

# Similarity calculation method for time series features based on rough set

**Jin Ling**

*Jiangxi University of Technology,  
NanChang 330098, JiangXi, China*

**Mu Zhendong\***

*Jiangxi University of Technology, NanChang 330098, JiangXi, China*  
*\*Corresponding author*

### Abstract

Informatization brings a large amount of time series information, and the method for extracting the feature of such information has been obsolete. In this article, the feature extraction between EEG signals of the same type was carried out by analyzing the EEG signal with an evident time series feature as a study object, by using the rough set thinking to define the upper approximation and lower approximation of a time point feature and via the method for defining similarity. The feature of visual evoked potentials (VEP) of five persons was calculated and verified using definition method. The result shows that the deviations of calculated similarity of 400 samples from each of five subjects are within three; and at the same time the sum of the 400 samples was calculated by taking the superimposed average EEG as a standard. The sample feature of the five subjects was calculated using the method designed herein, and then matching calculation was done with superimposed feature. As a result, the number of samples whose fit is above 0.7 accounts for 80% of the number of all samples.

Key words: TIME SERIES, FEATURE EXTRACTION, ROUGH SET, ELECTROENCEPHALOGRAPH (EEG) SIGNAL, SIMILARITY

### Introduction

With the development of information technology nowadays, the analysis of mass data with time features has become possible. Therefore, the feature extraction and study of mass time series data have become the main research direction of data mining. A time series feature has two opposite characteristics: certainty and uncertainty. The certainty refers that in general many times of overall analysis of time series information corresponds to different results with a constant

feature of time series. The uncertainty refers that for single time series information, such constant information is not complete or the time point when it occurs is uncertain. For solving the uncertainty, three mathematical theories were generated and they are fuzzy set theory, rough set theory and quotient space theory. These theories have been successfully applied in the fields such as data mining, problem solving, pattern recognition, etc. Because corresponding features can be extracted from information itself using the rough set theory

and this process needs no other knowledge, the extraction of many features is done using a rough set method.

During study of data mining, time series information includes much time-related information such as meteorology, finance, management, astronomy, earthquake, etc. Such information is discrete and continuous and has different study methods respectively. For example, Tim Shimeall and Phil Williams[1] made trend analysis of computer network information security and divided the trend into: the trend of internal and external causes, temporal trend, space trend, associated trend and mixed trend. And they introduced trend analysis methods for different types of trend analyses. For the study of such different trends, successful study methods are weighted average, degree of freedom, and least square method.

Up to now, some achievements have been obtained by using the rough set theory to analyze temporal data. Ostroff S[2] put forward a real-time temporal logic frame, which indicates temporal series using event variables. Berndt J D[3] detected the mode of time series using dynamic programming. Bazan Jan G[4] et al. made analysis of market data using rough set and dynamic reduction and succeeded. Golan R[5,6] made analysis of Canada stock data and put forward the thinking for converting time series to a traditional information system. Anders T B[7] brought forward a time series information system and the concept to implement the system, and formalized the thinking for converting time series to an information system. When rough set was used to process temporal data, previous study mainly paid attention to the time series features of mined temporal data, i.e. a strict time sequence is kept between objects. Time series information is divided into two types[8]: the time series without real-time restriction and the time series with real-time restriction. The former can be considered as a time string arranged according to time, and the time interval (sampling rate) between events is constant; while the latter must consider the time intervals between events, which differ from each other greatly.

With the development of biological information technology, complex human body information can be known by collecting various kinds of signals from human body. The knowledge of such information helps to treat many complex diseases and realize the development of many human body imitating techniques. Many of such signals are the ones of time series, such as electrocardio[9], electromyogram[10] and EEG

signals[11], etc. Complex time series include many physiological signals. The feature extraction of these signals is one of the main subjects in biological information study. Such feature extraction is often done by introducing a signal analysis method to biological identification technology, such as Fourier transformation[12], wavelet analysis[13], etc. However, physiological signals are special signals. Therefore, how to extract features from signals themselves is a difficulty baffling many study team.

In this article, the feature extraction of time series is realized by taking EEG signal as a study subject, collecting EEG signals via a proven EVP test mode, dynamically defining upper and lower approximation and a similarity calculation by using rough set method, and removing noise data gradually by adding time windows. The analysis result of 400 collected EEG signals of the same mode shows that the similarity among features of the 400 samples is up to 80%.

### Rough set theory

In 1982, Professor Z. Pawlak et al. in Poland Warsaw University of Technology came up with the concept of rough set, and up to now a perfect rough set theory system has been formed. In this system, the information in the real world can be generally indicated by a piece of information. Each line in the information is called an instance (entity, object), whose nature is reflected by assigning some variables. The main components of an information table knowledge expression system is a set of studied objects. The knowledge of such objects is described by assigning their attributes (features) and attribute values (characteristic values). Generally, an information table knowledge system can be expressed as  $S = (U, R, V, f)$ , where

$U$  = the set of objects;

$R$  = ; Subsets  $C$  and  $D$  are called a conditional attribute set and a resulting attribute set respectively;

$V$  = , the set of attribute values;  $V_i$  refers to the attribute value range of Attribute  $i$ , i.e. the range of  $r$ .

$f$ :  $f: U \times R \rightarrow V$  is an information function, which assigns the attribute value of each object  $x$  in  $U$ .

For each attribute subset  $B \subseteq R$ , we define an indiscernibility binary relation (indiscernibility relation)  $IND(B)$ , i.e.

$$IND(B) = \{(x, y) \mid (x, y) \in U^2, \forall b \in B(b(x) = b(y))\} \quad (1)$$

Definition 1: Assume . When  $X$  can be described exactly by the attribute subset,  $B$  (i.e. the merge of indiscernibility set on  $U$  determined according to attribute subset  $B$ ), we call that  $X$  can

be defined by B. Otherwise X cannot be defined by B. The definable set of B is also called the exact set of B, and the indefinable set the inexact set or rough set of B (rough set for short).

Definition 2: For each definition X (subset of instances) and indiscernibility relation B, both the maximum definable set included in X and the minimum definable set including X can be determined according to B. The former is called the lower approximation set of X (indicated by  $\underline{B}(X)$ ), and the latter the upper approximation set of X (indicated by  $\overline{B}(X)$ ).

Definition 3: Given knowledge system  $S = (U, R, V, f)$ . For each subset  $X \subseteq U$  and indiscernibility relation B, the upper and lower approximation sets of X can be expressed by the basic definition of B respectively as follows:

$$\underline{B}(X) = \bigcup \{Y_i \mid (Y_i \in UIND(B) \wedge Y_i \subseteq X)\}; \quad (2)$$

$$\overline{B}(X) = \bigcup \{Y_i \mid (Y_i \in UIND(B) \wedge Y_i \cap X \neq \emptyset)\}; \quad (3)$$

Where,  $U/IND(B) = \{U_i\}$  is the division of U by the indiscernibility relation B, that is to say, the basic set B of discourse domain U.

## EEG signal

The EEG signal is a bioelectrical signal which is generated by human brain itself or due to external stimulus and can be collected by an exclusive instrument. It is very difficult for us to purely attribute consciousness or thinking to the change of the organization, cell or neurotransmitter in some part of brain through objective evaluation of knowledge process, because it is hard to treat specific psychological activities by using specific microscopic natural scientific methods such as molecular neurobiology and neurobiochemistry. In 1960s, Sutton put forward the definition of event-related potential, which reflects the nerve electrophysiology change in the brain during knowledge process by recording the potential evoked by the brain on the surface of skull by means of average superimposed technology, so it is

called the "window" to "peek" the psychological activities. The development of nerve electrophysiology technology provides new methods and ways to study the process of brain knowledge activities.

The main components of a classic ERP include  $P_1$ ,  $N_1$ ,  $P_2$ ,  $N_2$  and  $P_3$ . The first three are called exogenous components and the last two endogenous components. The main features of these components are: firstly, they are not only the reflection of pure psychological activities in the brain, but also reflect some aspects of psychological activities; secondly, they must be induced by more than two special stimuluses or stimulus changes. Among the main components,  $P_3$  in the ERP is the most focused and studied and the most key indicator for lie detection. Therefore,  $P_3$  becomes the synonym of ERP to some extent. The differences between ERP and general evoked potential are as follows:

(1) The subject (tested person) is required to be conscious generally;

(2) All stimuluses are not single repeated flashes and short sound stimuluses, but have at least two or more than two stimulus series (Stimulus signals are indefinite and can be visual, hearing, digital, language, or image);

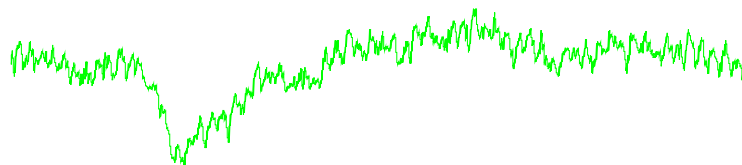
(3) The components include not only exogenous components easily affected by stimulus physical properties, but also the endogenous components not affected by physical properties;

(4) Endogenous components are closely related to knowledge process.

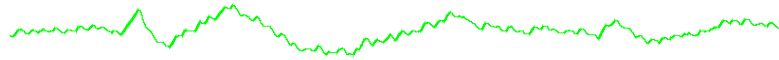
The occurrence of ERP components is the result of many superimpositions of the same kind of test, which can make the uncertainty in EEG signals aggregate into certain features. As shown in Figure 1, when disorderly EEG signals are superimposed, evident ERP components will occur.

Subject:  
EEG file: original.eeg Recorded : 11:12:34 30-Dec-2014  
Rate - 1000 Hz, HPF - 0 Hz, LPF - 300 Hz, Notch - off

Neuroscan  
SCAN 4.3  
Printed : 09:42:55 18-May-2015



(a)



(b)

**Figure 1.** Orderly Features in Disorderly EEG signals.

Figure 1 (a) is original EEG signal. Figure 1 (b) is the diagram of EEG signals superimposed for many times. From the comparison diagram of Figure 1, the original signal makes EEG signal feature not evident due to noise; however, the oscillogram shows that the original signal includes the traces of some features.

**Preprocessing method**

The EEG data used herein were collected by the brain-machine interface laboratory in the Jiangxi University of Technology from the students of Jiangxi University of Technology. Each subject sat relaxedly on a soft chair without handrails in a quiet shielded room, watching a computer screen in front of him, and conducted an EEG test according to the tester's arrangement and stimulus instructions on the screen. EEG data were obtained using a 40-sample Neuroscan amplifier and software scan4.3. Right mastoid was taken as a reference electrode; the sampling rate of 1000Hz used; 200 Hz low pass, 0.05Hz high pass and 50Hz trapped wave used for collecting wave bands.

The method for preprocessing data is as follows:

(1) Removing EEG signals with a great drift: During acquisition of EEG signals, the initial EEG signals drift greatly due to the movement and absent-mindedness of the subject, external noise, etc., which can result in effect on subsequent EEG signal processing. Therefore, these initial EEG signals must be removed before EEG signal processing;

(2) Removing electrooculogram (EOG): In the original EEG signals, the EOG produced due to blinking or looking left and right causes effect to EEG signals. Therefore, such effect should be removed before feature extraction and

classification. The effect of vertical EOG is mainly removed herein.

(3) Segmentation: examine the interval of stimulus, which generally is 10%-20%; the common values are -50 and -100; and the data segment taken herein is -100 - 923ms on the principle of not exceeding the start of next event.

(4) Baseline correction: Because the data processed sectionally are not on a baseline, two baseline corrections and one linear correction are carried out herein.

(5) Removing artifact: After collected EEG signals are classified, some segments cause bad data due to various kinds of reasons. They are not good for data analysis but affect data analysis. Therefore a certain limit should be selected for screening. The limit used herein is -80 - 80.

**Feature extraction method**

The EEG signal is a time series in the field of real numbers. As shown in Figure 1, although the original EEG signal includes certain ERP features, they are hard to extract. The reason is that the EEG signal is a weak electric signal, the features of which don't occur completely in one or several signals and are easy to be disturbed by noise data. The feature extraction herein is done by referring to information table, using wave crest and valley change and adding windows. The detailed algorithmic method is as follows:

Step 1: n-sample EEG signal.

Step 2: Calculate the gradient of input EEG signals, keep crest values, and zero remaining signals, and the crest value vector on a time series is obtained in this way. Define the crest value to be 1 and the valley value to be -1, and the waveform vector of time series is obtained in this way. Classifying n-sample EEG signals: Classify the

EEG signals of the same event into one category and mark them with 1, 2, 3... , respectively. The merging time span, TW is equal to 1.

Step 3: Define the window's size to be T and window's step to be s, and start calculating the similarity in the window of the similar type of EEG signals from start time point.

Step 4: According to the similarity calculated in Step 3, divide the wave crest and valley values of EEG signal into upper and lower approximation sets; and the span of merged points, TW is equal to TW plus 1;

Step 5: Make looping execution of Step 3 until all valley points are added to upper approximation or the merged span, TW is equal to 1/T.

Step 6: Output

The so-called event refers to the stimulus signals used to induce EEG signals from a subject during EEG signal collection;

The similarity used herein is divided into two parts: the shape similarity to indicate the change trend of time series and the value similarity to indicate the value of time series. If the definitions of two numbers  $a_i$  and  $a_j$  are operated as follows:

$$f(a_i, a_j) = a_i \oplus a_j = \begin{cases} 1 & a_i = a_j \\ 0 & a_i \neq a_j \end{cases} \quad (4)$$

The formula for calculating shape similarity  $Sis$  of n waveform vectors of time point x on a time series can be represented as follows:

$$Sis = \sum_{i,j=1}^{i,j=n} a_i \oplus a_j \quad (5)$$

For a classified threshold with a set similarity, if exists, the x time point of time series sample is of shape similarity.

For a time point x for n series samples, if  $mx1$  stands for the maximum value of the vector of such time point,  $mx2$  for the second maximum value,  $mi1$  for the minimum value, and  $mi2$  for the second minimum value, the similarity threshold, of n time series values corresponding to the time point x can be calculated as follows:

$$\omega_x = \frac{|mx1 - mx2|}{|mi1 - mi2|} \quad (6)$$

If the time point x is of shape similarity, the similarity  $Siv$  of n vector values of the time point x can be calculated using the following formula:

$$Siv = \frac{\sum_{i=1}^n (a_i / \sigma)}{n\omega_x} \quad (7)$$

Where,  $a_i$  refers to the vector value of the time point x, and  $\sigma$  the variance of vectors.

If the tolerance range interval for value similarity is  $[-Tx, Tx]$ , the definition of upper and lower approximation sets of n time series samples is as follows:

$$B_-(x) = \{x | Siv_x \in [-Tx, Tx]\} \quad (8)$$

$$B_+(x) = \{x | Sis \geq \lambda\} \quad (9)$$

When an EEG signal with a time series, the original signal, crest value vector and waveform vector are shown in Figure 2.

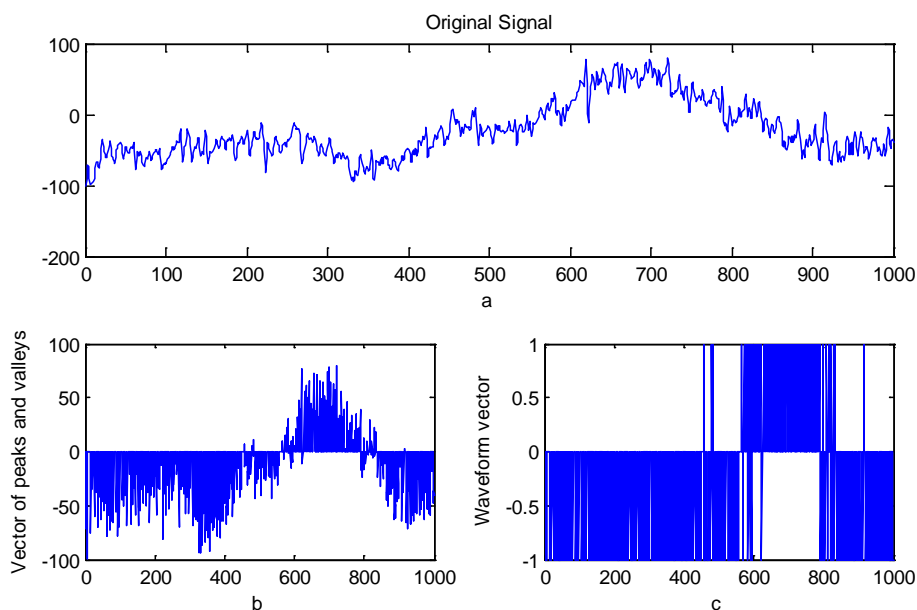


Figure 2. Comparison after EEG signal change

Figure 2 (a) is the original EEG signal; Figure 2 (b) is the crest and valley diagram after value transformation; and Figure 2 (c) is an oscillogram.

**Result**

In order to verify the effectiveness of the method herein, we chose five subjects and each subject chose 400 effective EEG signals in the same event. Firstly, the superimposed average of such signals was calculated using superimposed and taken as feature bases.

For samples with several time series, there is a certain time difference between the features occurred on each sample and overall features. Therefore, a certain step was set herein and calculation was done according to time window. The similarity value was calculated during time window sliding process at the step, by which the upper and lower approximations of each window were calculated. The calculation of the similarity of each time series was mainly carried out in the lower approximation set. After the time series was slid by the step window, a vector whose element is the lower approximation set was formed. For Sample A of time series, the finally calculated lower approximation set vector marking is  $TB = \{Win_1, Win_2, Win_3 \dots Win_k\}$  (Where,  $k$  is the number of windows). The similarity SiD of  $Wini$  of Sample  $A_1$  and  $A_2$  can be calculated using the following method:

$$SiD_i(A_1, A_2) = \frac{card(Win_{A1(i)} \cap Win_{A2(i)})}{card(Win_{A1(i)} \cup Win_{A2(i)})} \quad (10)$$

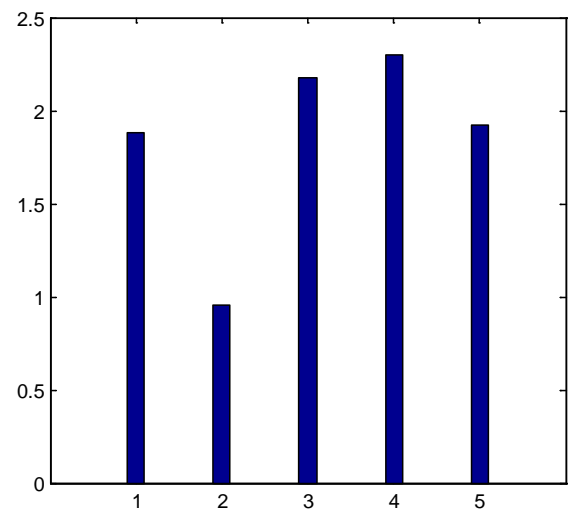
Where,  $card(x)$  refers to the radix of Set  $x$ , and  $i$  the lower approximation in the Window  $i$  of Vector  $A$ .

The similarity SiD among  $n$  vectors can be calculated using the following formula:

$$Sid = \frac{\sum_{i=1}^k (\mu_i / \sigma_i)}{k} \quad (11)$$

Where,  $\mu$  and  $\sigma$  refers to the average and variance of values in SiDi.

For 400 samples from five subjects, the similarity of such subjects calculated using the above method is 1.8836, 0.9536, 2.1737, 2.3010 and 1.9181 respectively, as shown in Figure 3. This indicates that the similarity calculation of time series features was successfully realized using the method.



**Figure 3.** EEG Similarity

Superimposed average signals were compared and the features of 400 samples were compared too. The result is shown in Table 1.

**Table 1.** Superimposed Average Feature Complies with Sample Number Distribution

	< 70%	70%-80%	80%-90%	> 90%
Subject 1	58	129	190	23
Subject 2	98	162	191	12
Subject 3	59	118	209	14
Subject 4	33	145	151	51
Subject 5	40	111	205	44

Table 1 shows the sample distribution after comparison of the features between the EEG signal sample and superimposed average sample of the five subjects. Where, > 70% refers to the number of samples whose superimposed average comparison similarity feature is below 70%. According to Table 1, the calculated number of samples below 70% of the five subjects accounts for 14.50%, 13.75%, 14.75%, 13.25% and 10.00%

of the number of all samples respectively, while the calculated number of samples above 90% accounts for 5.75%, 3.00%, 3.50%, 12.75% and 11.00%. Regardless the maximum and minimum number of samples, the sample distribution indicates that the proportion between 70%-90% is 79.75%, 83.25%, 81.75%, 74.00% and 79.00% respectively.

## Conclusions

The feature extraction of time series information is the main study direction of information feature extraction all the time. With the rise of EEG signal study, the time series information feature analysis of this special signal also has become the study direction of many researchers. In this article, the EEG signals of five subjects' response to the same event were intercepted and 400 segments of EEG signals of each subject were selected by taking EVP as a study object. Calculation was done by using the method designed herein. The result shows that superimposed average feature can be well matched. This lays a foundation for single feature analysis of EEG signals.

Certainly, the most disadvantage of the article is the requirements for input signals. Because the EEG signals with a great drift cannot be removed automatically, it is required to check EEG signals manually when the method is used for feature extraction.

## Acknowledgements

This work was financially supported by project of Technology Department of Jiangxi Province [No 20143BBM26048] and project of Jiangxi University of Technology [No. xtcx201312].

## References

1. Shimeall, T. J., Williams, P. (2002) Models of information security trend analysis. *Sensors Command Control Communications Intelligence Technologies for Homeland Defense Law Enforcement*, p.p.43-52.
2. Ostroff, J. S. (1989) Temporal logic for real-time systems. *Research Studies Press*, p.p.22-28.
3. Bernad, D. J. (1996). Finding patterns in time series: a dynamic programming approach. *Advances in knowledge discovery and data mining*, p.p.229-248.
4. Bazan, J., Skowron, A., Synak, P. (1994) Market data analysis: A rough set approach. *ICS Research Reports*, 6, 94.
5. Golan, R., Edwards, D. (1994) Temporal rules discovery using datalogic/R+ with stock market data. In *Rough Sets, Fuzzy Sets and Knowledge Discovery*, Springer, London, p.p.74-81.
6. Golan, R. H., Ziarko, W. (1995) A methodology for stock market analysis utilizing rough set theory. In *Computational Intelligence for Financial Engineering, 1995., Proceedings of the IEEE/IAFE 1995*, p.p. 32-40.
7. Anders, T. B. (1997) Mining Time Series Using Rough Set-A Case Study. In *Proceeding of the First European Symposium*, pp. 256-263.
8. Yin, X., Shang, L.(2001) Study of Time Series Mining by Rough Set. *Journal of Nanjing University*, 37(3), p.p.182-187.
9. Słowiński, R., Greco, S., Matarazzo, B. (2014) Rough-set-based decision support. In *Search Methodologies*, Springer, US, p.p. 557-609.
10. Jia, X., Tang, Z., Liao, W., Shang, L. (2014) On an optimization representation of decision-theoretic rough set model. *International Journal of Approximate Reasoning*, 55(1), p.p. 156-166.
11. Qian, Y., Zhang, H., Sang, Y., Liang, J. (2014) Multigranulation decision-theoretic rough sets. *International Journal of Approximate Reasoning*, 55(1), p.p. 225-237.
12. Le Van Quyen, M., Martinerie, J., Baulac, M., Varela, F. (1999) Anticipating epileptic seizures in real time by a non-linear analysis of similarity between EEG recordings. *Neuroreport*, 10(10), p.p. 2149-2155.
13. Stassen, H. H., Lykken, D. T., Bombent, G. (1988) The within-pair EEG similarity of twins reared apart. *European archives of psychiatry and neurological sciences*, 237(4), p.p. 244-252.

