Sensors & Transducers, Vol. 159, Issue 11, November 2013, pp. 415-421

Sensors & Transducers

© 2013 by IFSA

http://www.sensorsportal.com

Vector Quantization Codebook Design and Application Based on the Clonal Selection Algorithm

[1] **Mengling Zhao,** [2] **Hongwei Liu**

[1] School of Science, Xidian University, Xi'an University of Science and Technology, 710054, China

[2] A School of Science, Xidian University, 710054, China

[1] Tel.: 13363927092, fax: 029-85509942

E-mail: zhaomengling@126.com

Received: 22 November 2013 /Accepted: 29 November 2013 /Published: 30 November 2013

Abstract: In the area of digital image compression, the vector quantization algorithm is a simple, effective and attractive method. After the introduction of the basic principle of the vector quantization and the classical algorithm for vector quantization codebook design, the paper, based on manifold distance, presents a clonal selection code book design method, using disintegrating method to produce initial code book and then to obtain the final code book through optimization with the clonal selection cluster method based on the manifold distance. Through experiment, based on manifold distance, compared the clonal selection codebook design algorithm (MDCSA) with the hereditary codebook design algorithm and LBG algorithm. According to the result of the experiment, MDCSA is more suitable for the evolution algorithm of the image compression.

Keywords: Vector quantization, Codebook design, Clonal selection, Manifold distance, Image compression.

1. Introductoin

Along with the high-speed development of computer and digital communication technology, people need to transfer more and more image information which contains a large amount of redundant information. The purpose of image compression coding lies in eliminating all kinds of redundancy with the premise of ensuring the quality of the image and in representing and reconstructing image by using as few bits as possible in a given distortion conditions. Research on image compression coding is considered to be one of the most active fields in the information technology. In 1948, Olive proposed TV signal digitization in the field of image compression coding technology, and then a lot of new compression methods and international standards were put forward successively. In the early 1960s and its metaphase, there appeared the earliest theory on vector quantization, which gradually became perfect in the 1980s. Steinhaus, in 1956, first systematically developed the problem about the best vector quantization. In 1978, Buzo first put forward the actual vector quantizer. Until 1980 Linde, Buzo and Gray introduced clustering algorithm into the vector quantizer design and proposed a famous algorithm in vector quantization codebook design - LBG algorithm [1-2], which pushed the research and application of the vector quantization technology to a climax and which became a milestone in the development of this field. Although some achievements have been made in the research on the basic vector quantization technology, there still exist many problems. Therefore, a long way is ahead in the research on vector quantization technology.

2. Relevant Background

2.1. Theoretical Basis of Vector Quantization

Vector quantization is based on the theory of Shannon rate distortion [3]. In 1959, Shannon defined the rate-distortion function R (D) as the minimum code rate coding system can achieve in a condition not beyond a given distortion D, and defined distortion rate function D (R) as the minimum distortion coding system can achieve in a condition not beyond a given code rate R. Based on vector quantization, coding performance can be close to the rate-distortion function. Rate distortion theory suggests that vector quantization is always superior to the scalar quantization because vector quantization can effectively apply the mutual correlation properties of components of the vector to eliminate the redundancy of data. However, rate distortion theory is a theory of existence rather than a structural one for it does not specify how to structure vector quantizer.

2.2. The Definition of Vector Quantization

Vector quantization is to replace the value of a group of input samples (input vector) with the best-match set of output value (code word) of the output group set (code book) in the quantitative process based on the relationship between the adjacent sampling. The advantage of vector quantization lies in its simple decoding and the high ratio of compression.

The basic vector quantizer can be defined as a mapping Q from European space R^K to a limited subset C : $Q\colon R^K \to C$. $C = (\hat{x}_i; i = 1, 2, \cdots, N)$ is called code book, $\hat{x}_i$ is called code word and N is code word size. The mapping satisfies $Q(x) = \hat{x}_p$, and meets:

$$d(x, \hat{x}_p) = \min_{1 \le i \le N} d(x, \hat{x}_i) \tag{1}$$

where $d(x, \hat{x}_i)$ is the measure for the similarity between x and $\hat{x}_i$.

The basic coding and decoding process of vector quantization is shown in Fig. 1. According to certain measure of similarity, vector quantization encoder will search out the minimum code word between distortions of input vector in the code book and only transmit the index of this code word in transmission. Vector quantization decoding process is very simple. First, find out the code word in the code book according to the received code word index and then will it as the reconstruction vector of input vector.

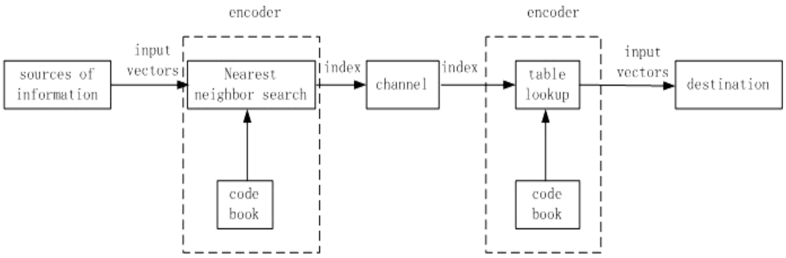

Fig. 1. Diagram of VQ encoding and decoding

2.3. Performance Indicator

The distortion between the input signal and the reconstructed signal is often described by using mean square error (MSE), signal-to-noise ratio (SNR) and peak signal-to-noise ratio (PSNR) [4] to describe. MSE，SNR and PSNR are defined as follows:

$$MSE = \frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(x_{ij} - y_{ij}\right)^2}{M \times N} \tag{2}$$

$$SNR = 10 \times \log_{10} \frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} x_{ij}^2}{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(x_{ij} - y_{ij}\right)^2} \tag{3}$$

$$PSNR = 10 \times \log_{10} \frac{(L-1)^2}{MSE} \tag{4}$$

where $M \times N$ is the size of gray image, L is the grayscale, x_{ij} For the pixels of original image, y_{ij} is the pixels of reconstruction image. $0 \le i \le M-1$, $0 \le j \le N-1$. The PSNR is used in this article.

2.4. The Key Technology of Vector Quantization

The key technology of vector quantization includes three aspects: code book design algorithm, code word search algorithm and distribution algorithm of code word index.

2.4.1. Code Book Design

The first problem of vector quantization is to design a code book with good performance because the performance of code book directly determines the quality of the reconstructed vector and the performance of vector quantization system. Code book design process is actually seeking an optimal scheme to divide M training vectors into N classes in the condition of minimum average distortion. All kinds of centroid vectors are the code words. The purpose of the code book design is to seek an effective method to make code book close to or reach optimum and reduce the computational complexity.

2.4.2. Code Word Search

For a given input vector, code word search algorithm of vector quantization can be used to search the code word with minimum distortion among the input vectors in the code book already designed. In the given code book $C=(\hat{x}_i;i=1,2,\cdots,N)$, $\hat{x}_i$ is the code word and N is the size of code book. The code word with minimum distortion meets $d(x,\hat{x}_p)=\min\limits_{1\le i\le N} d(x,\hat{x}_i)$.There is a high complexity of calculation in searching the input vector to the end, so the code word search's main purpose is to seek the rapid and effective algorithm to reduce the computational complexity.

2.4.3. Code Word Index Distribution

In vector quantization coding system, channel noise often causes inconsistency between the code word from the sending end and that from the receiving end. The distribution algorithm of code word index can redistribute the codeword index and reduce the distortion. The local optimal search algorithm is often used now and to seek the optimal or near global optimal algorithm has become a new branch in the field of vector quantization technique.

3. The Algorithm of Clonal Selection Code Book Design Based on the Manifold Distance

This algorithm is based on the clonal selection code book design and each antibody is a code book in a given antibody population. LBG algorithm is more sensitive to the initial code book, so clonal selection algorithm [5-6] is introduced in this paper. by which splitting method will be applied to generate the initial code book. That is, all of initial antibody forms the initial antibody population. The algorithm of clonal selection code book design based on the manifold distance can effectively search out the global optimal solution — the optimal code book.

3.1. The Formation of Initial Antibody Population

In this algorithm, real number coding mode based on the clustering center is adopted for the initial antibody group and each antibody represents a code book. If a training vector group X and the size of the design code book M are set, splitting method will be used to initialize code book. First, the whole training sample will be taken as a cell, which will be divided into two and two into four via the optimal partition hyperplane until it becomes M cells. They stand for the size of the code book M , which is regarded as the clustering number to generate an initial code book consisting of M code words. According to the size of the antibody population, the same method will be used to generate Size of the antibody population, every antibody is the initial code book as a need to optimize a final code book.

The specific steps of the splitting method:

1) Calculating the center of mass of the training samples X and it is set as C_0 —— the first code word;

2) Choosing disturbance vector δ , taking $\{C_0-\delta,C_0+\delta\}$ as the initial code book and applying LBG algorithm to design a code book containing only two code words;

3) Taking $\{C_1-\delta,C_1+\delta,C_2-\delta,C_2+\delta\}$ as the initial code book and applying LBG algorithm to design a code book containing only four code words.

Through repeating the above steps and designing $\log_2^M$ times, the initial code book consisting of M code words required can be obtained. The calculation method for disturbance vector is: $\delta=a\sigma=(a\sigma_1,a\sigma_2,\cdots,a\sigma_k)$. $a>0$ is the scale factor σ_i is standard variance of the i^{th} component among the training vectors; $i=1,2,\cdots,k$; k is the dimension of training vectors. In this algorithm, for a diversity of the antibody population, the first code word will be generated at random when the splitting method is applied to produce the antibody and the other steps are the same as those of the standard splitting method.

3.2. The Algorithm of Clonal Selection Code Book Design Based on the Manifold Distance

3.2.1. Clonal Selection Theory [7-9]

Due to the gene mutation of genetic and immune cell proliferation, a diversity of immune cells is formed, the proliferation of which will become a cloning system. In 1958, Burnet and other researchers proposed the theory of immune clonal selection, which argues that lymphocytes, in addition to being amplified or differentiated into plasma cells, can also be differentiated into B memory cells with longer life. Meeting the corresponding antigen, memory cells will be selected in advance by the immune system, being rapidly activated, proliferated and differentiated into cells of antibody generation and implementing efficient and durable immune function. Clonal selection is a dynamic process during which organism immune system tries to adapt itself to antigenic stimulation.

Inspired by the clonal selection theory, De Castro and other researchers put forward a kind of clonal selection algorithm (CSA), which is a new approach for artificial immune system based on the biological characteristics of antibody clonal selection. With the aid of antibody clonal selection mechanism of the immune system, Clonal selection algorithm structures clone operator suitable for artificial intelligence. In terms of clone operator, clonal selection algorithm is a kind of group search strategy characterized by parallelism and randomness of the search change, which helps it avoid falling into local optimal value and which can get the global optimal solution with the larger probability. Du Haifeng and others put forward the algorithm of immune clone strategy mainly suggesting the clone operator consisting of three steps: cloning, clone variation and clonal selection. In the antibody group $A(t) = \{A_1(t), A_2(t), \cdots, A_N(t)\}$, the state transition can be expressed in a random process as following:

$$C: A(k) \xrightarrow{\text{clone}} A'(k) \xrightarrow{\text{mutation}} A''(k) \xrightarrow{\text{selection}} A(k+1) \tag{5}$$

3.2.2. Production of Antibody Coding and Initial Antibody Group

Real number coding mode on the basis of clustering center is adopted in the Clonal selection algorithm. In n data samples, dimension d, every antibody A_i is composed of k clustering centers, which can be as a real number coding with the length $l = k \times d$. An antibody can be shown below:

$$A_i(t) = \{\underbrace{a_{11}a_{12}\cdots a_{1d}}_{c_1}\cdots\underbrace{a_{i1}\cdots a_{id}}_{c_i}\cdots\underbrace{a_{k1}a_{k2}\cdots a_{kd}}_{c_k}\} \tag{6}$$

In this equation, the corresponding coding of $c_1, c_2, \cdots, c_k$ is the coordinate for the clustering center in the sample space. The results of K-MEANS iterated 5 times will be applied to initialize antibody group. Of course, using random number to initialize antibody group is also feasible but the accuracy and convergence speed are both inferior to the results of K-MEANS iterated many times. More times of K - MEANS iteration cannot improve the clustering results, but takes more time.

3.2.3. Clone Operator

3.2.3.1 Cloning Operation

In order to control the optimal antibody proliferation, first antibody affinity and dimension of antibody population will be arranged according to the descending order. The corresponding antibody population also arranged by the descending order of antibody affinity degree sequence, and then the equation (3) is used to calculate the number of each optimal antibody to be cloned. Because the individual cloning number is relevant to antibody affinity and the values of it, the lower the antibody affinity and the value are, the smaller the number of cloning will be.

$$n_i = round\left(\frac{\beta \cdot n_b}{i}\right), \quad 1 \le i \le n_b, \tag{7}$$

where $round(\cdot)$ is the integer function, β is the cloning coefficient and is a constant, n_b is the total number of individuals selected to be cloned.

3.2.3.2 Clone Variation Operation

According to immunology, the improvement of antibody affinity and the appearance of antibody diversity mainly rely on high frequency variation of antibody but not on cross or reorganization. Hence, different from the general genetic algorithm suggesting that cross is a major operator and mutation is background operator, clonal selection algorithm puts more emphasis on the role of variation. Variation mimics gene mutation in natural biological evolution and it chooses several gene mutations in one or several genes of the antibody with a certain probability, randomly changing the gene value so as to change the structure of the antibody. In this paper the uniform mutation operator will be adopted and the operation process is: for each variation point, a random number will be chosen from the value range of the corresponding gene to replace the original gene value. That is,

$$P' = U_{\min} + r(U_{\max} - U_{\min}) \tag{8}$$

r is the random number within the scope of $(0,1)$; $U_{\max}, U_{\min}$ are the upper and lower limits of the value of this gene.

3.2.3.3 The Clonal Selection Operation.

Different from the options in genetic algorithm, clonal selection operation chooses the preeminent individuals from the cloned progenies of each antibody to form a new population—a neuter selection process.

3.2.3. Fitness Function

Manifold distance measure is adopted in the fitness function of the algorithm and the specific approach is described as below [10-15]:

Data clustering is characterized by local and global consistency. Local consistency refers to the high similarity between the adjacent data points in the space position; global consistency refers to the high similarity between the data points in the same manifold. Similarity metric based on Euclidean distance can only reflect the local consistency of clustering results — the high similarity of the adjacent data in the space position, but cannot reflect the clustering global consistency—the high similarity located in the same manifold sample. A simple example can illustrate this as is shown in Fig. 2. The similarity between data point a and data point e is expected to be larger than the similarity between data point a and data point f . So just using Euclidean distance as the similarity measure will seriously affect the performance of the clustering algorithm.

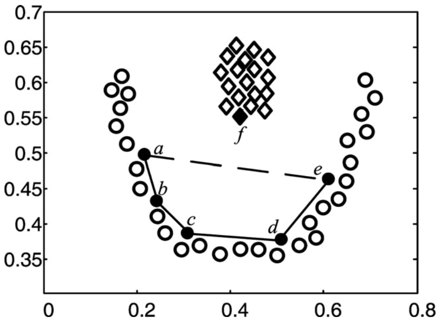

Fig. 2. Illustration of European measure unable being to meet the globally consistency.

In order to guarantee the performance of the clustering algorithm, meeting the distance of clustering global consistency does not necessarily mean meeting the triangle inequality of European measure. In other words, meeting the distance of clustering global consistency cannot definitely lead to the shortest direct path between two points. The path length in the same manifold connected with the short side must be made shorter than the distance between two points directly connected through the low density area, as is shown in Fig. 2.: $\overline{ab}+\overline{bc}+\overline{cd}+\overline{de}<\overline{ae}$. To achieve this objective, first a manifold line length will be defined.

Definition 1: Manifold Line Length

$$L(x_i,x_j)=\rho^{d(x_i,x_j)}-1 \tag{9}$$

$d(x_i,x_j)$ is the Euclidian distance between x_i and x_j ; $\rho>1$ is the expansion factor.

Definition 2: Manifold Distance Measure

The data point is taken as the culmination of graph $G=(V,E)$; $p\in V^l$ represents the path between the two connections p_1 and $p_{|p|}$ with the length $l=|p|-1$ in the figure; the side is (p_k,p_{k+1}), $1\le k<|p|$. P_{ij} stands for the set of all paths of connection data points x_i and x_j ; the manifold distance between x_i and x_j can be calculated via the following equation:

$$D(x_i,x_j)=\min_{p\in P_{i,j}}\sum_{k=1}^{|p|-1}L(p_k,p_{k+1}) \tag{10}$$

Obviously, manifold distance measure meets four conditions of measure, namely, symmetry: $D(x_i,x_j)=D(x_j,x_i)$, non-negativity: $D(x_i,x_j)\ge 0$: triangle inequality: for any x_i,x_j,x_k , $D(x_i,x_j)\le D(x_i,x_k)+D(x_k,x_j)$, self-reversity: $D(x_i,x_j)=0$, if and only if $x_i=x_j$. Manifold distance measure can measure the shortest path along the manifold and this makes the two points in same manifold can be connected by many short sides. At the same time, the two points located in different manifolds should be connected by the relatively longer side, finally achieving the purpose of lengthening the distance between data points in different manifolds and shortening the distance between data points in the same manifold. Based on the manifold distance, a fitness function of calculation similarity can be designed and the fitness function of P_i(the ith antibody):

$$f_{P_i}=\frac{1}{1+Dev(C)}=\frac{1}{1+\sum_{C_k\in C}\sum_{i\in C_k}D(i,c_k)} \tag{11}$$

C is the center set of all clustering; c_k is the class center of clustering C_k ; $D(i,c_k)$ is the manifold distance between data point i of clustering C_k and the clustering c_k .

3.2.4. The Process of the algorithm Of Clonal Selection Code Book Design Based on the Manifold Distance

Specific process of the algorithm is listed below:

Step 1: initialize antibody population. The scale of antibody group is set as N ; $t=0$; splitting method is used to initialize the code book and each code book represents an antibody group; a total of N initial antibodies is produced to form an initial antibody population $C(t)$.

Step 2: According to the clustering center code of the antibody, each sample point will be grouped into different clustering via manifold distance measure based on the nearest neighbor principle; antibody affinity and degree value is calculated.
Step 3: Cloning operation. After clonal variation operation and Clonal selection operation, a new generation of antibodies $C(t+1)$ is generated.

Step 4: Calculating fitness of the updated antibody population.
Step 5: $t = t + 1$. If the clustering error rate is less than stop threshold e, or reaches to the upper limit of iterations ($t$max), the progress will terminate and output the optimal antibody-optimal code book. Otherwise, the process will turn to Step 3.

4. The Analysis of Experimental Results

4.1. Experimental Setup

In the experiment, the main parameters are shown in Table 1 and Table 2 below.

Table 1. Parameters of the algorithm of genetic code book design selection matrix rules.

Parameters	Genetic Algorithm
Maximum iteration	100
Population size	40
Termination condition	$\varepsilon = 1e-5$
Crossover probability	$p_c = 0.8$
Mutation probability	$p_m = 0.5$

Table 2. Parameters of the algorithm of clonal selection code book design.

Parameters	Clonal Selection Algorithm
Maximum iteration	100
Population size	40
Termination condition	$\varepsilon = 1e-5$
Cloning coefficient	$\beta = 3.5$
Numbers to be cloned	$n_b = 5$
Mutation probability	$p_m = 0.5$

4.2. Analysis of the Results

In order to validate the performance and effectiveness of the algorithm, in this paper the four pictures of natural image are compressed and reconstructed. These pictures and their size are shown in Fig. 3. In the experiment, the three pictures of boat , bird, Barbara and build with eight bits per pixel produce code books of all sorts of size as the training images, with their respective vector dimension 16 (4x4). In Fig. 3, peak signal-to-noise ratios (PSNR) produced when the code books of different sizes generated by all kinds of algorithms (based on the manifold distance clonal selection algorithm (MDCSA), based on the manifold distance genetic algorithm (MDGA) and the traditional LBG algorithm) perform image coding on the four pictures respectively are compared.

According to Table 3, for the four images tested, PSNR value of MDCSA algorithm is much higher than that of MDGA and LBG algorithm when the code book size is 128; when the code book size is 256, PSNR value of MDCSA and MDGA is the same, but it is much higher than that of LBG algorithms, when the code book size is 512, in addition to PSNR value of MDGA algorithm of fruit image is high, PSNR values of MDCSA algorithm of others is much higher. Therefore, a conclusion can be reached that the method of clonal selection code book design based on the manifold distance is characterized by better performance and effectiveness.

(a) Boat (512×512) (b) Bird (481×321)

(c) Barbara (512×512) (d) Build (481×321)

Fig. 3. Images tested.

Table 3. Image tested PSNR (dB).

Images	**Code book size**	**MDCSA**	**MDGA**	**LBG**
Boat	128	**24.943**	24.8419	20.1869
	256	**24.8418**	**24.8418**	24.5579
	512	**24.8431**	**24.8431**	24.565
Bird	128	**23.45**	23.3815	19.5655
	256	**23.3781**	**23.3781**	23.1114
	512	23.385	**23.3859**	23.1495
Barbara	128	**23.1356**	23.0691	19.4034
	256	**23.0688**	**23.0688**	22.8644
	512	**23.0703**	23.0701	22.8773
Build	128	**25.1284**	25.058	20.0653
	256	**25.12**	**25.12**	24.5258
	512	**25.1421**	25.1419	24.5186

5. Conclusion

This paper is mainly about the image compression problem of image clustering, in which the basic principles of vector quantization and classic LBG algorithm are introduced. Besides, a method of clonal selection code book design based on manifold distance is put forward. During the process, splitting method is adopted to produce initial code book and clonal selection clustering approach based on manifold distance is applied to generate the final code book. In the experiment, algorithm of clonal selection code book design, algorithm of genetic code book design and the traditional LBG algorithm are compared. The results show that for code books of different sizes relatively high PSNR value can be achieved via MDCSA, which proves that this algorithm has good performance and effectiveness.

Acknowledgements

This work is supported partly by the National Natural Science Foundation of China under Gran No. 61072144 and No. 61179040.

References

[1]. B Ramamurthi, A Gersho, Classified Vector Quantization of Images, *IEEE Trans. Communication*, Vol. 34, No. 11, 1986, pp. 1105-1115 .

[2]. K. W. Chan, K. L. Chan, Subband VPIC with Classified Joint Vector Quantization, *Signal Processing: Image Communication,* Vol. 13, No. 13, 1998, pp. 145-153.

[3]. Dong-Li, Chang Souguil Ann, Choong Woong Lee, A Classified Split Vector Quantization of LFS Parameters, *Signal Processing*, Vol. 59, No. 3, 1997, pp. 267- 273.

[4]. Rafael C. Gonzalez, Richard E. Woods, Steven L. Eddins, Digital Image Processing, *Publishing House of Electronics Industry*, China, 2008.

[5]. Yingchun Zhang, Juan Cao, Bohong Su, Image Registration Based on Genetic Algorithm, *JCIT: Journal of Convergence Information Technology*, Vol. 8, No. 8, 2013, pp. 37 - 45.

[6]. Zhou Yang, Ma Li, Bai Lin, Clustering Method with Evolutionary Immune Network Based on Polyclonal Algorithm, *Computer Engineering and Applications*, Vol. 45, No. 27, 2009, pp. 146-150.

[7]. Zhu Feng, Zhang Qun, Bai You-Qing, et al., A Novel Reconstruction Method Based on Genetic Algorithm in CS Theory and Its Application in SAR HRRP reconstruction, *Control and Decision*, Vol. 27, No. 11, 2012, pp. 1669-1675.

[8]. Liu Ruo-Chen, Immune Clonal Strategy Algorithms and Their Appligcation, *Xidian University*, 2007.

[9]. Wang Juan, Li Fei, A Quantum-inspired Immune Clonal Algorithm Based on Real-Encoding, *Computer Engineering*, Vol. 38, No. 18, 2012, pp. 133-136.

[10]. Dong Hong-Bin, Yang Bao-Di, Liu Jia-Yuan, et al., A Co-Evolutionary Algorithm for Encoding, *Paghtedttern Recognition and Artificial Intelligence*, Vol. 25, No. 04, 2012, pp. 676-683.

[11]. Zhang Ying-Jie, Fan Chao-dong, Improved Clonal Immune Algorithm for All Solutions of SAT Based on Multi-Population, *Information and Control*, Vol. 40, No. 01, 2011, pp. 34-38.

[12]. Yang Rui-Rui, Niu Jian-Qiang, Meng Hong-Fei, Pavement Crack Extraction Using Iterative Clustering Algorithm Based on Manifold Distance, *Computer Engineering,* Vol. 37, No. 12, 2011, pp. 212-214.

[13]. Kelvin Lo Yir Siang, Siak Wang Khor, Path Clustering using a Mixture of Dynamic Time Warping Technique and Cluster Searching Index Method, *IJIIP: International Journal of Intelligent Information Processing*, Vol. 3, No. 4, 2012, pp. 1- 10.

[14]. Yang Jin-Fu, Song Min, Li Ming-Ai, A Novel Reduction K-Nearest Neighbor Classification Method Based on Weighted Distance, *Journal of Electronics & Information Technology*, Vol. 33, No. 10, 2011, pp. 2378-2383.

[15]. Song Dan, Lai Xu-Zhi, Wu Min, Clonal Selection Algorithm Based on Grade Variation, *Paghtedttern Recognition* and Artificial Intelligence, Vol. 24, No. 03, 2011, pp. 438-443.