Pet. Sci. (2015) 12:135–147
DOI 10.1007/s12182-014-0006-5

ORIGINAL PAPER

Fault diagnosis for down-hole conditions of sucker rod pumping systems based on the FBH–SC method

Kun Li · Xian-Wen Gao · Hai-Bo Zhou · Ying Han

Received: 21 October 2013 / Published online: 23 January 2015

Abstract Dynamometer cards are commonly used to analyze down-hole working conditions of pumping systems in actual oil production. Nowadays, the traditional supervised learning methods heavily rely on the classification accuracy of the training samples. In order to reduce the errors of manual classification, an automatic clustering algorithm is proposed and applied to diagnose down-hole conditions of pumping systems. The spectral clustering (SC) is a new clustering algorithm, which is suitable for any data distribution. However, it is sensitive to initial cluster centers and scale parameters, and needs to predefine the cluster number. In order to overcome these shortcomings, we propose an automatic clustering algorithm, fast black hole–spectral clustering (FBH–SC). The FBH algorithm is used to replace the K-mean method in SC, and a *CritC* index function is used as the target function to automatically choose the best scale parameter and clustering number in the clustering process. Different simulation experiments were designed to define the relationship among scale parameter, clustering number, *CritC* index value, and clustering accuracy. Finally, an example is given to validate the effectiveness of the proposed algorithm.

K. Li (✉) · Y. Han
College of Engineering, Bohai University,
Jinzhou 121013, Liaoning, China
e-mail: tubiekun@163.com

X.-W. Gao · H.-B. Zhou
College of Information Science and Engineering, Northeastern University, Shenyang 110819, Liaoning, China

Edited by Yan-Hua Sun

Keywords Sucker rod pumping systems · Fault diagnosis · Spectral clustering · Automatic clustering · Fast black hole algorithm

1 Introduction

Sucker rod pumping systems are the main artificial lift methods in oil production. In practical oil production, monitoring of the down-hole working conditions of the sucker rod pumping systems mainly relies on manual methods. This precludes real-time monitoring and the operating costs are high. In order to monitor down-hole working operations continuously in normal running, it is necessary to use computers to replace the manual work. In oilfields, dynamometer cards are commonly used to analyze down-hole working conditions. Nowadays, many computer diagnosis methods have been used to achieve intelligent identification of the dynamometer cards. These include expert systems (Derek et al. 1988; Martinez et al. 1993), artificial neural networks (Rogers et al. 1990; Xu et al. 2007; Tian et al. 2007a, b; de Souza et al. 2009; Wu et al. 2011), rough set theory (Wang and Bao 2008), support vector machine (Tian et al. 2007a, b; Li et al. 2013a; Yu et al. 2013), fuzzy theory (Li et al. 2013b, c), and designed component analysis (Li et al. 2013b). The current research mainly focuses on supervised learning methods, which rely on manual work to select training samples. Although the supervised learning maybe has higher classification accuracy, it is mainly dependent on high quality training samples. In oil production, this has two shortcomings: firstly, it is easily affected by the subjective experience of the technical staff; secondly, there are larger workloads in manual classification of the training samples. So, in order to reach a higher efficiency of monitoring the down-hole working conditions, the unsupervised learning method is useful to decrease the dependence on the training samples. In this paper, an unsupervised classification method based on the FBH–SC clustering algorithm is discussed.

The clustering algorithm is a commonly used unsupervised learning method, which can realize automatically classification of the data according to their characteristics in a self-learning way. However, many commonly used clustering algorithms (like: K-mean, fuzzy C-means (FCM)) have some deficiencies, such as: (1) the clustering results are sensitive to the shape of the data set; and (2) the clustering number is needed to be set in advance according to the prior knowledge.

The spectral clustering (SC) algorithm is a new type of clustering algorithm based on graph theory. It can be applied to data sets with any distribution shape and is independent of the dimension of the data; the global optimal solution can be obtained using the standard linear algebra method by a clustering criterion in a relaxed continuous domain through the eigen decomposition of the similarity matrix (Von Luxburg 2007). SC is a focus research of the clustering analysis method (Alzate and Suykends 2012; Fujiwara et al. 2012; Mirkin and Nascimento 2012; Tasdemir 2012; Frederix and Van Barel 2013; Lv and Feng 2013). However, it still has some deficiencies, such as (1) the clustering results are sensitive to the initial clustering centers; (2) the algorithm is sensitive to the scale parameter (σ) and (3) the clustering number (k) needs to be set in advance. For the first one, Liu et al. (2012) proposed a method based on fuzzy K-harmonic means which could reduce the sensitivity to the initial centers to a certain extent; however, the local optimal value cannot be solved completely as the algorithm still uses the iterative optimization technique which is similar to the K-mean method. For the second one, the clustering ensemble-based methods (Zhang et al. 2008; Vega-Pons and Ruiz-Shulcloper 2011) are used to solve the problem that the clustering results rely on precise choice of the scale parameters, which uses different scale parameters in a given interval to obtain different clustering results and then merges them by an integrated approach. For the third one, the clustering validity index is used to evaluate the clustering results, of which the maximum value or the minimum value is taken when a most suitable cluster number is obtained (Bezdek and Pal 1998; Breaban and Luchian 2011; Saha and Bandyopadhyay 2012); however, this method also has some deficiencies (Wang et al. 2012), such as (1) exhaustive search of each k may bring huge computational cost and (2) for each k, the global optimal solution of the clustering results cannot be guaranteed.

Now, in current studies of the SC algorithm, dynamic adjustment of k and optimal choice of σ in the clustering process is expected to be achieved. However, although the clustering results are affected by both k and σ, there is no effective way to consider them together. So, in our study, the clustering process is considered to be a combinatorial optimization problem. A specified validity index is considered as the optimization goal, k and σ is considered as the solutions of the optimization process, of which the best validity index corresponds to the best k and σ. In this paper, the fast black hole (FBH) algorithm is proposed to replace the K-mean method in the SC algorithm. As the initial clustering centers throughout the whole solution space, it is effective to make the clustering results insensitive to the initial centers. The *CritC* index function (Breaban and Luchian 2011) is used as the optimization target whose optimal value is used to determine the best k and σ.

2 Background

In actual oilfield production, dynamometer cards are used to analyze down-hole conditions of sucker rod pumping systems according to the production experience of technical staff. Different fault types are reflected by different graph shapes of the dynamometer cards. Taking some typical dynamometer cards in actual production for examples, “normal running” is reflected by “the above and bottom of the curve is nearly parallel, and the same of the left and right”; “liquid shortage in the pump” is reflected by “lack of right-bottom corner of the curve, the loading is normal and the unloading is slower”; “parting of rod” is reflected by “shape of the curve is near flat, and the loading drops”; “oil of high viscosity” is reflected by “shape of the curve is fat (round and convex)”; “travelling valve leakage” is reflected by “like parabola, the loading is slower and the unloading is faster”; “standing valve leakage” is reflected by “the upper portion of the curve is upward and the two sides are round, the unloading is slower and the loading is faster”; “pump bumping (upstroke)” is reflected by “the loading at the top dead point has a sudden increase”; “pump bumping (downstroke)” is reflected by “the loading at the bottom dead point has a sudden increase”; “sand production” is reflected by “zigzag pattern of the curve with crest tips and rapid changes”; “piston goes outside of the cylinder” is reflected by “lack of right-top corner of the curve, the unloading happens quickly”.

In order to eliminate the effects of deformation, viscous resistance, vibration and inertia of the sucker rod string, the surface dynamometer card is first transformed into a down-hole dynamometer card which can truly reflect the working conditions of the subsurface pump. In this paper, we use the Fourier coefficient method (Chen 1988; Li et al. 2013a) to solve the one-dimensional wave equation proposed by

Gibbs (Gibbs and Neely 1966) to complete this transformation. Then, graphic feature vectors of the down-hole dynamometer card are used as the inputs of the fault diagnosis model (in this paper, all feature vector sets we use are extracted from the down-hole dynamometer card). The transformation result of one surface dynamometer card is shown in Fig. 1.

In some of our earlier works, two feature extraction methods have been discussed, one is the curve moment-based method (Li et al. 2013a) and the other is Freeman chain code based method (Li et al. 2013b). These are briefly introduced here.

1 *Curve moment-based method* According to the "Four-point" analysis method which is commonly used in actual oil production, the down-hole dynamometer card is divided into four parts and then 7 invariant moment features are extracted. So, a total of 28 eigenvectors are used to describe the dynamometer card.

2 *Freeman chain code-based method* The boundary of the down-hole dynamometer card is represented by the Freeman chain code with 8 directions. 10 eigenvectors are calculated by changes of the curvature and 2 eigenvectors are calculated from the maximum and minimum load. So, these 12 important eigenvectors are used to describe the dynamometer card.

In this paper, the unsupervised classification method is used to diagnose down-hole working conditions of sucker rod pumping systems. This does not rely on training samples for artificial classification. The classification error will be reduced as the automatic classification of the data set can be realized according to its global distribution modes and data attributes. One key problem is to find a feature vector set (data attributes) to distinguish different types of down-hole dynamometer cards (distribution modes). All of them with the same fault type can be divided into a class by the automatic clustering algorithm. The above two feature extraction methods are used to extract the curve eigenvectors of the experimental data (down-hole dynamometer cards) to structure the data set (different fault types of down-hole dynamometer cards with different data distribution). Then the data set can be used to verify the effectiveness of the proposed algorithm.

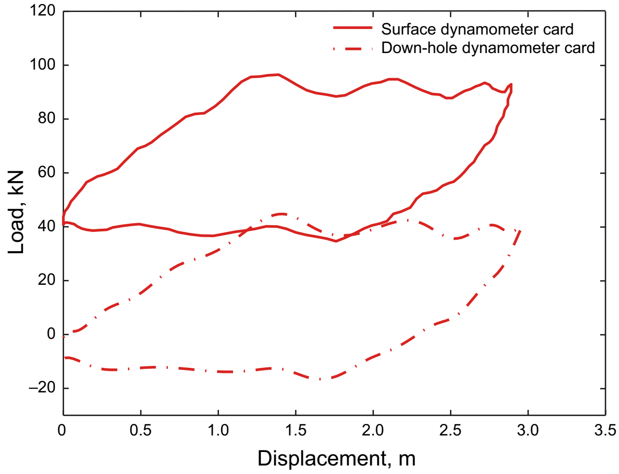


Fig. 1 Transformation of one surface dynamometer card to the equivalent down-hole dynamometer card

3 Basic theory

3.1 Spectral clustering (SC) algorithm

A SC algorithm based on a normalized Laplacian matrix L_{rw} is used in this paper. Two goals can be achieved using this algorithm: (1) finding a segmentation to make the different types of data dissimilar and (2) finding a segmentation to make the same type of data similar (Von Luxburg 2007).

Algorithm 1 SC algorithm based on normalized Laplacian matrix L_{rw}.

```
Input: similarity matrix $S \in R^{n\times n}$, the number of cluster categories $k$.
Output: the final clustering results $A_1, A_2, \ldots, A_k$, that is $A_i \in \{j|y_j \in C_i\}$.
Step 1: Establish a similarity connection diagram of samples, where $W$ is its connection weights matrix;
Step 2: Calculate the normalized degree matrix $D$ and the normalized Laplacian matrix $L_{\mathrm{rw}}$;
Step 3: For a generalized characteristic problem $L_{\mathrm{rw}}u = \lambda Du$, calculate eigenvectors $u_1, u_2, \ldots, u_k$ of the first $k$ minimum eigenvalues;
Step 4: Let $U \in R^{n\times k}$ be the matrix of $u_1, u_2, \ldots, u_k$ by column arrangement;
Step 5: For $i = 1, 2, \ldots, n$, let $y \in R^k$ correspond to the $i$th row vector of $U$;
Step 6: Use the K-mean algorithm to divide the samples $(y_i)_{i=1,\ldots,n}$ in $R^k$ into $k$ classes $C_1, C_2, \ldots, C_k$.
```

3.2 Black hole (BH) algorithm

The black hole (BH) algorithm was proposed by Hatamlou (2013) named after the "black holes" in astrophysics. It is a new heuristic optimization algorithm which has better performance than the particle swarm optimization (PSO) algorithm and the gravitational search algorithm (GSA). In the initialization stage, a group of candidate solutions is generated randomly, called "star". Then the fitness function values of the population are calculated and the optimal candidate solution is chosen whose fitness function value is the best in the population, called “black hole”. After initialization, the stars around the black hole are absorbed by it and movement of the stars toward the black hole can be formulated by

$$x_i(t+1) = x_i(t) + rand \times (x_{\text{BH}} - x_i(t)), \quad (1)$$

where $i = 1, 2, \ldots, N$; $x_i(t)$ is the location of the ith star at t iteration, $x_i(t+1)$ is the location of the ith star at $t+1$ iteration; x_{BH} is the location of the black hole in the search space; *rand* is a random value in [0, 1]; N is the number of stars.

If the fitness function value of one star is better than that of the black hole when the stars move toward the black hole, the black hole moves to the location of that star. Then the BH algorithm will continue and the stars will move towards the black hole in the new location. In some cases, the stars may be sucked in by the black hole when the stars move toward the black hole. If a star is sucked in, a new star generates randomly in the search space at the same time and then a new search is started. This way can keep the number of stars constant. The radius of the black hole can be formulated by

$$R = f_{\text{BH}} \Big/ \sum_{i=1}^{N} f_i, \quad (2)$$

where f_{BH} is the fitness function value of the black hole; f_i is the fitness function value of the ith star; N is the number of stars. The star will be sucked in by the black hole when the distance between it and the black hole is less than R.

Algorithm 2 BH algorithm

```
Initialization: Randomly generate a population of stars in the search space.
Step 1: For each star, calculate their fitness function value;
Step 2: Select the star that has the best fitness function value as the black hole;
Step 3: Change the location of each star according to Eq. (1);
Step 4: If the fitness function value of one star is better than the black hole, interchange their locations;
Step 5: If a star is sucked in by the black hole, a new star will generate randomly in the search space;
Step 6: If the terminal condition is satisfied, end the iteration; otherwise return to Step 1.
```

4 Automatic clustering based on fast black hole-spectral clustering (FBH–SC) algorithm

4.1 Fast black hole (FBH) algorithm

In the BH algorithm, a large amount of computation is needed when calculating fitness function values of all stars. When a star is sucked in by the black hole, a new star generates randomly at the same time and the next iteration starts. Then their fitness function values are recalculated and the new black holes are repositioned. In this process, it can be found that except only a few sucked stars, the position of most stars do not change, so their fitness function values have not changed. If they are recalculated in a new iteration, the computation obviously increases. So, in order to solve this problem, we propose the FBH algorithm. The fitness function values of all stars are calculated in the initialization and then only the fitness function values of new stars are recalculated in Step 5.

Algorithm 3 FBH algorithm

```
Initialization: Randomly generate a population of stars in the search space, and calculate their fitness function values.
Step 1: Select the star that has the best fitness function value as the black hole;
Step 2: Change the location of each star according to Eq. (1);
Step 3: If the fitness function value of one star is better than the black hole, interchange their locations;
Step 4: If a star is sucked in by the black hole, a new star will generate randomly in the search space, and their fitness function values are recalculated;
Step 5: If the terminal condition is satisfied, end the iteration; otherwise return to Step 1.
```

Next, the computation time of the BH algorithm and FBH algorithm are compared in the following. Suppose that both the number of iterations of two algorithms is *MaxLoop*, each computation time of the fitness function values of all stars is T, each computation time of the moving stars in the FBH algorithm is t, as the number of the moving stars is far less than the total number of the stars, so $t \ll T$. It is considered that each t is basically the same for *MaxLoop* iterations. So, the computation time of the BH algorithm is about $T_1 = MaxLoop \times (T + T) = 2 \times MaxLoop \times T$ and the FBH algorithm is about $T_2 = T + MaxLoop \times (T + t) = (MaxLoop + 1) \times T + MaxLoop \times t$. t can be considered zero as $t \ll T$ and T_2 is approximately equal to $(MaxLoop + 1) \times T$, which is obviously smaller than T_1.

4.2 FBH–SC algorithm

Now, we use the proposed FBH algorithm to replace the K-mean method in the SC algorithm to solve the problem that the clustering results are sensitive to the initial centers. The proposed FBH–SC algorithm is as follows:

Algorithm 4 FBH–SC algorithm

```
Input: similarity matrix $S \in R^{n \times n}$, the number of the cluster categories $k$.
Output: the final clustering results $A_1, A_2, \ldots, A_k$, that is $A_i \in \{j | y_j \in C_i\}$.
Step 1: Establish a similarity connection diagram of the samples, where $W$ is its connection weights matrix;
Step 2: Calculate the normalized degree matrix $D$ and the normalized Laplacian matrix $L_{\mathrm{rw}}$;
Step 3: For $L_{\mathrm{rw}} u = \lambda D u$, calculate eigenvectors $u_1, u_2, \ldots, u_k$ of the first $k$ minimum eigenvalues;
Step 4: Let $U \in R^{n \times k}$ be the matrix of $u_1, u_2, \ldots, u_k$ by column arrangement;
Step 5: For $i = 1, 2, \ldots, n$, let $y \in R^k$ correspond to the $i$th row vector of $U$;
Step 6: Use the FBH algorithm to divide the samples $(y_i)_{i=1,\ldots,n}$ in $R^k$ into $k$ classes $C_1, C_2, \ldots, C_k$.
```

4.3 FBH–SC-based automatic clustering algorithm

The proposed FBH–SC algorithm is an improvement of the traditional SC algorithm, however, the problem of having to pre-set of the number of cluster categories (k) and scale parameter (σ) still exists. So, in this paper, the clustering process is considered to be a combinatorial optimization problem and a proper validity index function is used as the optimization target; the best k and σ are obtained by searching the optimal validity index value. In this paper, the *CritC* index function is used as the optimization index function which is defined as follows,

$$CritC = (a \cdot F)^{le(k)}, \tag{3}$$

where $F = 1/(1 + W/B)$; $B = \sum_{i=1}^{k} |C_i| \delta(C_i, g)$ is a measure among different classes, $W = \sum_{i=1}^{k} \sum_{d \in C_i} \delta(C_i, d)$ is a measure within a class; $a = 2 \cdot m/(2 \cdot m + 1)$, under the same k, the greater m is, the greater *CritC* is, here $m = 2$; $le(k) = \log_2(k + 1) + 1$; the value range of the *CritC* index function is in [0, 1], and the greater *CritC* represents the better clustering results.

Algorithm 5 FBH–SC-based automatic clustering algorithm

```
Initialization: $k$ and $\sigma$ are taken as solutions in the clustering process, where $k \le \sqrt{n}$ ($n$ is the number of samples) (Yu and Chen 2002; Wang et al. 2012), $\sigma \in [0, 1]$; randomly generate a population of stars in the search space, and recalculate their fitness function values.
Step 1: Select the star whose CritC is the greatest as the black hole;
Step 2: Change the location of each star according to Eq. (1);
Step 3: If CritC of one star is better than the black hole, interchange their locations;
Step 4: If a star is sucked in by the black hole, a new star will generate randomly in the search space, and its CritC is recalculated;
Step 5: If the terminal condition is satisfied, end the iteration; otherwise return to Step 1.
```

5 Experiments

5.1 Computing environment and data set

The computer configuration is as follows: Microsoft Windows 7, Intel Core 2 Duo E6570 @ 2.66 GHz, Samsung 2 GB DDR2 667 MHz. Five data sets are used, iris, wine, and seeds in the UCI data base, data 1 and data 2 are composed separately by the characteristic data of 126 dynamometer cards using two methods in Sect. 2 ("data 1" stands for the characteristic data extracted by the curve moment based method; and "data 2" stands for the characteristic data extracted by the Freeman chain code based method), shown in Table 1. The data distribution of data 1 and data 2 is shown in Fig. 2.

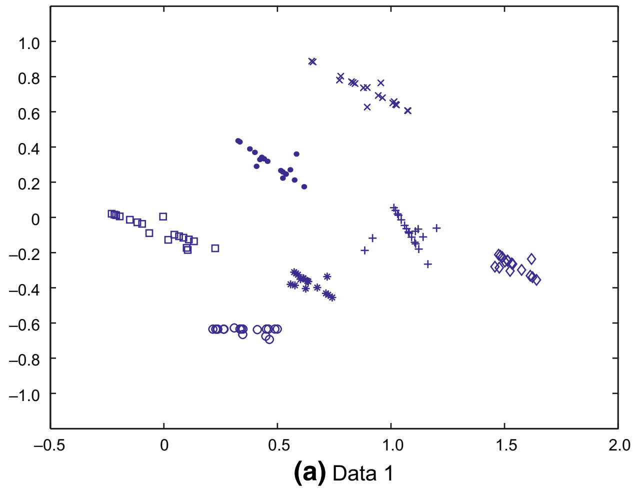
5.2 Comparison of BH–SC algorithm and FBH–SC algorithm

The FBH algorithm is proposed to decrease the computational complexity of the BH algorithm. The comparison results of two algorithms of the clustering accuracy and computation time are given in Tables 2 and 3 (running 50 times respectively).

It can be seen from Tables 2 and 3, the clustering accuracy of the FBH–SC algorithm is basically the same as the BH–SC algorithm. However, its computation time is much less than that of the BH–SC algorithm. So, while retaining reliable calculation results, the FBH–SC algorithm can effectively decrease the computation time.

Table 1 Characteristics of different data sets

Data sets	Class	Dimension	Number
Iris	3	4	150
Wine	3	13	178
Seeds	3	8	210
Data 1	7	28	126
Data 2	7	12	126



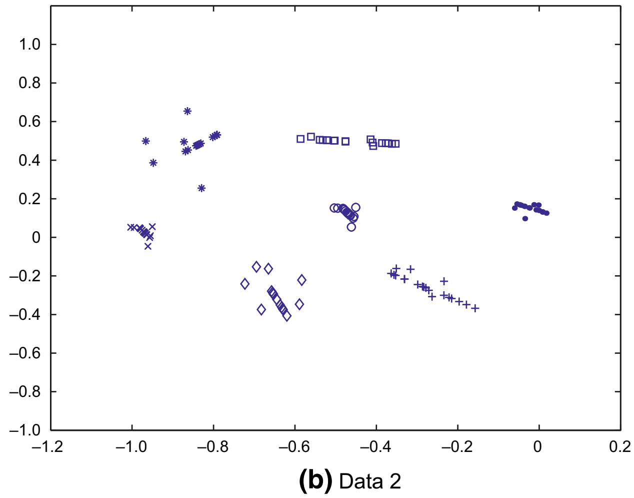


Fig. 2 Data distribution of data 1 and data 2

5.3 Relationship between k and *CritC*

An F index function is used as the fitness function, which is defined as follows (Hatamlou, 2013),

$$F(O,\ Z) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \left\| O_i - Z_j \right\|^2, \quad (4)$$

Table 3 Comparison of the computation time (mean ± std)

Data set	BH–SC	FBH–SC
Iris	289.39 ± 37.289	163.96 ± 5.04
Wine	388.96 ± 99.36	235.36 ± 66.54
Seeds	545.87 ± 161.95	311.47 ± 48.13
Data 1	439.39 ± 96.26	235.27 ± 51.51
Data 2	425.23 ± 40.08	233.63 ± 20.92

where n is the number of samples; k is the number of cluster categories; $\left\| O_i - Z_j \right\|^2$ is the distance between O_i and its clustering center; w_{ij} denotes the weight of O_i belonging to the jth class, if the sample belongs to the jth class, w_{ij} is 1, else 0; w_{ij} has arbitrary value in interval (0, 1) in fuzzy clustering.

However, in our research, it is found that the correct number of cluster categories cannot be obtained when the F index function has its optimal value, shown in Fig. 3.

As shown in Fig. 4, the F index value of five data sets gradually reduces along with increases of k when $\sigma = 0.22$, which is inconsistent with the actual classification. In fact, for iris, wine and seeds, F is minimum when $k = 3$, and for data 1 and data 2, F is minimum when $k = 7$. For the *CritC* index function, the relationship between its value and k is shown in Fig. 5.

As shown in Fig. 5, when $\sigma = 0.22$, for iris, wine and seeds, *CritC* is maximum when $k = 3$, and for data 1 and data 2, it is maximum when $k = 7$. Then, when σ takes different values, their relationships are discussed below, which are shown in Fig. 5.

As shown in Fig. 5, for the five data sets, although σ takes different values, the maximum value of *CritC* corresponds to the optimal k. So there exists a correspondence relationship between the *CritC* index function and the correct clustering number.

5.4 Relationship between k and clustering accuracy

Now, when σ takes different values, the relationships between k and the clustering accuracy of five data sets are discussed below, which are shown in Fig. 6.

Table 2 Comparison of the clustering accuracy (mean ± std)

Data set	SC	BH–SC	FBH–SC
Iris	0.9109 ± 0.0014 ($\sigma = 0.21$)	0.9279 ± 0.001339 ($\sigma = 0.21$)	0.9277 ± 0.001299 ($\sigma = 0.21$)
Wine	0.9742 ± 0.0011 ($\sigma = 0.29$)	0.9788 ± 0.001617 ($\sigma = 0.29$)	0.9787 ± 0.001557 ($\sigma = 0.29$)
Seeds	0.9161 ± 0.0019 ($\sigma = 0.53$)	0.9213 ± 0.007065 ($\sigma = 0.53$)	0.9211 ± 0.006906 ($\sigma = 0.53$)
Data 1	0.6844 ± 0.098 ($\sigma = 0.51$)	0.8098 ± 0.03321 ($\sigma = 0.51$)	0.8095 ± 0.03214 ($\sigma = 0.51$)
Data 2	0.6912 ± 0.093 ($\sigma = 0.29$)	0.8139 ± 0.01 ($\sigma = 0.29$)	0.8114 ± 0.08104 ($\sigma = 0.29$)

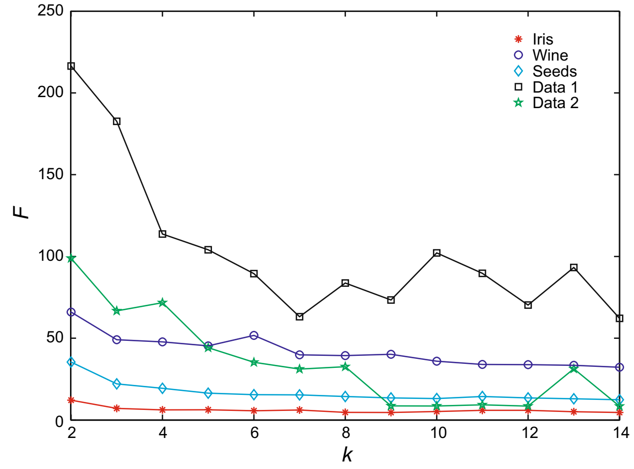


Fig. 3 *F* index value of five data sets ($\sigma = 0.22$)

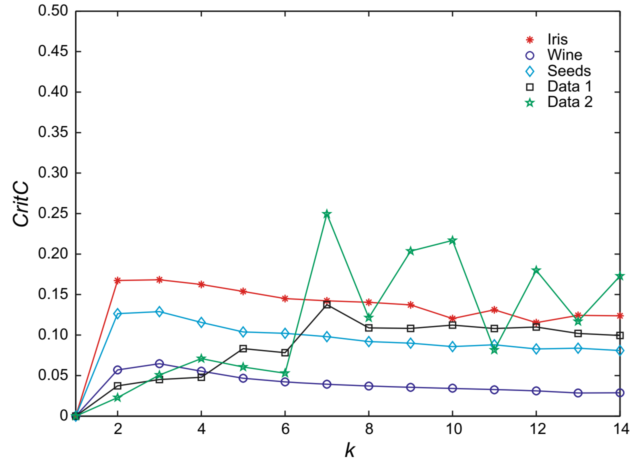


Fig. 4 *CritC* index value of five data sets ($\sigma = 0.22$)

As shown in Fig. 6, when σ takes different values, the optimal k corresponds to the highest clustering accuracy. So, there exists a correspondence relationship between the clustering accuracy and the correct clustering number.

5.5 Relationship between σ and *CritC*

Next, the influences of different σ on the *CritC* of five data sets are discussed below.

For the sake of discussion, we take values of k are 2, 3, 4, and 5 in iris, wine, and seeds, and values of k are 6, 7, 8, and 9 in data 1 and data 2. It can be seen from Fig. 7, for iris, wine, and seeds, when $k = 3$, *CritC* has the optimal values if σ is taken values in a certain range; and for data 1 and data 2, when $k = 7$, *CritC* has the optimal values. So, the *CritC* index function will have the optimal value if σ is chosen reasonably.

5.6 Relationship between σ and the clustering accuracy

The influences of different σ on the clustering accuracy of five data sets are discussed below.

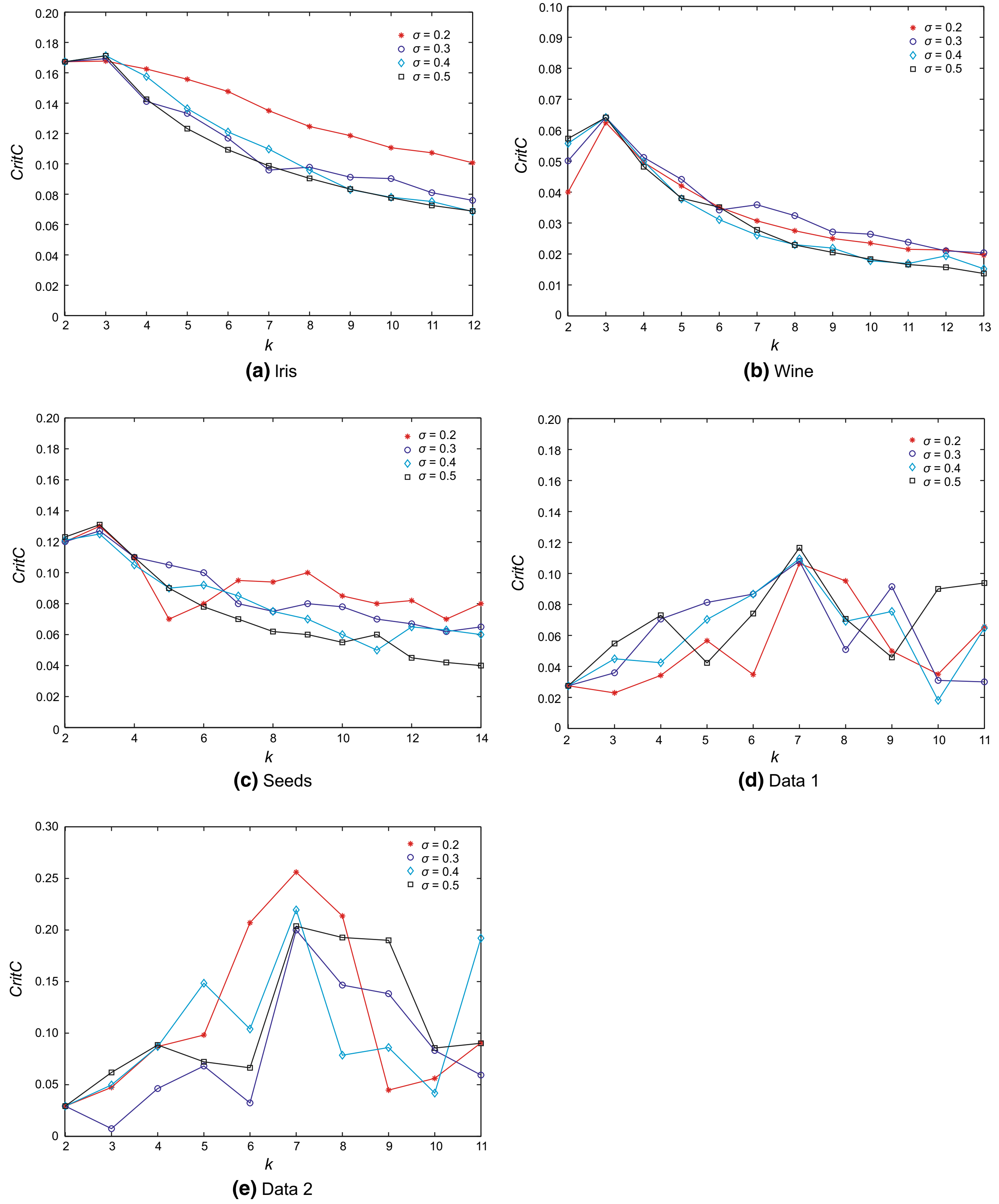
As shown in Fig. 8, for each data set, the optimal value or suboptimal value of the clustering accuracy can be obtained when a reasonable σ is chosen. A comparison of Figs. 8 and 9 shows that the interval of σ having the optimal index function values is contained in the interval of σ having the highest clustering accuracy, which illustrates that it is reasonable to use the index function as the fitness function to select the proper σ. However, to some extent, it can be seen that the clustering accuracy corresponding to the interval of σ having the optimal index value is not the highest value, but the next highest one. It is because the measurement of *CritC* is different from the clustering accuracy, but its next highest value is also greater than the clustering accuracy of other k under the same conditions. Automatic clustering results of five data sets using FBH–SC algorithm are given in Table 4.

It can be seen from Table 4, although the clustering accuracy of the FBH–SC-based automatic clustering algorithm is slightly below the highest clustering accuracy of the BH–SC algorithm and the FBH–SC algorithm, k and σ need not to be set in advance in the proposed automatic clustering algorithm. Moreover, parameters in the optimal algorithm need not be set in advance. So the proposed algorithm possesses better comprehensive performance.

6 Case study

The number of samples (down-hole dynamometer cards) of the experimental data set is expanded to 228 of 11 classes, which were collected from one oil well in one oilfield, China. They are: “normal running” (for the purpose of discussion, “normal running” is considered as one fault type), 24 samples; “gas obstruction”, 20 samples; “liquid shortage in the pump”, 24 samples; “parting of rod”, 20 samples; “oil of high viscosity”, 20 samples; “travelling valve leakage”, 20 samples; “pump bumping (upstroke)”, 20 samples; “pump bumping (downstroke)”, 20 samples; “standing valve leakage”, 20 samples; “sand production”, 20 samples; “piston goes outside of the cylinder” 20 samples. The second feature extraction method in Sect. 2 is used to extract their characteristics, named by data 3. The data distribution of the set data 3 is shown in Fig. 9.

One diagnostic dynamometer card is taken as an example to show the effectiveness of the proposed method. The diagnostic dynamometer card is shown in Fig. 10, which is first transformed into the down-hole dynamometer card.



(a) Iris

(b) Wine

(c) Seeds

(d) Data 1

(e) Data 2

Fig. 5 *CritC* index value of five data sets with different σ

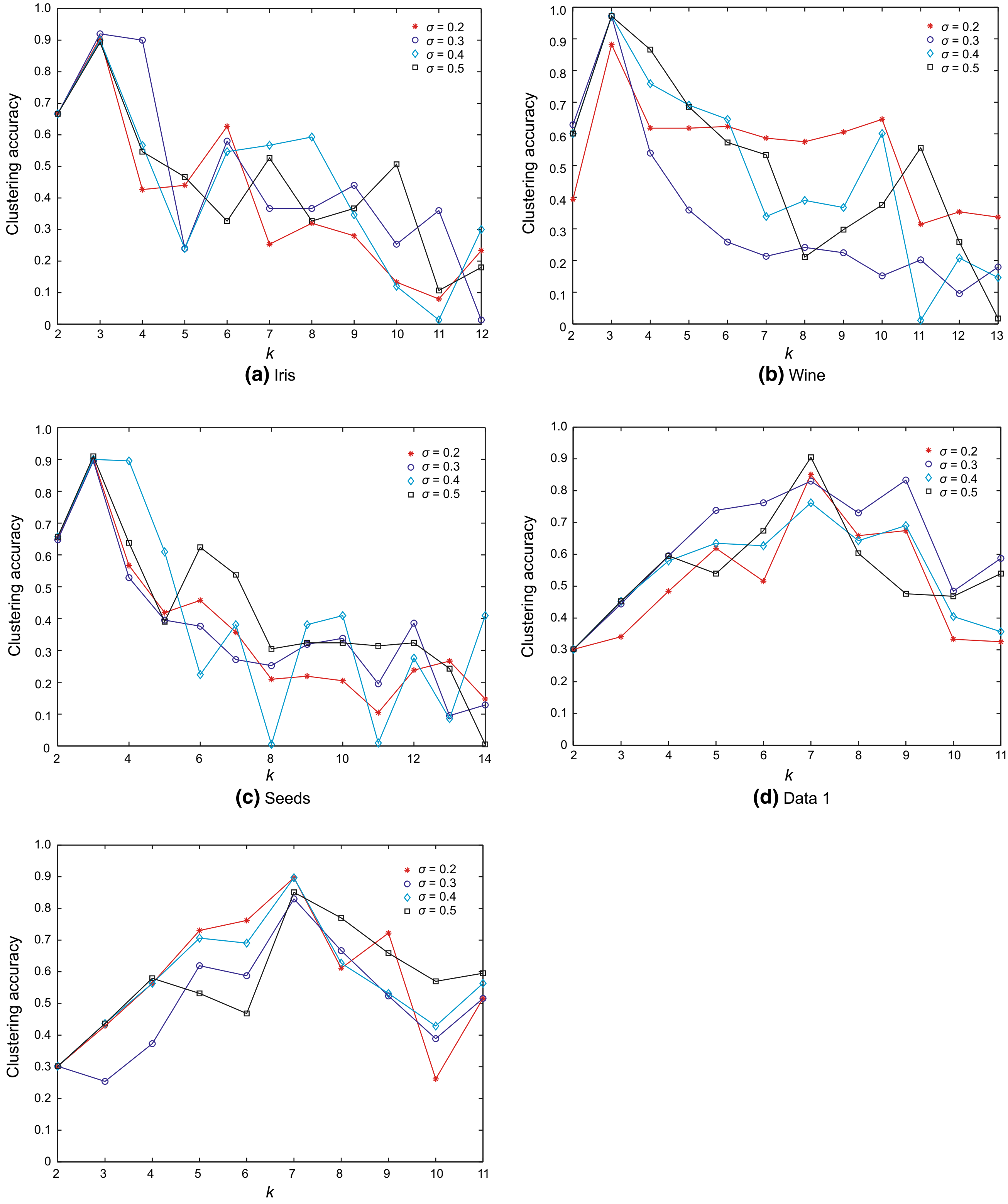
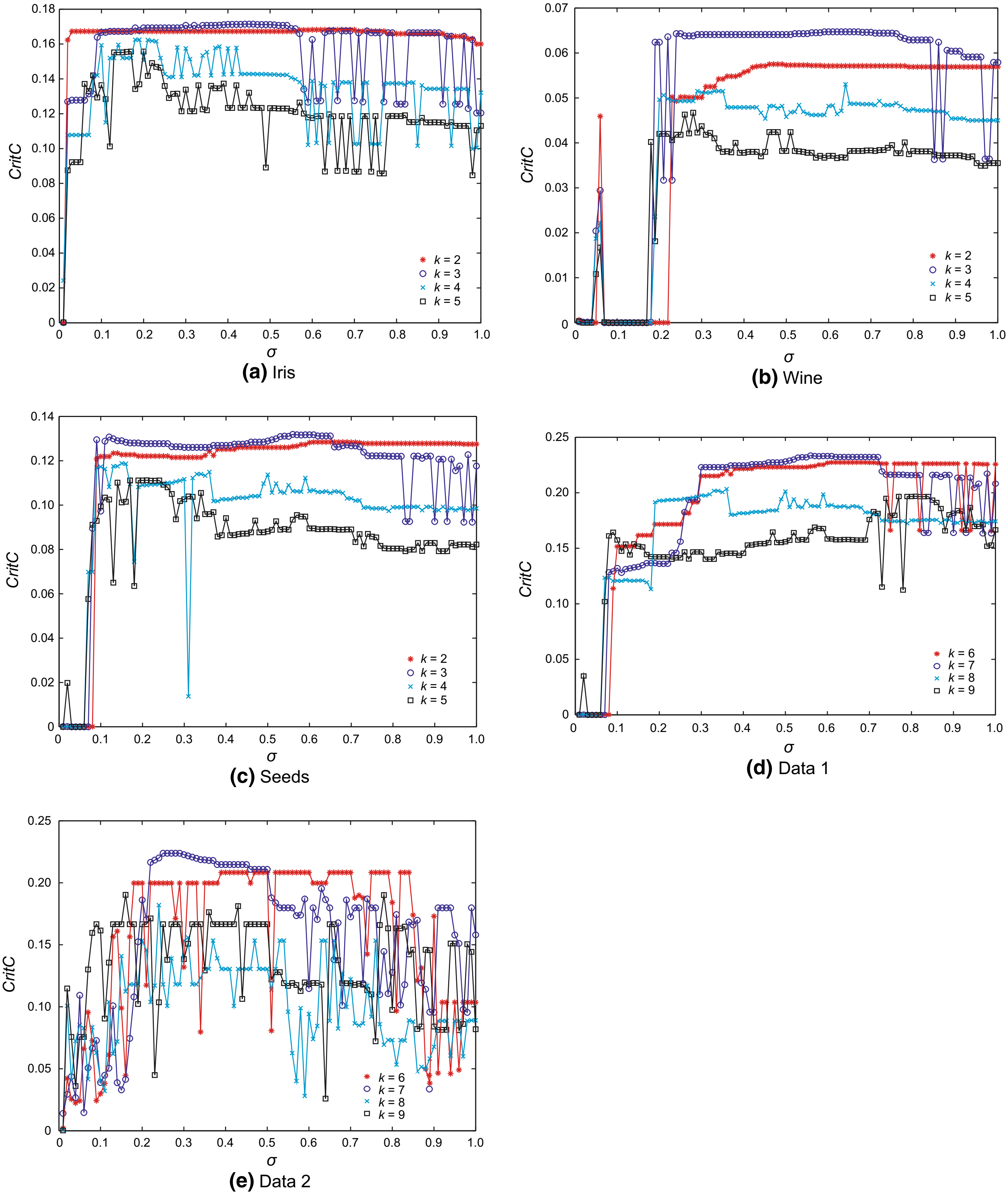


Fig. 6 The clustering accuracy of five data sets with different σ



(a) Iris

(b) Wine

(c) Seeds

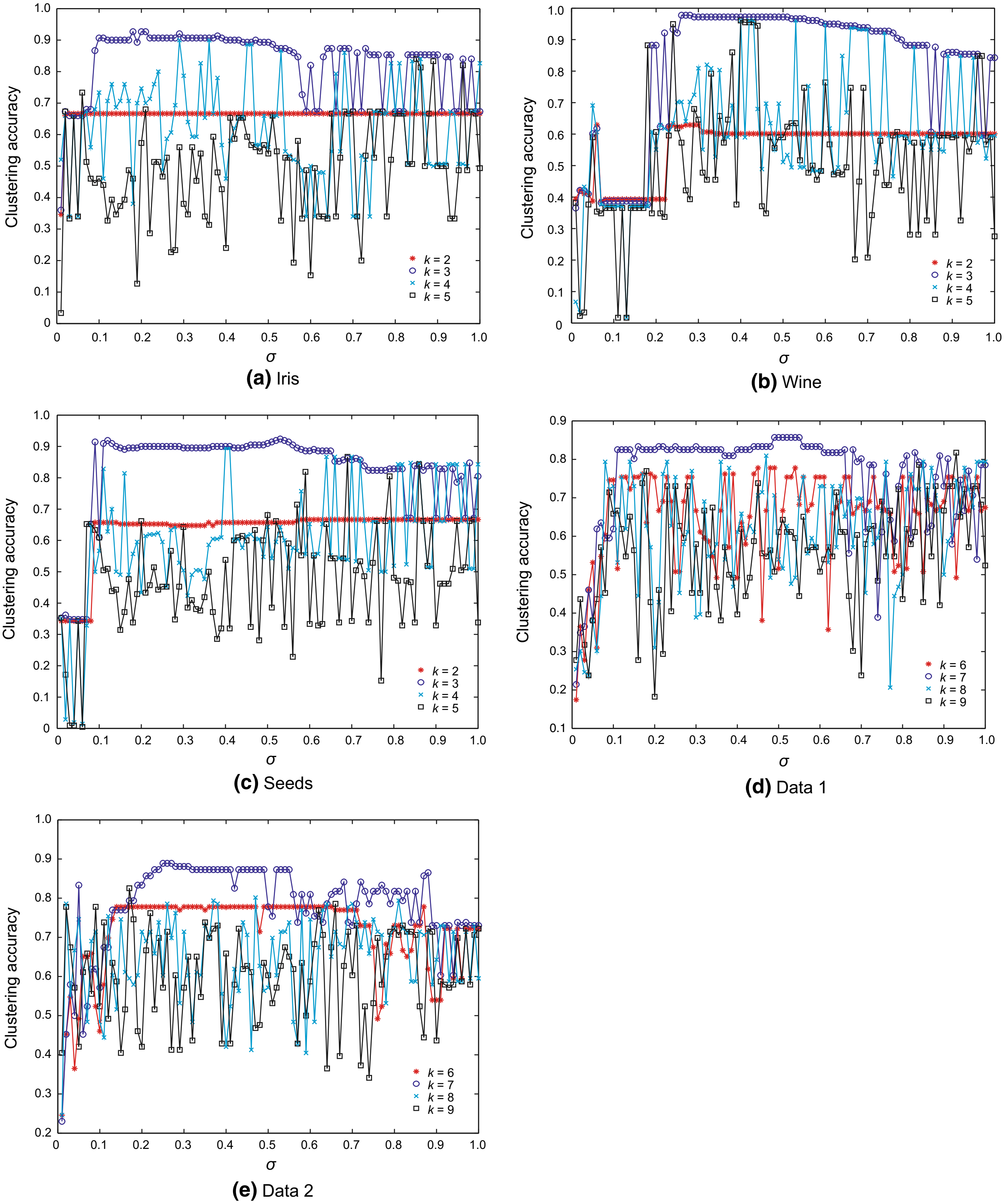
(d) Data 1

(e) Data 2

Fig. 7 Relationship of *CritC* and σ with different k

In Table 5, 12 eigenvector parameters are obtained using the second feature extraction method in Sect. 2.

Let us consider the data 3 (228 samples) and the diagnostic sample together as a new data set data 4 (229 samples). Then automatic clustering of the data 4 is completed (running 10 times and taking the maximum value of *CritC*) and the results are shown in Table 6 and Fig. 11.



(a) Iris

(b) Wine

(c) Seeds

(d) Data 1

(e) Data 2

Fig. 8 Influence of σ on the clustering accuracy with different k

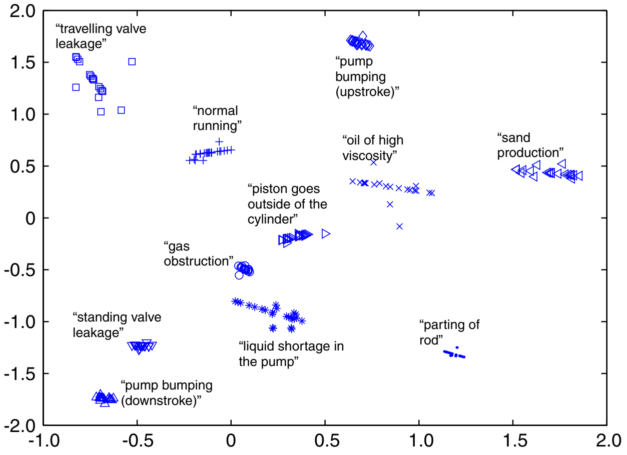


Fig. 9 Data distribution of the set data 3

Table 4 Automatic clustering results of five data sets

Data set	*CritC*	Clustering accuracy	Optimal clustering number (*k*)	Optimal scale parameter (*σ*)
Iris	0.1714	0.9133	3	0.4482
Wine	0.0647	0.9551	3	0.5906
Seeds	0.1318	0.9286	3	0.5205
Data 1	0.2322	0.8412	7	0.5102
Data 2	0.2238	0.8730	7	0.2946

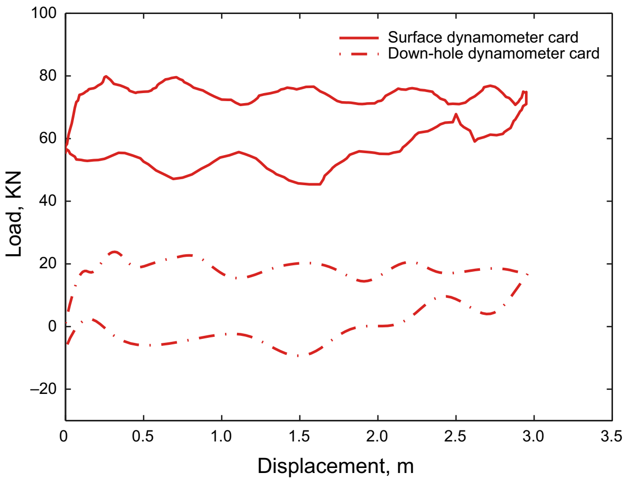


Fig. 10 The diagnostic dynamometer card

Table 5 12 eigenvector parameters of the diagnostic sample

Eigenvector parameters	Values
Degree of zigzag	0.0600
Degree of bulge of the left-bottom corner	0.1144
Degree of bulge of the right-top corner	0.1270
Degree of flatness	0.1288
Degree of lack of the left-top corner	0.076
Degree of lack of the right-top corner	0.0250
Degree of lack of the right-bottom corner	0.1127
Degree of lack of the left-bottom corner	−0.033
Degree of sharp-load of the left-top corner	0.1020
Degree of sharp-unloading of the right-bottom corner	0.1575
Degree of rapid-unloading of the right-top corner	0.0526
Degree of fatness	−0.0031

Table 6 Automatic clustering results of data 4

Data set	*CritC*	Clustering accuracy	The optimal clustering number (*k*)	The optimal scale parameter (*σ*)
Data 3	0.234	0.9167	11	0.1346

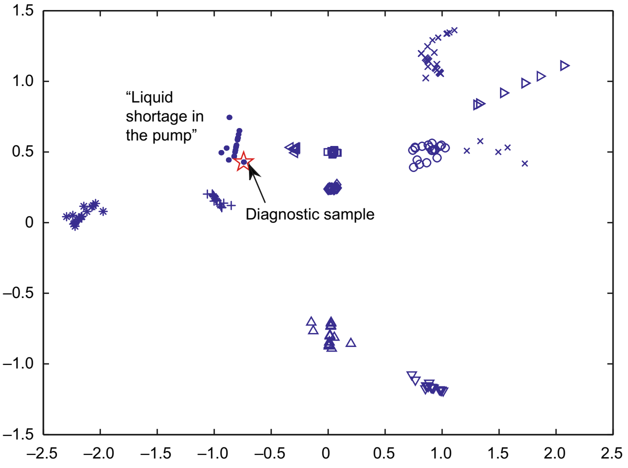


Fig. 11 The diagnosis result

As shown in Table 6, the correct clustering number and the satisfied clustering accuracy can be obtained.

It can be seen from Fig. 11, the diagnostic sample (marked by “☆”) is classified into the fault type of “liquid shortage in the pump” automatically by the proposed automatic clustering algorithm. In order to check the accuracy of the diagnosis result, we analyze the curve shape of the diagnostic sample according to the artificial production experiences. The graph’s main characteristics are described as “lack of the right-bottom corner, the faster loading and the slower unloading”, which has the main characteristics of “liquid shortage in the pump”. So, it is basically consistent with the computer diagnosis results.

7 Conclusions

An automatic clustering algorithm based on the FBH–SC algorithm is proposed in this paper. It is used to solve the problem that in current studies fault diagnosis for downhole conditions of sucker rod pumping systems relies too much on training samples. This algorithm is not limited by any data distribution shapes and is insensitive to the initial cluster centers. Besides, any parameters need not be set in an optimal process. The *CritC* index function is used as the target function to find the best k and σ. According to the analysis of the relationships among *CritC*, σ, k and the clustering accuracy of the five data sets, the proposed algorithm is useful in an unsupervised learning mode. However, although the proposed FBH algorithm has nearly halved the operation time of the BH algorithm, it still takes about 330 min for the set data 3 as the initial solution throughout the whole search space and the mechanisms of "suck" and "generation" used in the solving process. So, in the further work, we will focus on how to reduce the operation time of the algorithm as much as possible. We emphasize it is still a difficult problem to use an unsupervised learning method to diagnose multiple faults for down-hole conditions of sucker rod pumping systems.

Acknowledgments The authors are grateful for financial support from the National Natural Science Foundation of China (Grant No. 61403040). The authors also thank for the support from the Liaohe Oilfield of China, providing research and experimental conditions.

References

Alzate C, Suykends JAK. Hierarchical kernel spectral clustering. Neural Netw. 2012;35:21–30.

Bezdek JC, Pal NR. Some new indexes of cluster validity. IEEE Trans Syst Man Cybern. 1998;28(3):301–15.

Breaban M, Luchian H. A unifying criterion for unsupervised clustering and feature selection. Pattern Recognit. 2011;44(4): 854–65.

Chen JL. A fast algorithm for down-hole dynagrams in sucker rod pumping wells. Acta Pet Sin. 1988;9(3):105–13 (in Chinese).

de Souza A, Bezerra M, Filho M, et al. Using artificial neural networks for pattern recognition of downhole dynamometer card in oil rod pump system. Proceedings of the 8th WSEAS international conference on artificial intelligence, knowledge engineering and data bases, February 2009, Cambridge.

Derek HJ, Jennings JW, Morgan SM. Sucker rod pumping unit diagnostics using an expert system. Paper SPE 17318 presented at permian basin oil and gas recovery conference, 10–11 March 1988, Midland.

Frederix K, Van Barel M. Sparse spectral clustering method based on the incomplete Cholesky decomposition. J Comput Appl Math. 2013;237:145–61.

Fujiwara K, Sawada H, Kano M. Input variable selection for PLS modeling using nearest correlation spectral clustering. Chemom Intell Lab Syst. 2012;118:109–19.

Gibbs SG, Neely AB. Computer diagnosis of down-hole conditions in sucker rod pumping wells. J Pet Technol. 1966;18(1):91–8.

Hatamlou A. Black hole: a new heuristic optimization approach for data clustering. Inf Sci. 2013;222:75–184.

Li K, Gao XW, Tian ZD, et al. Using the curve moment and the PSO-SVM method to diagnose downhole conditions of a sucker rod pumping unit. Pet Sci. 2013a;10(1):73–80.

Li K, Gao XW, Yang WB, et al. Multiple fault diagnosis of down-hole conditions of a sucker rod pumping unit based on Freeman chain code and DCA. Pet Sci. 2013b;10(3):139–48.

Li K, Gao XW, Zhou HB, et al. Fault diagnosis for down-hole conditions in beam pumping units based on an improved fuzzy iterative self-organizing data analysis technique. Proceedings of the 10th international conference on fuzzy systems and knowledge discovery. 23–25 July 2013c, Shenyang.

Liu N, Xiao ZB, Lu MY. Spectral co-clustering documents and words based on fuzzy K-harmonic means. Control Decis. 2012;27(4): 501–6 (in Chinese).

Lv SG, Feng YL. Consistency of coefficient-based spectral clustering with l1-regularizer. Math Comput Model. 2013;57:469–82.

Martinez ER, Moreno WJ, Castillo VJ, et al. Rod pumping expert system. Paper SPE 26246 presented at SPE petroleum computer conference, 11–14 July 1993, New Orleans.

Mirkin B, Nascimento S. Additive spectral method for fuzzy cluster analysis of similarity data including community structure and affinity matrices. Inf Sci. 2012;183:16–34.

Rogers JD, Guffey CG, Oldham WJB. Artificial neural networks for identification of beam pump dynamometer load cards. Paper SPE 20651 presented at annual technical conference and exhibition, 23–26 September, 1990, New Orleans.

Saha S, Bandyopadhyay S. Some connectivity based cluster validity indices. Appl Soft Comput. 2012;12(5):1555–65.

Tasdemir K. Vector quantization based approximate spectral clustering of large datasets. Pattern Recognit. 2012;45:3034–44.

Tian JW, Gao MJ, Li K, et al. Fault detection of oil pump based on classify support vector machine. Proceedings of the international conference on control and automation, 30 May–1 June, 2007a, Guangzhou.

Tian JW, Gao MJ, Liu YX, et al. The fault diagnosis system with self-repair function for screw oil pump based on support vector machine. Proceedings of the international conference on robotics and biomimetics, 15–18 December, 2007b, Sanya, Hainan, China.

Vega-Pons S, Ruiz-Shulcloper J. A survey of clustering ensemble algorithms. Int J Pattern Recognit Artif Intell. 2011;25(3):337–72.

Von Luxburg U. A tutorial on spectral clustering. Stat Comput. 2007;17(4):395–416.

Wang J, Wang ST, Deng ZH. Survey on challenges in clustering analysis research. Control Decis. 2012;27(3):321–8 (in Chinese).

Wang JP, Bao ZF. Study of pump fault diagnosis based on rough sets theory. 2008 3rd international conference on Innovative Computing Information and Control (ICICIC), 18–20 June 2008, Dalian.

Wu W, Sun WL, Wei HX. A fault diagnosis of sucker rod pumping system based on wavelet packet and RBF network. Adv Mater Res. 2011;189–193:2665–9.

Xu P, Xu SJ, Yin HW. Application of self-organizing competitive neural network in fault diagnosis of sucker rod pumping system. J Pet Sci Eng. 2007;58(1–2):43–8.

Yu DL, Zhang YM, Bian HM, et al. A new diagnostic method for identifying working conditions of submersible reciprocating pumping systems. Pet Sci. 2013;10(1):81–90.

Yu J, Chen QS. Search range of the optimal cluster number in fuzzy clustering. Sci China (Ser E). 2002;32(2):274–80 (in Chinese).

Zhang XG, Jiao LC, Liu F, et al. Spectral clustering ensemble applied to SAR image segmentation. IEEE Trans Geosci Remote Sens. 2008;46(7):2126–36.