Journal of Artificial Intelligence
ISSN: 2229-3965 & E-ISSN: 2229-3973, Volume 3, Issue 3, 2012, pp.-117-122.
Available online at http://www.bioinfo.in/contents.php?id=71

RESEARCH ON IMPROVED MEAN SHIFT ALGORITHM BASED ON LOCAL DISTRIBUTION IN EEG SIGNAL CLASSIFICATION

ZHANG SHAO-BAI[1*], HAN YAN-BIN[2], LI JIN-PING[2] AND CHENG XIE-FENG[1]

[1]College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu-210046, China
[2]School of Information Science and Engineering, Jinan University, Jinan, Shandong-250022, China
*Corresponding Author: Email- hyb309@gmail.com

Received: June 19, 2012; Accepted: September 10, 2012

Abstract-In the study of brain computer interface, Mean Shift (MS) which was improved by the local samples distribution was applied to classifier design for electroencephalography (EEG). The raw EEG was extracted the different frequency band feature by the discrete wavelet transformation (DWT), furthermore the characteristic component would be selected by the generic algorithm (GA) to promote the classification efficiency. Compared with the traditional MS, the direction of shift was guided by the local samples distribution, which could reduce the risk of across classifying boundary. The performance of new classifier was tested using the dataset form "BCI 2003 competition", which recognition rate is better than linear discriminate analysis (LDA) and support vector machine (SVM). The accuracy of new method is 92.1% and is better than the best result of the competition.

Keywords- brain-computer interface, discrete wavelet transformation, electroencephalography, mean shift, classifier

Citation: Zhang Shao-bai, et al (2012) Research on Improved Mean Shift Algorithm Based on Local Distribution in EEG Signal Classification. Journal of Artificial Intelligence, ISSN: 2229-3965 & E-ISSN: 2229-3973, Volume 3, Issue 3, pp.-117-122.

Introduction

A brain-computer interface (BCI) is also called brain-machine interface (BMI). Primary research is how to make a direct communication between the brain and an external device without using nervous system. The BCI can private help to patient with paralysis and language barriers because of nervous system [1]. So the main research in BCI field is how to accurately analysis the electroencephalogram (EEG) which can represent the behavior intentions. Recently the research of BCI includes feature extraction and feature classification.

The analysis process of BCI includes 4 stages: EEG collection, EEG preprocessing, feature extraction, feature classification [3]. Because the EEG contains brain activities in a period of time and the information from multi-electrodes, the data dimension is higher and the data is larger. So the analysis process of BCI also uses feature selection to reduce data dimension[4]. Because the EEG is a time variant non stationary admonitory, the feature extraction method of EEG mainly include discrete wavelet transform (DWT), Lipschitz exponent and entropy[5-8]. The effect of feature extraction base on the wavelet depends on wavelet families and selection of appropriate wavelet is very important for the analysis of signals [6-7]. The feature extraction algorithm base on entropy also depends on selection of the entropy. Currently classifier design is mainly support vector machine (SVM), linear discrimination analysis (LDA) etc.

The goal of this present work is to design a classification base on mean shift (MS) to discriminate the different EEG. The local samples distribution is considered in new method, which is used to estimate the new probability density center of samples. The local samples distribution can decrease reduce the risk of across classifying boundary to guide the direction of shift and then accurately find out the class center of samples. The new classifier was called LSD-MS (Local Sample distribution-MS). In the present study, features are extracted from raw EEG data and then obtained the eigenvector construct of different frequency band and then the features are input the classification designed by LSD-MS to recognize the raw EEG. In order to evaluate the effectiveness of new classification, the BCI competition Dataset-Ⅲ and we used the GA to reduce the features which were from three scales of raw signal. The process of feature reduction can find the relation between frequency band and motor image to improve the classification ability. Comparing with LDA, SVM and the result of competition, the recognition rate of new method is 92.1%.

The paper is organized as follows: section 2 describes the feature extraction, feature selection; section 3 describes LDS-MS classifier; Section 4 reports the analysis results and makes some discussion;

Section 5 concludes the paper.

Features Extraction and Features Selection

Materials

This dataset was recorded from a normal subject (female, 25 yrs.) during a feedback session. The subject sat in a relaxing chair with armrests. The task was to control a feedback bar by means of imagery left or right hand movements. The experiment consists of 7 runs with 40 trials each. Given are 280 trials of 9s length. Three bipolar EEG channels were measured over C3, Cz and C4 [Fig-1]. The EEG was sampled with 128 Hz, it was filtered between 0.5 and 30Hz.

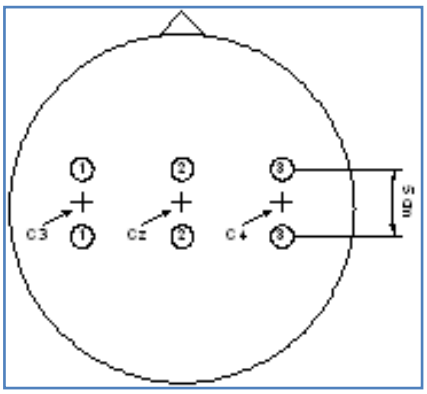

Fig. 1- The electrode positions

The first 2s was quite, at t=2s an acoustic stimulus indicates the beginning of the trial, the trigger channel (#4) went from low to high and a cross "+" was displayed for 1s; then at t=3s, an arrow (left or right) was displayed as cue. At the same time the subject was asked to move a bar into the direction of the cue. The process is shown by the [Fig-2].

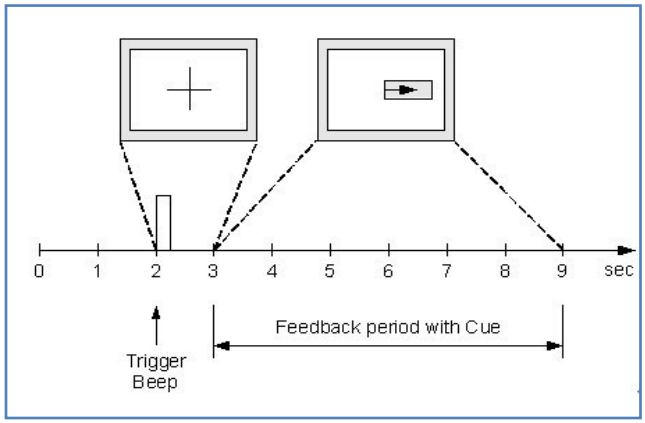

Fig. 2- Timing scheme

Features Extraction

EEG signals can real-time express the complex thinking activity of brain, so it is complex and non-stationary signal and it is spatial-temporal dependence [2]. So the DWT is applied to extract features of EEG and it gives a highly complete representation of EEG signals in the time-scale domain. It is the algorithm essence of DWT that the raw signal is decomposed into deferent frequency band (band-pass filters) to analysis the relation between EEG frequency and motor imagery [6,7]. The normal wavelet and scale basis functions φjk(x), ϕjk(x) can be defined as

$$\phi_{j,k}(x) = 2^{j/2} h(2^j x - k) \quad (1)$$

$$\varphi_{j,k}(x) = 2^{j/2} g(2^j x - k) \quad (2)$$

Each experiment consists of two motor imagery (left, right), which were identified by "1" and "2". So the recognition algorithm of the EEG is converted into two-category and it is the first step that how to extract the features of raw EEG.

According to the equation(1) and (2),the expression of the raw EEG can be defined as

$$A(x) = \sum_{k=-\infty}^{\infty} S_j(k) 2^{j/2} \varphi(2^j x - k) + \sum_{j=1}^{J} \sum_{k=-\infty}^{\infty} D_j(k) 2^{j/2} \phi(2^j x - k) \quad (3)$$

Where the $S_j(k)$ and $D_j(k)$ are the detail coefficients and the approximation coefficients. The expression can be define as

$$S_j(k) = < A(x), 2^{j/2} \varphi(2^j x - k) > \quad (4)$$

$$D_j(k) = < A(x), 2^{j/2} \phi(2^j x - k) > \quad (5)$$

Using the expression (1-5), raw EEG is decomposed into the detail coefficients and the approximation coefficients on different scales.

According to Subasi [14], the selection of the number of decomposition levels and the wavelet basis function is very important step in the analysis of signals by using the DWT. In this paper, the data sampling frequency is 128HZ, so the number of levels the EEG signals were decomposed into is chosen to be three. The decomposition scale is 3 and the frequency range distribution is shown by the [Table-1].

Table 1- Scale domain

Frequency band	Band range
D_1	32-64
D_2	16-32
D_3	8-16
A_3	0-8

According to the [Table-1], the raw EEG signal is decompose different scale (different frequency band) and the DWT decomposition algorithm is a band filter. The [Fig-3] shows the decomposition result, which the signal data is from electrode C3.

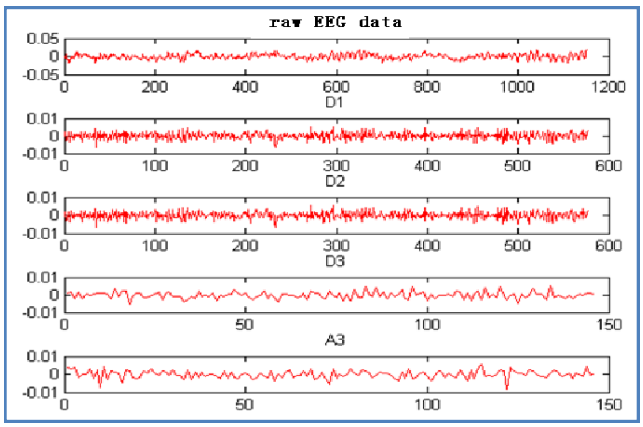

Fig. 3- The result of 3 levels scales wavelet decompose

Each sub-band data are extracted three statistical features and the EEG features is composed of all sub-band statistical feature. Three statistical features are as follows:

- **W_Mab:** Mean of the absolute values of the wavelet coefficients in each sub-band.
- **W_PAv:** Average power of the wavelet coefficients in each sub -band.
- **W_Std:** Standard deviation of the wavelet coefficients in each sub-band.

According to the [Table-1], each channels EEG signal are decomposed into 4 sub-band and 3 features are extracted on the each sub-band, so ach channels EEG signal are described by 12(3×4). In this experiment, the EEG data are from 3 electrodes (C3, Cz, C4), so some whole imagery movement can be replaced by 36 (3×12) feature vector. Now the description about feature vector are as follows:

- The dimension of the feature vector is 36 T;
- The 36 components are divided into 3 groups Ai (each groups is consist of 12 component), each groups are response to each electrode, i≤3 and T=[A1,A2,A3];
- The Ai are divided into 4 blocks (D1,D2,D3,A3), so each block is response to each sub-band [Table-1];
- Each block includes 3 components (W_Mab, W_PAv, W_Std) and the [Fig-4] shown those information.

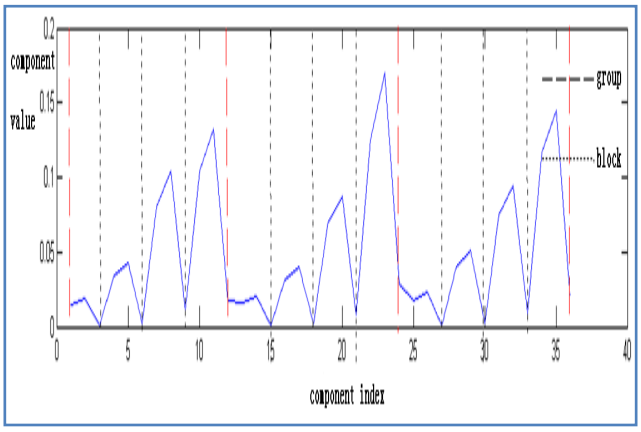

Fig. 4- The description of the feature vector

Features Selection

Features selection not only reduce data dimension to improve recognition effect, but also analysis the relation between electrode, sub-band and imagery movement. By feature selection, we can find which electrode and sub-band play an import role in BCI research. In this research, features selection is used to reduce not electrode number [4] but feature components [11] and improve the relative between feature components and imagery movement.

Though the frequency rang which is closely related with sport conscious is [8Hz-12hz] (α frequency) and [19Hz-30Hz] (β frequency), the sub-band frequency of movement plan and sense production are reserved to check the effect of LDS-MS. The features selection algorithm is GA and the algorithm flow is as follows.

- Many individual solutions are randomly generated to form an initial population; the population size is n; the set of individual solutions is G; each solutions (Gi, i≤n) is binary encoding and the 36 bit (Gij, j≤36) corresponds to 36 components of feature vector; if some component is selected and the corresponding bit is 1(Gij=1), else Gij=0;
- According to individual solution, some components are selected to calculate similarity of EEG to classify all signals, the classification accuracy of all EEG is act as the individual fitness measures.
- Generate the next generation by the strategy of selection which includes “roulette” and “elitism preserving”.
- If fixed number of generations don’t reach, then go to (2), else finish the iteration.

Then the individual solution which has the best fitness measures is the result of the features selection and then if Gij=1, the corresponding feature components are selected.

The Classifier Design Based LSD-MS

Analysis of Traditional MS

The mean-shift procedure was originally presented in 1975 by Fukunaga and Hostetler and then it was improved by Cheng in 1995. It was applied to cluster analysis and global optimization when it was modified kernel function and weighting function[9]. It make every point shift to the position which has local maximum of PDF (probability density function). That position is defined “shift center” and then one center represents one class. This algorithm is used in image classification and video tracking. Cheng propose the common expression of mean-shift. If X is the set of all samples, which dimension is m and the X={x1,x2,…xn}, n is the sample size, xi is m-dimension vector and the expression of the samples’ PDF is as follows:

$$f(x) = \sum_{i=1}^{n} w_i k(\beta \|x_i - x\|^2) \tag{6}$$

The general expression of mean shift is as follows:

$$\bar{x}_{k+1} = \sum_{i=1}^{n} w_i k'(\beta \|x_i - \bar{x}_k\|^2) x_i / \sum_{i=1}^{n} w_i k'(\beta \|x_i - \bar{x}_k\|^2) \tag{7}$$

Where, $\bar{x}_k$ is the center of the probability density; $\bar{x}_{k+1}$ is the new center of the probability density which is calculated by the current center of the probability density and PDF; wi is the weight, which satisfaction is $\sum_{i=1}^{n} w_i = 1$; the weight show samples contribution of construction of probability density distribution and if the samples distribution is unknown, the wi is 1/n;k(x) is kernel function, which satisfaction is $\int k(x)dx < \infty$; $\|x_i - \bar{x}_k\|^2$ is distance measure, which is the similarity measure between samples and $\bar{x}_k$;β is the windows radius of kernel function.

According to expression (6) and (7), the classification algorithm based on MS is as follows: the new center of probability density

$\bar{x}_{k+1}$ is calculated by $\bar{x}_k$ and this neighborhood; if the $\bar{x}_{k+1}$ and $\bar{x}_k$ doesn't overlap each other, the center of probability density is approximate to $\bar{x}_{k+1}$ and the center of probability density is shifted; the samples which are shifted the same center of probability density is regarded as the same class. So the direction which increased the density is the shifting direction. The kernel function is the key of the astringency.

The Defects of Traditional MS

According to the expression (6) and (7), the traditional MS utilizes all samples in the iteration method and the weight of the samples is calculated by k(x) (kernel function). This shifting strategy ignores distribution of local samples and the local distribution is the key of discriminating samples class boundary. In order to talk about the defects of traditional MS, we give the hypothesis as flaws:

The size of sample is n, the weight is 1/n; all samples are divided into two classes (Xa, Xb),X=(Xa, Xb); the kernel function is Gauss truncated function and the expression is seen as follows:

$$k(x)=\begin{cases} e^{-x} & x\in[0,h] \\ 0 & x>h \end{cases} \tag{8}$$

The expression (8) can select some samples which are used to calculate the new center of probability density and the kernel is satisfied by $\int k(x)dx = 2(1-e^{-h}) < \infty$. The samples set of Gauss truncated function is composition of some samples which similarity distance with xk is less than h ; the set is shown by $X_k{}^h = \{x_k{}^{(1)}, x_k{}^{(2)}, ..., x_k{}^{(n_k)}\}$; h is the Gauss truncated threshold; $x_k{}^{(i)}$ is the index of the $X_k{}^h$, $n_k \le n$, $x_k{}^{(i)} \in X$. According to $X_k{}^h$ and the expression (6), the PDF is shown as follows:

$$f(x_i | x_i \in X_k{}^h) = k(\beta \left\| x_k{}^{(i)} - \bar{x}_k \right\|^2) / \sum_{x_j \in X_k{}^h} k(\beta \left\| x_j - \bar{x}_k \right\|^2) \tag{9}$$

Where k(x)=e^{-x} and the center of probability density is calculated by the expression (10).

$$\bar{x}_{k+1} = \frac{\sum_{x_i \in X_k{}^h} w_i k(\beta \left\| x_i - \bar{x}_k \right\|^2) x_i}{\sum_{x_j \in X_k{}^h} w_i k(\beta \left\| x_j - \bar{x}_k \right\|^2)} = \sum_{i=1}^{n} f(x_k{}^{(i)}) x_i \tag{10}$$

For two-classifier, the samples in the $X_k{}^h$ are classified into A_1, A_2, ($X_k{}^h = \{A_1, A_2\}$); and x_k that is the center of probability density belongs to A_1 and X_a and A_1 is called correlated neighborhood set, so the new center of probability density is calculated as follows:

$$\bar{x}_{k+1} = \sum_{x_k{}^{(i)} \in A_1} f(x_k{}^{(i)}) x_i + \sum_{x_k{}^{(i)} \in A2} f(x_k{}^{(j)}) x_j \tag{11}$$

If $A_1 = \phi$ or $A_2 = \phi$, the shift direction of $\bar{x}_{k+1}$ is to the center of probability density of X_a or X_b ,according to convergence of MS[9]. If $A_1 \neq \phi$ and $A_2 \neq \phi$, the shift direction of depends on the value of $\sum_{i=1}^{m} f(x_k{}^{(i)} i) x_i$ and $\sum_{j=m+1}^{n_k} f(x_k{}^{(j)} i) x_j$, which does not depend on the classification that the $\bar{x}_k$ belongs to. So the direction which increased the density is the shifting direction, but it is lack of analysis of local distribution, the shift process will happen to across the classifying boundary. The phenomenon which $\bar{x}_k \in A$ but $\bar{x}_{k+1} \in B$ is called across the classifying boundary. [Fig-5] shows some local distribution which exits risk of the across the classifying boundary.

[Fig-5] shows some local distribution and according to MS, the shifting path is across the classifying boundary. The new center of probability density which is calculated by all samples in the neighborhood moves one class to another class. In order to ignore the risk, the samples local distribution is considered and the new center of probability density is calculated by subset which contains the current center of probability density and this subset is the correlated neighborhood set.

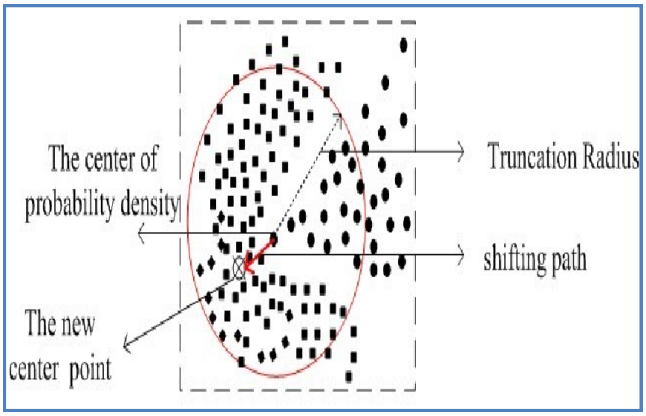

Fig. 5- The scheme of local distribution.

The Correlated Neighborhood Set

The key of LDS-MS is the new center of probability density using correlated neighborhood set, so the import work is to find the correlated neighborhood set from the neighborhood samples. The correlated neighborhood has the similar distribution to the current center of probability density to reduce the risk of across classifying boundary. The analysis method of local samples distribution is graph [13], so in this paper, calculated method is proximity graph and minimum spanning tree. If the Gauss truncated set of $\bar{x}_k$ is $X_k{}^h$, the cor-

related neighborhood set can be calculated as follows:

Caculate the weighted graph G by the distance between sample and sample of X_k^h;

- Convert G into minimum spanning tree T;
- Calculate diameter of tree (T) and sample depth on the diameter;
- According to local minimum depth, divide the sample into two class X_a and X_b;
- If X_a contains the $\bar{x}_k$, X_a is the correlated neighborhood set, else it is X_b.

The correlated neighborhood set save the local sample distributing in the shift process.

The Algorithm Convergence of LDS-MS

The key which influence on convergence of the traditional MS is kernel function and according to section 3.2, the kernel only ensured that the probability density increased, but it doesn't assure the shift path in the same class samples. So it will cause the across classifying boundary. The LDS-MS take the local samples distribution (the correlated neighborhood set) as the prior information of the shift direction, so its convergence is influenced by kernel function and local samples analysis method. Now the algorithm convergence of LDS-MS will be talk about as follows:

X_k^h is Gauss truncated set of $\bar{x}_k$ and the all samples are divided into two class(X_a, X_b). X_a is the correlated neighborhood set. According to the section 3.3, the correlated neighborhood set is X_a' and the rest samples is X_b, and $X_c \in X_a' \bigcap X_a$. According to expression (11), $\bar{x}_{k+1}$ which is the expected new center of probability density is calculated by expression (12) and $\bar{x}'_{k+1}$ which is the new center of probability density by the local samples distribution is calculated by expression (13)

$$\bar{x}_{k+1} = \sum_{x_i \in X_c} f(x_i)x_i + \sum_{x_j \in X_a - X_c} f(x_j)x_j \tag{12}$$

$$\bar{x}'_{k+1} = \sum_{x_i \in X_c} f(x_i)x_i + \sum_{x_j \in X'_a - X_c} f(x_j)x_j \tag{13}$$

So the difference between $\bar{x}_{k+1}$ and $\bar{x}'_{k+1}$ is calculated by follows expression:

$$\Delta = \sum_{x_i \in X_a - X_c} f(x_i)x_i - \sum_{x_j \in X'_a - X_c} f(x_j)x_j \tag{14}$$

Where Δ shows that the difference is between local samples actual distribution and n local samples estimation distribution. The value of Δ is smaller and the risk across the classifying boundary is smaller. The center of probability density converges to same class samples. So the convergence of LDS-MS is depended on local samples distribution estimation modal. The estimation rule is related to less classification rate and over classification.

Algorithm Flow of LDS-MS

1 According to section 3.3 and 3.4, the algorithm flow of LDS-MS is designed as the follows:

1. Randomly select sample as the shifting initial point; k which is the iteration number is 0; h is the radius of Gauss truncated set.
2. According to section 3.3, the X_a which is correlated neighborhood set of x is calculated.
3. According to section X_a, the PDF is constructed and $\bar{x}_k$ is calculated by expression (11).
4. $d = \left\| x - \bar{x}_k \right\|$,if $d<t$, the flow goto step (5) else $x = \bar{x}_k$, $k=k+1$ and goto step (2).
5. Take the samples which is shift the same $\bar{x}_k$ as same class and by which the samples are classified.

The Analysis of Result

The Analysis of Classification Accuracy

In this paper, LDS-MS is validated by Dataset-Ⅲ of "BCI Competion 2003". In order to analyze the effect of α and β frequency band in the classification of EEG， the four frequency bands of three electrodes are encoded and the deferent frequency band is the deferent feature component. [Fig-6] shows the classification accuracy of MS and LDS-MS in the GA iteration. The iteration number is 100; the population size is 100; X-coordinate is the iteration times; Y-coordinate is the classification accuracy. The best accuracy of MS is 89.1% and the best accuracy of LDS-MS is 92.1%. The accuracy is increased 3%, which reason is the optimum of class boundary.

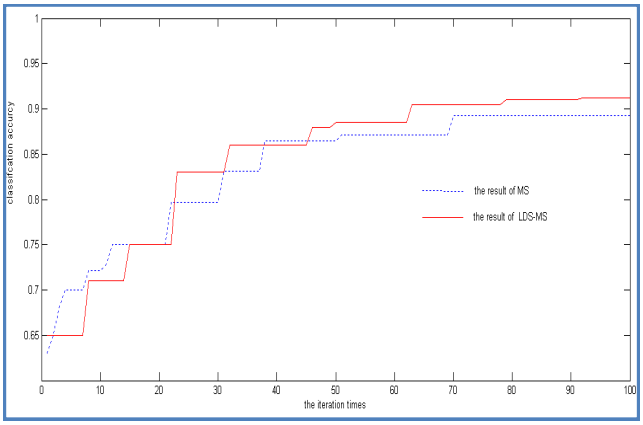

Fig. 6- The classify result of MS and LDS-MS in the GA

The classification performance of LDS-MS depends on the local samples distribution estimation. [Table-2] shows the deferent result of the deferent distribution estimation. The result which [Table-2]

shows is the best result and the kernel function is e^{-x}.

In [Table-2], the K nearest-neighborhood distribution modal be regarded as truncation radius adaptive extension of Gauss truncated function, So the result of the K nearest-neighborhood distribution modal is better than latter. LDS-MS divided the neighborhood samples further and uses the some samples which distribution is same to the center point to calculate the new shifting point, which can optimize class boundary. So the performance of LDS-MS depends on kernel function and neighborhood size. In this research, we use LDA and SVM to classify the EEG based the feature which extracted by section 2.2, the result is 88.1% and 90.0% respectively.

Table 2- The classification accuracy of deferent distribution estimation

Local sample distribution modal	Accuracy
Gauss truncated function	86.70%
K nearest-neighborhood	89.10%
LDS-MS	92.10%

The Analysis of Feature Selection

In this research, the feature is selected by GA which is described by section 2.3. the feature selection is not only used to be reduce the feature dimension, but also to analysis the relation between feature component and classification performance. [Fig-7] show the stat of selecting feature components.

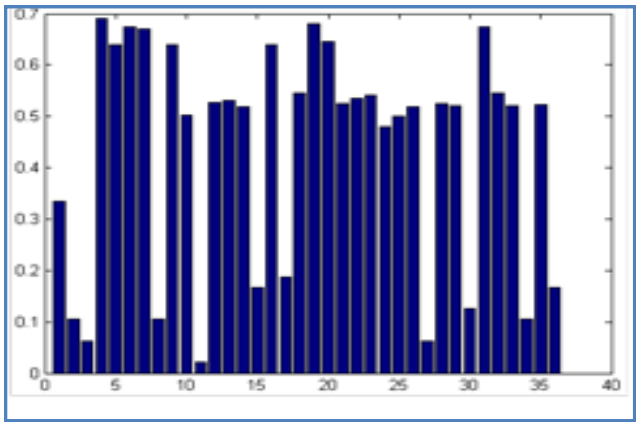

Fig. 7- The statistical histogram of feature selection

Where the x-coordinate is the index of feature components and the y-coordinate is the selected times of feature components (normalization). The result shows that the index of feature components are selected is 4-9, 16-21, 28-33. According to section 2.2, the feature components selected are from frequency bands of 8-32Hz. The result of feature components selection verifies the relation between imagery movement `and frequency band [8Hz-12Hz], [19Hz-30Hz]. So the classification is efficient.

Conclusion

In this research, the local samples distribution is used in the MS to improve the performance of classification. In order to verify the new method, the feature of EEG is extracted by DWT and the classification accuracy of new algorithm is 92.1%, which is better than the result of the BCI competition Dataset-Ⅲ. The result of feature components selected by GA shows the new algorithm is efficient too. The performance of new method depends on feature extraction, kernel function and nearest-neighborhood size, so the future work is to analysis the feature selection algorithm, the relation between the distribution of local samples and the distribution of all samples, which can improve the classification accuracy of EEG.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No.61073115) & (No.61271334)

References

[1] Vidal J.J. (1973) *Annual Review of Biophysics and Bioengineering*, 2, 157-180.

[2] Ali Bashashati, Mehrdad Fatourechi, Rabab K. Ward (2007) *J. Neural Eng.*, 4(2), 32-57.

[3] Lotte F., Congedo M., Lecuyer A., et al. (2007) *J. Neural Eng.*, 4(2), 1-13.

[4] Ling Guo, Daniel Rivero, Julián Dorado (2011) *Expert Systems with Applications*, 38(7), 10425-10436.

[5] Kiymik M.K., Güler I., Dizibüyük A., et al. (2005) *Comput. Biol. Med.*, 35(7), 603-616.

[6] Tapan Gandhi, Bijay Ketan Panigrahi, Sneh Anand (2011) *Neurocomputing*, 74(17), 3051-3057

[7] Clodoaldo A.M. Limaa, André L.V. Coelhob (2011) *Artificial Intelligence in Medicine*, 53(2), 83-95.

[8] Cheng Y.Z. (1995) *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(8), 790-799.

[9] Deng Wang, Duoqian Miao, Chen Xie (2011) *Expert Systems with Applications*, 38(11), 14314-14320.

[10]Shiliang Sun, Changshui Zhang, Yue Lu (2008) *Pattern Recognition*, 41(5), 1663-1675.

[11]Li Xiang-Ru, Wu Fu-Chao, Hu Zhan-Yi (2005) *Journal of Software*, 16(3), 365-374.

[12]Limei Zhang, Songcan Chen, Lishan Qiao (2012) *Pattern Recognition*, 45(3), 1205-1210.

[13]Subasi A. (2005) *Expert Systems with Applications,* 28, 701-11.