Clustering Time Series by Wavelet and Evolutionary Computing

Antonio Ramírez, Ricardo Barrón, Giovanni Guzmán, Edgar. A. García

Artificial Intelligence Laboratory, Center for Computing Research
National Polytechnic Institute, Mexico
{tonotron, barron3121, giovanni.guzman, zavael}@gmail.com

Abstract. The classification of information can take various hues, depending on the approach to be performed [1], however to achieve this classification is one of the most important tasks of computational science and several of its branches [2]. In this paper a classification of time series by calculating numerical wavelet coefficients is presented, since it has been shown to be a very good tool for signal analysis [3]. A new approach in which tuning parameters of a classification function are optimized in a supervised way by an evolutionary algorithm proposed is also presented. With the use of different wavelets the best way to get descriptors for each data set was deducted, then sorted and grouped by the use of the evolutionary algorithm, combining these two tools achieved a better classification of the information presented in numeric strings, considered as time series.

Keywords. Wavelets, classification, time series, evolutionary computing, metaheuristics.

1 Introduction

Over the last five decades, preferences concerning with evolutionary algorithms have been growing. This framework, offers a wide set of techniques for solving the problem of searching optimal values by using computational models inspired by Nature, particularly evolutionary processes, the called Evolutionary Algorithms (EAs) which are population based optimization techniques and are designed for searching optimal values, in big data sets [5]. On the other hand, the signals analysis has been optimized thanks to employ of wavelets, which are considered one of most accurate ways to locate minutiae over signals [4]. Techniques, EA and wavelets are the most robust methods to acquire accurate information from a signal, which able us, to make a better interpretation of such data [16]. This paper applies the combination of EA and wavelets to generate a spectral signature descriptor based on the wavelet calculation [2], taking into account the time series of productivity of some oil wells. The information of such wells, which is grouped by a bio-inspired PSO algorithm [3], is the basis for characterization and classification.

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 33–44
Paper Recived 13-09-2013 and Acepted 21-10-2013

2 Theoretical backgrounds

This section provides some preliminaries about the state of the art of the tools used in this work. The different methods or algorithms for clustering, and their relationship with the time series, are also treated. It will be established some clustering algorithms from single to widely used, even the most modern and sophisticated based on evolutionary algorithms. Finally, we present a brief introduction to wavelet theory, which combined with meta-heuristics, serve to optimize the time carried out by the clustering algorithm. Also we conducted a comparison against the brute force algorithm proposed by [17], for classification of big data sets, in which good results were obtained.

2.1 Clustering

One of the best ways to handle large concentrations of data is clustering. A cluster is a collection of elements that share common characteristics, which can be separated and also a part of a group. The study area now known as clustering, is a method of unsupervised learning, and such clustering without a previous pattern separation or preclassification. One of the first clustering algorithms is the classic K-means [14], which is quick and easy to be implemented. In contrast, such algorithm is also very sensitive to the initialization of the centroids and easily falls into the so-called "local optimum". A less sensitive algorithm initialization was proposed by Zhang, B. [15] called K-harmonic means (KHM), on a technical report for HP Labs, and this algorithm employs harmonic means for calculating centroids and is in some cases more efficient. For its part, the classical Pattern Recognition algorithm, also been successfully applied for the definition of various strategies for regulating the exploration of large search spaces.

Aforementioned algorithms are basically processes that partitions a population into k n-dimensional sets, based on calculations of average distance between each of the elements and the number of partitions obtained according to its variance. The difference between them is that one is based on the calculation and their means distances (KM) and the other based in their harmonic means (KHM), in order to determine their membership in a given cluster. Clustering is considered as an unsupervised learning method, and it is commonly used for the analysis of statistical data in many disciplines, including data mining, pattern recognition, image analysis, machine learning and bioinformatics to mention only some [3 7 8].

2.2 Particle swarm optimization

Particle Swarm Optimization (PSO) is a meta-heuristic based on populations of individuals as those found in nature, as swarms of ants, banks of fishes, flocks of birds, etc. The algorithm developed in 1995 by Kennedy and Eberhart, is based mainly on psychosocial approach known as "social metaphor" [16], PSO algorithm can be summarized as: "every individual can change their opinion and participate or not in a group, with based on three factors, environmental knowledge, experience, and

knowledge of experiences of individuals in his neighborhood. This individual then adapts his schemes, beliefs and opinion in accordance with individuals who have better experiences in their environment and based on interaction rules laid down"[13].

The heuristic techniques for exploring large search spaces, used by these two disciplines have been used successfully to define data clustering algorithms, allowing troubleshooting data clustering and classification of large volumes of data, or in a dynamic context [8].

Using heuristic algorithms for defining cluster has some important advantages over the use of traditional construction algorithms [7]. The use of heuristics requires a programmer to stake a grouping or classification problem as one of optimization (either looking for the optimal position of the centroid of each group, or optimizing the topological properties of the final clustering) which in turn permits to define clustering algorithms that provide a certain probability of quality response as data volume expands.

2.3 Wavelets

The wavelets are short wave functions with compact support that can represent both a signal or time series in either the time or the frequency domain [5]. Unlike the Fourier transform, wavelet transform consists of a set of basic functions, often called mother functions, which allows a local description of the behavior in frequency (spectrum of the signal). Moreover, wavelets can capture transient data and not only the average frequency behavior, as occurs in the Fourier transform.

The wavelet transform of a function $\varphi(t)$, corresponds to the decomposition of that function, in a group of daughters functions with the form $\psi_{(m,n)}(t)$, called wavelets. As it occurs in the continuous case, there is also the wavelet transform for the discrete case that can be defined as.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\,\psi\left(\frac{t-b}{a}\right) \tag{1}$$

where b ∈ R is a translation parameter and ∈ R, (a > 0), is regarded as a scale factor. In the development of this work, tests were conducted with the families of Haar wavelet [6] and Daubechies [7], using the criterion of energy minimum variation. This criterion refers that the coefficients resulting from the scaling factor of any signal must preserve energy from the signal.

2.4 Time series properties

For concreteness, we begin with a definition of our data-type interest of a time series. Definition: A time series $T = t_1, \ldots, t_j$ is an ordered set or not, of j real-valued elements. Assume a time series $\vec{X}\ (\vec{X} \in \mathbb{R}^n)$ is located in the J scale, this series can be decompose into a j specific scale $(j \in [0, 1, \ldots, J-1])$, then the coefficients $H_j(\vec{X})$ correspond to the scale j and can be represented by a series $\{\overrightarrow{A_j}, \overrightarrow{D_j}, \ldots, \overrightarrow{D_{j-1}}\}$; where

the $\overrightarrow{A_j}$ must be considered an approximation coefficients which are the projection of $\vec{X}$ in V_j and the $\overrightarrow{D_j}, \ldots, \overrightarrow{D_{j-1}}$ are the wavelet coefficients in $W_j, \ldots, W_{j-1}$ representing the detail information of $\vec{X}$, where W_{j-1} is considered a sub-space or complement of $V_j = V_{j-1} \oplus W_{j-1}$, since by defining the $\varphi_{j,k}$ and $\psi_{j,k}$ are the orthogonal basis of W_j; $\{\varphi_{j,k}, \psi_{j,k}; j \in \mathbb{Z}, k \in \mathbb{Z}\}$ [6 17].

2.5 Spectrar signature

Each element over earth's surface such as forests, crops, rivers, lakes, buildings, plains, human default erosions, etc., have the ability to transform differently the electromagnetic radiation receiving of the sun. Each type of these objects, have a level of specific response and this is measured as spectral signature.

The amount of energy variation that one object can reflect (reflectance), is a function of the wavelength that each emits body and is called spectral signature or spectral firm [11]. The spectral signature is then a quantitative measure of the spectral properties of an object in one or more spectral bands. Also known as spectral behavior. Thereby, any object can vary in quality and / or quantity of their spectral signature, depending on weather conditions, seasons of the year and essentially lighting conditions. Thereby, any object can vary in quality and / or quantity of their spectral signature, depending on weather conditions, seasons of the year and essentially lighting conditions. So it is possible to identify in a image for instance, the organic and physical nature of a particular object. In the analysis of data sets may also obtain a spectral response of the objects that represent such information. It means in a data set, we could identify small variations and to interpret info about them.

3 Proposed algorithm

For spectral characterization, which we call 'signature', calculate the wavelet transform of the signal to identify active frequencies and then choose a representative part of the signal, based on an energy criterion or threshold limit. Taking into account the information provided for the analysis and implementation of present work, we considered a data matrix in which first column is the code name for each well and the rest of the information on each line corresponds to data supplied each of them. In order to apply the proposed wavelet family, it was necessary to standardize the data in a matrix in which the data should be calculated in 2n basis, always considering the importance of using as closely as possible the information provided. Following the next criteria made this:

(a) The data elements of each row, represent one record of monthly production reading date from October 1980 to April 2011 yielding 367 columns of data.
(b) With intention that the data table was uniform and his columns were also a multiple of 2n, this matrix was completed with a zero in the empty positions (before and after the column 368).

(c) When data (columns) nonzero did not exceed 256, that figure was adjusted to apply the wavelet calculation.

The treatment of the lack of information that occurred in nearly 50% of the data matrix, was performed with the algorithm that shows in Table 1.

The heuristic technique used during this first phase of experimentation is Particle Swarm Optimization (PSO) [10]. The choice was due to the proven ability of this technique to explore large search spaces using a relatively small number of particles.

Furthermore, this technique allows to combine various search strategies, both on global and local criteria, which, inevitably leads to the possibility of running the parallel computational model with the consequent increase in performance.

Table 1. Pseudo code of treament of lack of information.

```
1.  Read line
2.    count Data;
3.      if Data > 255
4.        For each blank,
5.              delete blank
6.              if Data <= 256 break
7.            end for
8.      else
9.         while Data < 512 add '0'
10.    end if
11. End
```

When the volume of data is considerably large in size according to the type of information that is analyzed, it is advisable to group the data by calculating the standard deviation typical and obtain uniformity coefficients, which will give rise to patterns leaders or centroid by calculating the variance. Overall building a digital distribution commonly consists of three stages: 1) to determine the "classes" and their intervals, 2) classifying (or distribute) the data in the appropriate classes, and 3) count the number of cases of each class, and set the corresponding "class mark" or centroid.

The algorithm implemented in this work uses two cycles (external and internal) that optimize the number of clusters and clustering quality using a cluster validation index and overall variance to form the groups. The outer loop of the algorithm, which measures the quality of clustering and optimizes the number of groups, used a model of Particle Swarm Optimization. The inner loop uses a hybrid algorithm between microPSO, and local adjustment algorithm. This hybrid algorithm form groups passing to outer loop. The solution that delivers the inner loop is a grouping, which passes to outer loop and this last one, measures the group quality and this way optimizes the number of groups.

3.1 External cycle

The fitness function of PSO algorithm performed by external cycle use the index j, which measures the quality of clustering and thus optimize the number of groups. The index j is described as follows:

$$J(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K\right)^P \tag{2}$$

where K is the number of groups, and P is a real number greater than or equal to 1, that controls the contrast between different cluster configurations. In the above equation E_k y D_k are described as follows:

$$E_K = \sum_{k=1}^{K} \sum_{j=1}^{n} u_{Kj} \|x_j - z_k\| \tag{3}$$

$$D_K = max_{i,j=1,\dots k} \|z_i - z_j\| \tag{4}$$

Where *n* is the total number of points in the data set, z_k is the centroid of *k-th* group, x_j is the *j-th* item from *k* group. The *index I* is a three factors composition called $\frac{1}{K}$, $\frac{E_1}{E_k}$ y D_k. The first factor decreases linearly as the value of *k* increases. Thus, this factor is to reduce the value of the *index i* to increase the value of *k*. The second factor is the radius of E_1, which is a constant value. The factor E_k, decreases when *k* value increases. The *index I* value increases with E_k decreasing. This, indicates that more groups should be formed that should be compact in nature. The third factor (which measures the maximum separation between two groups) will increase with the value of *k*. So while the first factor is decrease the value of *k*, the second and third factors try to increase the value of *k* favoring compact well-separated groups.

3.2 Internal cycle

The hybrid algorithm microPSO and local adjustment, used the fitness function as global variance, which is defined as follows:

$$f(x) = \frac{1}{K} \sum_{j=1}^{K} \sum_{i=1}^{n} D_E(o_i, C_j) \tag{5}$$

where *K* is the group number, *n* is the number of elements belonging to the group; D_E is the function of similarity / difference between patterns; O_i is the *i-th* pattern that belongs to the class *j* and C_j is the centroid of the class. A pseudo code of this proposal is shown in Table 2.

3.3 Comparing algorithms for times series classification

The brute force algorithm let us to appreciate that we simply take each possible sequence of numerical data and find the distance to the nearest (any other), is match and the subsequence that has the greatest such value is the discordant. This is achieved with nested cycles, where the external cycle considers each possible candidate subsequence, and the internal cycle in this case, is a linear scan to identify the candidate's nearest any-other match.

In the Table 3, we can see the pseudo code of the algorithm, by [17]. Such, requires just one parameter, the length of subsequences to consider. The algorithm is easy to

implement and produces exact results. However, the algorithm that we propose has not the excessive spending flaw, for data mining as brute force algorithm which has $O(n^2)$ time-complexity.

Table 2. Pseudo code clustering cycles (external and internal).

```
GenerateInitialPopulation(P1)
Evaluate(P1)
while not ExternalConvergence(P1) do
    foreach particle (p_i) in P1
        GeneratePopulation(P2)
        Evaluate(P2)
        while not InternalConvergence (P2) do
            foreach particle (x_i) in P2
                CalculateNewPosition(x_i)
                LocalAdjustment(x_i)
                Evaluate(x_i)
                ReplaceGlobalBest(P2)
            end foreach
        end while
        SelectInternalBestIndividuals(P2)
        CalculateNewPosition(p_i)
        Evaluate(p_i)
        ReplaceGlobalBest(P1)
    end foreach
end while
SelectBestSolution(P1)
```

Table 3. Pseudocode code Brute Force Discord algorithm.

```
1. Function [dist, loc ]= Brute_Force(T, n)
2.   best_dist = 0
3.   best_local = N
4.   for p = 1 to |T|-n+1
5.     nearest_neighbor_dist = infinity
6.     for q = 1 to |T | - n + 1
7.       if p−q ≥ n
8.         if Dist (t_p,…,t_p+n-1, t_q,…,t_q+n-1) < nearest_neighbor_dist
9.           nearest_neighbor_dist = Dist (t_p,…,t_p+n-1, t_q,…,t_q+n-1)
10.        end if
11.      end if
12.    end for
13.    if nearest_neighbor_dist > best_dist
14.      best_dist = nearest_neighbor_dist
15.      best_local = p
16.    end if
17.  end for
18.Return[ best_dist, best_local ]
```

3.4 Configuration parameters for experimentation

In the development of a configuration, experiments used for all tests, both for internal and external cycle. In Tables 4 and 5 shows the configuration data of their parameters.

Table 4. Configuration parameters of the external-cycle algorithm.

[1] Parameter	[2] Value
[3] Swarm size	[4] 20
[5] Search space dimensions	[6] 1
[7] Inertia factor	[8] 0.9
[9] Cognitive factor weight	[10] 1.2
[11] Social factor weight	[12] 1.4
[13] Limits to search space	[14] [2 N] where N is total number of data
[15] Max velocity	[16] 1.2

Table 5. Configuration parameters of the inner loop hybrid algorithm.

[17] Parameter	[18] Value
[19] Swarm size	[20] 3
[21] Search space dimensions	[22] 2
[23] Inertia factor	[24] 0.9
[25] Cognitive factor weight	[26] 1.6
[27] Social factor weight	[28] 1.8
[29] Limits to search space	[30] [0.0 3428835.7]
[31] Max velocity	[32] 1.5

3.5 Classification of wells using historical data

According to the information that had at that moment, was determined use the periodic information of each record, which in turn represent the wells production, as time series that can be analyzed as with digital signals. It was considered for this purpose, using wave-signal analysis, best known as wavelets analysis. This information should be standardized as a matrix, where each row corresponds to the registration of a well and this in turn should be composed of a number of 2^n-based data.

Thus, a small record an array of data, representing a signal from a well. At this signal (numerical series), was applied to the wavelet calculation for obtain a tuple of representative features, to this tuple (two values) we call descriptor. These descriptors are compared to each other to find the class it belongs to each well. This is determined by a minimum proximity criterion (calculating the Euclidean distance) with a centroid. This centroid, is a representative element for each cluster, is also a tuple of two numerical data. At this stage of development, the centroids are defined as statics.

Once you have found the optimal clusters of wells according to their productivity and their proximity to class centroids is possible to encode these classes and placed it in any of them. The production time series from wells, gives the grouping with formed clumps by algorithm known as ("clustering").

Algorithm to classify a well given their time series:

1. Standardize the sample size
2. Calculating the descriptor or spectral signature using wavelet transform.
3. Calculating the distance of the spectral signature of the well with respect to the centroid.
4. Assign the well label the corresponding class (classification).
5. Perform static grouping of wells (clustering)

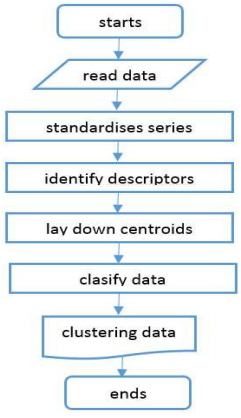

Fig. 1. Total classification algorithm flowchart.

4 Experimental results

To get the pooling of data from time series, were translated the original file format spreadsheet format to a flat file comma separated text (file.csv). Those data organized as a matrix then we applied the calculation algorithm based on wavelet spectral signature.

In the test phase we used two families of orthogonal wavelets: Haar and Daubechies, obtaining very similar results. We decided to use only the calculation of the spectral signature with the Haar, due this one was faster and offered better results.

With the spectral signature (time series analysis of the data) was obtained identification record for each of the wells then information was used as sample learning for clustering algorithm, generating different groups of data and different classes or groups.

Flat file was generated with each sediment records placed (Table 6). The final information is presented in four columns, the number of well, the coefficients calculated with the wavelet, which in this case are considered traits descriptors each well and the number of class to which they belong. These results may not be less than ideal, since the information available to the data file is very limited. In some cases it was necessary to assign a zero to the missing data.

Table 6. View the data analysis and classification.

POZO ID	DESC 1	DESC 2	CLASE
POZO 1	2261009.45965523	645128.454706114	CLASE 3
POZO 2	1114.4373721754	18696.6893687409	CLASE 6
POZO 3	0	17977.5910761071	CLASE 6
POZO 4	1263766.3033614	68687.3132890685	CLASE 5
POZO 5	1263766.3033614	164943.967887151	CLASE 5
POZO 6	0	0	CLASE 4
POZO 7	125.818652279295	1995.80321136723	CLASE 4
POZO 8	162.346705130885	2575.23126180834	CLASE 4
POZO 9	191603.720039493	42381.2007671788	CLASE 7
POZO 10	1205.24246354158	0	CLASE 6
POZO 11	109829.972788095	1006.87036987241	CLASE 6
POZO 12	1103145.18273855	43832.0783253027	CLASE 5

The experimental data was contained in a flat file, was treated as numerical series and subjects to mathematical calculus that result in a table of two columns in which are represented the descriptors of those series. These descriptors are evaluated to know to which class they belong. In Table 7, it shows a list of classes as was obtained from the PSO algorithm, within one of the end files. The final data within on figure 3, correspond to identification from each well, into each class.

The algorithm described above in Table 2 determined these seven types. Fixed data were used in each experiment shown in Table 8. The unique parameter that changed on the experiments was the fitness value.

When we compare the clustering by wavelet and evolutionary algorithm, with brute force algorithm, we can see a substantial difference in the time-convergence of clustering over similar data sets as shown in Figure 2.

Table 7. Classification of centroids obtained from PSO algorithm, a) best fitness values, b) coordinates of centroids.

1. 2214702.150332	**C1** (751072.158462, 196043.994000)
2. 2314530.151289	**C2** (3247609.92500, 423875.33500)
3. 2333896.527851	**C3** (2334484.40000, 330180.57900)
4. 2073020.127509	**C4** (31269.16174, 221666.88895)
5. 2080099.692564	**C5** (1258644.26400, 131729.75784)
6. 2333896.527851	**C6** (16855.87879, 22789.45322)
7. 2338401.671122	**C7** (314861.60150, 20148.04506)

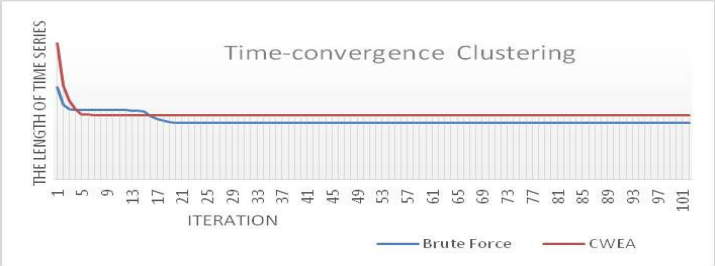

Fig. 2. Comparison of execution time between BF and CWEA algorithms.

Table 8. UGPSO Algorithm parameters.

1. Algorithm:	Ugpso
2. Adjust:	On
3. Groups:	7
4. Generations:	200
5. Popsize:	3
6. Inertia:	0.900000
7. Social:	1.800.000
8. Cognitive:	1.600.000
9. Vmax:	1.500.000
10. Range:	[0.000000, 3428835.700000]
11. PopKeep:	1
12. Fitness value	2073020.127509

5 Conclusions

In this paper, we presented a new approach of clustering, in which the tuning parameters a classification function is being optimized on supervised way by a combined heuristic technique. The proposed PSO-based technique being population-based random, then the search optimization technique does not require initialization of adjusting parameters.

Depending on some circumstances, big data sets would be considered like a temporary signal scannable as any other signal. In signal analysis area, there are several

techniques to identify significant differences that allow us to isolate the unique traits it need to represent a pattern. When it is possible to identify patterns in the signals, they can be sorted and grouped. For this work we considered the information that was counted as time series data as represented.

The technique used in this work was able to separate into classes a list of series, based on calculations with wavelet and using the evolutionary algorithm to optimize a classification function, finding that with the proposed algorithm, it was finally possible to properly clustering the oil wells in seven categories, based on the analysis of time series representing historical record of production.

References

1. T. Babdagli, Development of mature oil fields a review, Journal of Petroleum Science and Engineering, Elsevier, (2007).
2. Qian Tao, Vai Mang, Wavelet Analysis and Applications, Applied and Numerical Harmonic Analysis, Springer Science + Business Media, Germany, (2007).
3. G. Luque, E. Alba, Parallel Genetic Algorithms, Theory and Real World Applications, Studies in Computational Intelligence, Springer-Verling, Spain, (2011)
4. L. Guan, Y. Du, L. Li, Wavelets in Petroleum Industry: Past, Present and Future, Society of Petroleum Engineers Inc. SPE 89952. (2004).
5. K. Premalatha. A New Approach for Data Clustering Based on PSO whit Local Search, CCSE, Computer and Information Science. Vol. l, No. 4, (2008)
6. Hui Zhan, et al, Unsupervised Feature Extraction for Time Series Clustering Usign Orthogonal Wavelet Transform, Informatics 30, (305-319), (2006).
7. Ten lectures on Wavelets. Daubechies I. Society for Industrial and Applied Mathematics (1992).
8. Michael W. Berry, Survery of Text Mining: Clustering, Classification and Retrieval, Springer, (2004).
9. Swagatam Das, Amit Konar, Meta-heuristic Clustering". Studies in Computational Intelligence 178. Springer, (2010).
10. Xin-SheYang, Nature-Inspired Metaheuristic Algorithms. Luniver Press, Second Edition, (2010).
11. Sean Luke, Essentials of Metaheuristics. Second Edition, Lulu Publishing, (2013).
12. Particle Swarm Optimization: Theory, Techniques and Applications (Engineering Tools, Techniques and Tables). Andrea E. Olsson (E). Nova Science Pub Inc, (2011).
13. Kennedy J., Russell C., Yuhui S., (Swarm Intelligence). The Morgan Kaufmann Series in Evolutionary Computation, USA, (2001).
14. MacQueen, J. B, Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California, USA, (1967).
15. Zhang, B., Hsu, M., Dayal, U., K-harmonic means, a data clustering algorithm. Technical Report HPL-1999-124, Hewlett-Packard Laboratories, (1999).
16. Yuehui Chen, et al, A Local Linear Wavelet Neural Network, Proceedings of Congress on Intelligent Control and Automation, Hangzhou, P. R. China, (2004).
17. E. Keogh, J. Lin, HOT SAX: Finding the Most Unusual Time Series Subsequence: Algorithms and Applications. The Fifth IEEE International Conference on Data Mining, USA, (2005).