k)\} \tag{3}$$

3.2. A Gaussian mixed clustering model was built

GMM (Gaussian mixture Model) is a probabilistic model for clustering, which uses EM algorithm for iterative calculation. Gaussian mixture model assumes that the data of each cluster conform to Gaussian distribution (also called normal distribution), and the distribution presented by the current data is the result of the superposition of Gaussian distribution of each cluster.

With the random variable X, the mixed Gaussian model can be expressed by the following equation:

$$p(\mathrm{x}) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{4}$$

Actually, it can be argued that π_k is the weight of each component $N(x|\mu_k, \Sigma_k)$.

We define as four clusters, that K is 4, then the corresponding GMM form is as follows:

$$p(\mathrm{x}) = \pi_1 N(x|\mu_1, \Sigma_1) + \pi_2 N(x|\mu_2, \Sigma_2) + \pi_3 N(x|\mu_3, \Sigma_3) + \pi_4 N(x|\mu_4, \Sigma_4) \tag{5}$$

There are 12 unknown parameters in the formula, because it is not precise to specify the value of π_k directly, so we use the EM algorithm to automatically determine the value of π_k while iterating on the parameters in the GMM: (π_k, x_k, Σ_k).

3.3. Extraction and clustering results of the time-domain features

Through the data provided by the title, the abnormal data is subjected to wavelet packet transformation to obtain eight sub-bands. Seven time-domain features [7] (maximum, mean, square root amplitude, standard deviation, skewness, waveform, and index) are extracted. The formula is shown in Table 2. The total number of classifications was finally determined to be 4, with clustering using Gaussian mixture models, and the results are shown in Table 3.

Table 2. Statistical characteristics of the time domain

characteristics	**expression**
max	$MAX = \max(X(i))$
mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} X(i)$
Square root magnitude	$FGFZ = (\frac{1}{N}\sum_{i=1}^{N}\sqrt{\lvert X(i)\rvert})^2$
Standard Deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X(i)-\mu)^2}{N}}$

Kurtosis	$KU = \dfrac{\frac{1}{N}\sum_{i=1}^{N}(X(i)-\mu)^4}{rms^4}$
Skewness	$SK = \dfrac{\sum_{i=1}^{N}(X(i)-\mu)^3}{N\sigma^3}$
Shape Indicator	$SH = \dfrac{rms}{\frac{1}{N}\sum_{n=1}^{N}\lvert X(i)\rvert}$

Table 3. Cluster results of the Gaussian mixture model

category	The number of category files
1	211
2	255
3	133
4	270

3.4. Analysis of data results

The normal rhythm data and seven kinds of characteristic variables, by the average, at the same time, after the cluster four groups of abnormal rhythm data, according to the category of the average of seven characteristic variables, the average of the abnormal rhythm data and normal rhythm data comparison analysis (abnormal data-normal data), as shown in Table 4:

Table 4. Comparison table of heart rhythm data

		maximum	mean	Square root magnitude	standard deviation	Skewness	Waveform metrics	Steepness indicator
Mean normal heart rhythm		55.682	0.108	21.716	7.362	1.471	7.580	2.563
Mean abnormal heart rhythm	Class 1	51.713	0.654	20.638	8.144	1.182	6.355	2.505
	Class 2	55.509	0.139	21.681	7.343	1.444	7.564	2.560
	Class 3	60.391	0.116	22.856	7.053	1.981	8.600	2.641
	Class 4	52.701	0.214	21.058	7.570	1.079	6.966	2.503
Difference	Class 1	-3.969	0.546	-1.079	0.782	-0.289	-1.225	-0.057
	Class 2	-0.173	0.031	-0.035	-0.020	-0.027	-0.016	-0.003
	Class 3	4.709	0.008	1.139	-0.309	0.509	1.020	0.078
	Class 4	-2.981	0.106	-0.658	0.208	-0.393	-0.614	-0.060

From the above results, it can be seen that the maximum value of class 1 is small, that is, the DC component of the original data is small, the overall mean is large, and the data fluctuation of class 1 from the root value is weak. It can be seen from the standard deviation that the data is discrete, the wave peak is weak, and the cliff index <3 belongs to the normal range. Class 2 is less different from the normal heart rhythm data, which is a normal heart rhythm, but the beating frequency is slow, and the cliff index <3 belongs to the normal range. The maximum value of class 3 is large, that is, the DC component of the original data is large, the data of class 3 fluctuates strongly, the data is concentrated, the wave peak is prominent, and the cliff index <3 belongs to the normal range. Class 4 is small from the analysis of the maximum value, and the peak is shifted to the right. From the waveform index, the wave peak is weak, and the cliff index <3 belongs to the normal range.

3.5. The correspondence of the analytical results to the realistic situation of the arrhythmia

In reality, common arrhythmias include atrial fibrillation, premature ventricular beats, atrial premature beats, supraventricular premature beats and so on. Comparing the data results with the types of arrhythmia in the real world, The arrhythmia type corresponding to each cluster can be obtained, The DC component of class 1 is small and the wave peak is weak, However, the larger mean value indicates that the high-frequency peak is significantly decreased or disappeared, While the low-frequency peak enhancement, The corresponding arrhythmia type is atrial fibrillation; Class 2 has fewer fluctuations but fewer beats, The corresponding type of arrhythmia is sinus bradycardia; The DC component of class 3 is too large, Strong fluctuation, Peak highlights indicates its significantly enhanced high frequency component in the spectrum, The corresponding arrhythmia type is supraventricular tachycardia; Class 4 peak shifts to the right, The wave peaks become weak, Treat it as a right atrial conduction obstruction. The corresponding categories of the specific arrhythmia types are shown in Table 5:

In reality, common arrhythmias include atrial fibrillation, premature ventricular beats, atrial premature beats, supraventricular premature beats and so on. Comparing the data results with the types of arrhythmia in the real world, The arrhythmia type corresponding to each cluster can be obtained, The DC component of class 1 is small and the wave peak is weak, However, the larger mean value indicates that the high-frequency peak is significantly decreased or disappeared, While the low-frequency peak enhancement, The corresponding arrhythmia type is atrial fibrillation; Class 2 has fewer fluctuations but fewer beats, The corresponding type of arrhythmia is sinus bradycardia; The DC component of class 3 is too large, Strong fluctuation, Peak highlights indicates its significantly enhanced high frequency component in the spectrum, The corresponding arrhythmia type is supraventricular tachycardia; Class 4 peak shifts to the right, The wave peaks become weak, Treat it as a right atrial conduction obstruction. The corresponding categories of the specific arrhythmia types are shown in Table 5:

Table 5. Type of arrhythmia corresponding to each category

category	The number of category files	Arrhythmia type
1	211	Atrial fibrillation
2	255	Sinus bradycardia
3	133	Ventricular tachycardia
4	270	Right atrial block

4. Improved Topsis model based on entropy weight method

4.1. Topsis Establishment of the comprehensive evaluation model

Before the establishment of Topsis model [4], the entropy weight method is used to determine the value of the index. The entropy weight method [3], a hierarchical analysis method with strong subjectivity, adopts objective assignment, and the assignment result is more reasonable. The result of the entropy weight method comes from the data itself, that is, the less the variation of the index, the less information reflected, and the lower the weight. Therefore, we choose the entropy weight method to assign the weight value.

As shown in Table 6 and Figure 3, Although the maximum value is the largest in the index, But after a forward-forward standardization, Its certain weight is not large, That is, in the abnormal data, The degree of variation in the maximum value is too small, The reaction is also less informative; alike, Although the cliff index is small by the difference of the data, But it fluctuates greatly, So the weight of this index is 0.147; It is obvious that the mean has the largest weight in the radar map, It means that the mean is the most variable, The largest amount of information is reflected, In calculating the score, The largest proportion of the average value; The proportion of standard deviations is the second, For 0.209, This indicates that the dispersion of the data in the abnormal data varies greatly.

Table 6. The weights determined by the entropy weight method

	maximum	mean	Square root magnitude	standard deviation	Skewness	Waveform metrics	Steepness indicator	sue for peace
The weight determined by the entropy weight method	0.0514	0.3635	0.0435	0.2085	0.1413	0.0449	0.1469	1

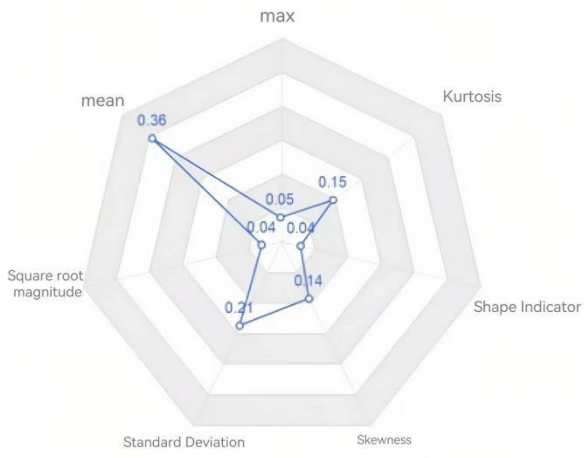


Fig 3. The weight of each index compares the radar chart

Topsis [3] Method (superior and inferior solution distance method) is a comprehensive evaluation method that approaches the ideal result. We can take the normal heart rate data as the ideal result, calculate and sort the score of the abnormal heart rate data. The following figure is the flow chart of the Topsis evaluation model.

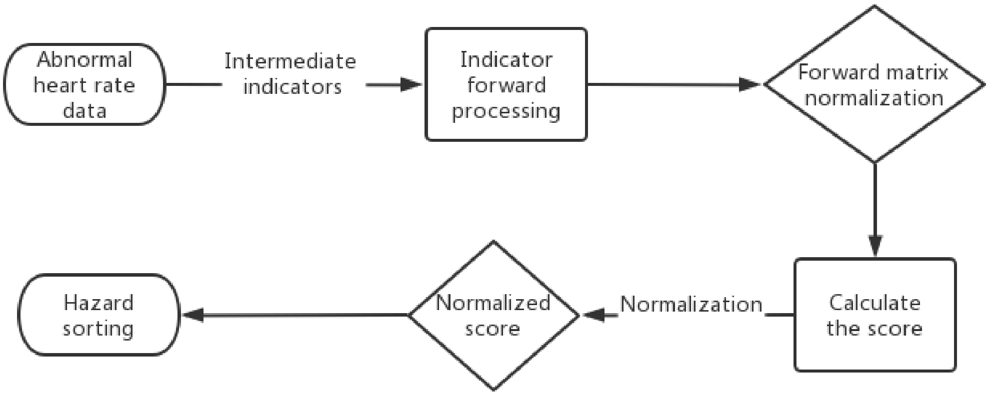


Fig 4. Topsis flowchart

4.2. Analysis of hazard results of abnormal rhythm data

Table 7 through the Topsis model to analyze the results: from the theory of the Topsis model, the score of the Topsis, the higher the comprehensive performance of the file, the data file, the closer to the normal rhythm, the standard, the lower, the risk, and the lower the comprehensive performance of the data file, the worse, the higher the risk degree. Table 7: Document 717619 has the highest score, the best overall performance, the lowest risk, the closer to the normal rhythm; the lowest symptoms in arrhythmia types; Document 6151175_1 has the lowest score, the worst overall performance and the highest risk; similarly, the top 10 patients have a high risk, which means that the patient's condition is critical and requires immediate treatment.

Table 7. Ranked 10 in Topsis scores and risk

filename	File sorting	Score ranking	Normalized score	Hazard ranking
6151175_1	546	869	0.00056	1
7121078_5	634	868	0.000572	2
711_411	618	867	0.000596	3
615857_8	583	866	0.000605	4
714935	728	865	0.00061	5
6121412	465	864	0.000615	6
7161273_8	752	863	0.000638	7
7191409	860	862	0.000655	8
615688_8	574	861	0.000656	9
609889_3	382	860	0.000684	10
……	……	……	……	……
717728_5	820	10	0.001413482	860
7181350	842	9	0.001415263	861
717149	789	8	0.001415431	862
419401	121	7	0.00141816	863
7121017_5	622	6	0.001421778	864
717497	794	5	0.001422467	865
714528	716	4	0.001422692	866
713671	694	3	0.00142325	867
6121202	446	2	0.001425275	868
717619	812	1	0.001427104	869

5. Evaluation of the model

5.1. Advantages of the model

(1) Avoid the subjectivity of the data, do not need the objective function, do not test, and can describe the comprehensive impact of multiple influence indicators.

(2) There is no strict limit on data distribution, sample size and index quantity, which is not only suitable for data with small samples, but also convenient to evaluate large systems with many units and indexes, which is more convenient.

(3) It can obtain good results with comparable evaluation ranking.

(4) Making full use of the original data can accurately reflect the distance of different evaluation schemes.

(5) The GMM algorithm runs fast and is the fastest algorithm in the hybrid model.

(6) The GMM algorithm has Agnostic properties

5.2. The shortcomings of the model

(1) It is required to have the data of each indicator, and it is difficult to select the corresponding quantitative indicators.

(2) Uncertain about the number of indicators to better describe the influence of each index.

(3) More than two study subjects were required.

(4) The GMM algorithm has singularities, making it difficult to estimate the covariance matrix when the number of each mixing point is insufficient, and unless someone artificially adjusts the covariance, the algorithm will diverge and find a solution with infinite possibilities.

(5) The number of components of the GMM algorithm is too large

5.3. Improvement of the model

(1) The clustering algorithm can't deal with the data fluctuation problem of normal and abnormal rhythm data, and should be forward and standardized in the comparative analysis of data.

(2) Can improve the arrhythmia data characteristic index weight, may improve place mainly reflected in the Topsis method index weight assignment, we only use the entropy weight method combined with the original data to assign, in fact, we can combine experts scoring using hierarchical analysis to make the weight more scientific significance, the entropy weight and hierarchical analysis weight after the corresponding proportion of the standard weight. The idea of combinatorial empowerment [6] is used to empower the indicators.

(3) Prior knowledge of the patient can be introduced to the [5]. The prior knowledge of the patient is summarized in the auxiliary diagnosis system. According to the prior knowledge, the corresponding representative characteristic components are selected to establish a more perfect standard library of arrhythmia types, so as to further improve the classification efficiency and universality.

5.4. The promotion of the model

The method used in the process of building the model can be extended to different fields. For example, in the brain wave [7] analysis, smooth wavelet transform and double threshold function can also be used to denoise the data, then conduct standard feature extraction, and use clustering method for classification, which can greatly improve the analysis efficiency. Of course, this idea can also be used in other fields in addition to medicine, such as industrial parts production standard judgment and product quality evaluation analysis [8].

References

[1] Wang Lingli. Research on the ECG signal detection and classification algorithm [D]. Shaanxi University of Science and Technology, 2019.

[2] Daubechies. [M]. Beijing: National University of Defense Technology Press, 2011:23.

[3] Yang Yueli, He Xingguang, Huang Zhen, Fan Jiangrong, Liu Qisheng. Application of entropy weight TOPSIS and comprehensive index method based on oncology department in comprehensive evaluation of DRGs [J]. Modern Hospital, 2023,23 (02): 244-246.

[4] Liang Lijun. Identification of key influencing factors of disease diagnosis and evaluation of treatment options based on individual patient differences [D]. Tianjin University, 2014.

[5] Yan Yu, Sun Cheng. An ECG beat classification algorithm based on cluster analysis [J]. Computer Applications, 2014,34 (07): 2132-2135 + 2139.

[6] Fu Jing. Study on the supplier evaluation model based on the improved TOPSIS [D]. Liaoning Normal University, 2011.

[7] Chen Lifan. Identification and prediction of the EEG time series [D]. And Fuzhou University, 2020.DOI: 10.27022/d.cnki.gfzhu. 2020.000359.

[8] Li Yuxin, Xing Na, Bai Haodong, etc. Research on the quality of notoginseng and its artillery products from different producing areas by TOPSIS method based on entropy weight method [J]. Chinese herbal medicine, 2023,54 (04): 1252-1259.