Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition Letters 27 (2006) 765–772

Pattern Recognition Letters

www.elsevier.com/locate/patrec

An efficient line symmetry-based *K*-means algorithm

Kuo-Liang Chung [*,1], Keng-Sheng Lin

Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology,
No. 43, Section 4, Keelung Road, Taipei 10672, Taiwan

Received 13 December 2004; received in revised form 11 November 2005
Available online 10 January 2006

Communicated by A.M. Alimi

Abstract

Recently, Su and Chou presented an efficient point symmetry-based *K*-means algorithm. Extending their point symmetry-based *K*-means algorithm, this paper presents a novel line symmetry-based *K*-means algorithm for clustering the data set with line symmetry property. Based on some real data sets, experimental results demonstrate that our proposed line symmetry-based *K*-means algorithm is rather encouraging.

Keywords: Clustering; *K*-means algorithm; Point symmetry; Line symmetry

1. Introduction

Partitioning a set of data points into some nonoverlapping clusters is an important topic in data analysis and pattern classification. It has many applications, such as codebook design (Gersho and Gray, 1992), data mining (Ng and Han, 2002), image segmentation (Jain and Dubes, 1988), data compression (Sayood, 1996), etc. Many efficient clustering algorithms (Fischer and Buhmann, 2003; Bajcsy and Ahuja, 1998; Hartigan, 1975; Zhu and Po, 1998; Fred and Leitao, 2003; Su and Chou, 2001) have been developed for data sets of different distributions in the past several decades. Most of existing clustering algorithms adopt the 2-norm distance measure in clustering process.

Among these developed clustering algorithms, Su and Chou (2001) first took the point symmetry issue (Zabrodsky et al., 1995; Kanatani, 1997) into account. Based on their proposed point symmetry distance (PSD) measure, they presented a novel and efficient clustering algorithm, which is very suitable for symmetrical intra-clusters; for convenience, their proposed clustering algorithm is named the PSK algorithm. Experimental results demonstrate that the previous PSK clustering algorithm outperforms the traditional *K*-means algorithm. In essence, the PSK algorithm not only inherits the simplicity advantage of the *K*-means algorithm, but it also can handle the symmetrical intra-clusters quite well. Recently, their proposed PSK algorithm was improved by Chung and Lin (in press) and extended to be able to handle both the symmetrical intra-clusters and the symmetrical inter-clusters; for convenience, their proposed clustering algorithm is called the IPSK algorithm. From the geometrical symmetry viewpoint, point symmetry and line symmetry are two widely discussed issues. The motivation of our research is to develop a new clustering algorithm for handling the data set with line symmetry property while preserving the advantages in the previous PSK algorithm and the previous IPSK algorithm.

In this paper, we propose a line symmetry-based *K*-means (LSK) algorithm for clustering the data set with line symmetry property while preserving the advantages in the previous PSK algorithm and the previous IPSK algorithm.

[*] Corresponding author. Tel.: +886 227301081; fax: +886 227301080.
E-mail address: klchung@cs.ntust.edu.tw (K.-L. Chung).

[1] Supported in part by the National Science Council of ROC under contract NSC92-2213-E-011-079.

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.11.006

Consequently, the proposed clustering algorithm can handle the data set with point symmetry property, line symmetry property, or both properties. Given a data set, the *K*-means algorithm is first used to obtain *k* temporary clusters. Second, the concept of centroid moment (Hu, 1962) is applied to determine the symmetrical line of each cluster which has been obtained by the *K*-means algorithm. Finally, the symmetry similarity level (SSL) operator is modified and extended to measure the line symmetry level between two data points. The modified SSL operator is called the MSSL operator for convenience. Utilizing the obtained symmetrical line of each cluster and the proposed MSSL operator, our proposed LSK algorithm can determine the most line-symmetrical data points when we are given a set of data points. Under some real data sets, experimental results demonstrate the feasibility of our proposed line-symmetry based *K*-means algorithm and the experimental results are rather encouraging.

The remainder of this paper is organized as follows. In Section 2, the previous PSK algorithm by Su and Chou is surveyed. In Section 3, the proposed MSSL operator is presented to measure the level of symmetry and it will be used in our propose LSK algorithm. In Section 4, our proposed LSK algorithm is described. In Section 5, some experimental results are demonstrated to show the effectiveness of the proposed LSK algorithm. In Section 6, some conclusion remarks are addressed.

2. The past work by Su and Chou

Different natural scenes usually have different features. Among these features, symmetry property is one of the most popular ones. Based on *K*-means algorithm, recently Su and Chou (2001) presented an efficient PSD measure to help partitioning the data set into the clusters where each cluster has the point symmetry property. In this section, the previous PSK algorithm by Su and Chou is surveyed.

Given N data points, $\{p_i \mid \text{for } 1 \leqslant i \leqslant N\}$, after running the *K*-means algorithm, let the obtained k temporary cluster centroids be denoted by $\{c_k \mid \text{for } 1 \leqslant k \leqslant K\}$. The PSD measure between the data point p_i and the data point p_j relative to the cluster centroid c_k is defined as

$$d_s(p_j, c_k) = \min \frac{\|(p_j - c_k) + (p_i - c_k)\|}{\|p_j - c_k\| + \|p_i - c_k\|} \tag{1}$$

for $i \neq j$ and $1 \leqslant i \leqslant N$ where $\|\cdot\|$ denotes the 2-norm distance. Using a similar data set as in (Su and Chou, 2001), Fig. 1(a) illustrates a set of data points which contains two point symmetrical clusters C_1 and C_2 associated with the centroids c_1 and c_2, respectively. Fig. 1(b) demonstrates the clustering result by running the *K*-means algorithm on Fig. 1(a). In Fig. 1(b), the data point p_2 is assigned to the cluster C_2 because the distance between p_2 and c_2 is less than the distance between p_2 and c_1. However, from human visualization, it will be better to assign the data point p_2 to the cluster C_1 due to the point symmetrical distribution of data points in C_1. Applying the PSD measure shown in Eq. (1) to Fig. 1(a), Fig. 1(c) illustrates the satisfactory clustering result after running the PSK algorithm on Fig. 1(a).

From Fig. 1, it is observed that the previous PSK algorithm by Su and Chou worked for clustering the point symmetrical data set and experimental results demonstrated that the PSK algorithm significantly outperforms the conventional *K*-means clustering algorithm for this kind of data set.

Next section presents our proposed modified symmetry level (MSSL) operator and the proposed MSSL operator will be used in our proposed LSK clustering algorithm for handling the data set with line symmetry property while preserving the advantage in the previous PSK algorithm.

3. The proposed modified symmetry similarity level operator

Given a set of data points, first the traditional *K*-means algorithm is used to obtain *k* temporary clusters. Next, we want to find the symmetrical line of each cluster by using the central moment technique (Gonzalez and Wood, 2002). The found symmetrical line will be used to measure the symmetry similarity level between two data points relative to that symmetrical line.

Suppose the given data set is covered by an $h \times w$ integer domain, the (p,q)th order moment is defined as

$$m_{pq} = \sum_{1 \leqslant x \leqslant h} \sum_{1 \leqslant y \leqslant w} x^p y^q f(x,y), \tag{2}$$

where $f(x,y)$ is set to 1 when $f(x,y)$ is the given data point at location (x,y) in one obtained cluster after running the *K*-means algorithm; otherwise $f(x,y)$ is set to 0. By Eq. (2), the centroid of the given data set for one cluster is defined to $\left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right)$. The central moment is defined as

$$u_{pq} = \sum_{1 \leqslant x \leqslant h} \sum_{1 \leqslant y \leqslant w} (x - \bar{x})^p (y - \bar{y})^q f(x,y), \tag{3}$$

where $\bar{x} = \frac{m_{10}}{m_{00}}$ and $\bar{y} = \frac{m_{01}}{m_{00}}$. According to the calculated centroid and Eq. (3), the major axis of each cluster can be determined by the following two items:

(a) The major axis of the cluster must pass through the centroid.
(b) The angle between the major axis and the x axis is equal to $\frac{1}{2}\tan^{-1}\frac{2u_{11}}{u_{20}-u_{02}}$.

Consequently, for one cluster, its corresponding major axis is thus expressed by $\left(\left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right), \frac{1}{2}\tan^{-1}\frac{2u_{11}}{u_{20}-u_{02}}\right)$.

The obtained major axis is treated as the symmetric line of the relevant cluster. We now define the proposed MSSL operator to measure the symmetry level between two data points relative to the same major axis. Our proposed MSSL operator contains two suboperators, namely, the modified distance similarity level (MDSL) operator and the modified orientation similarity level (MOSL) operator. The pro-

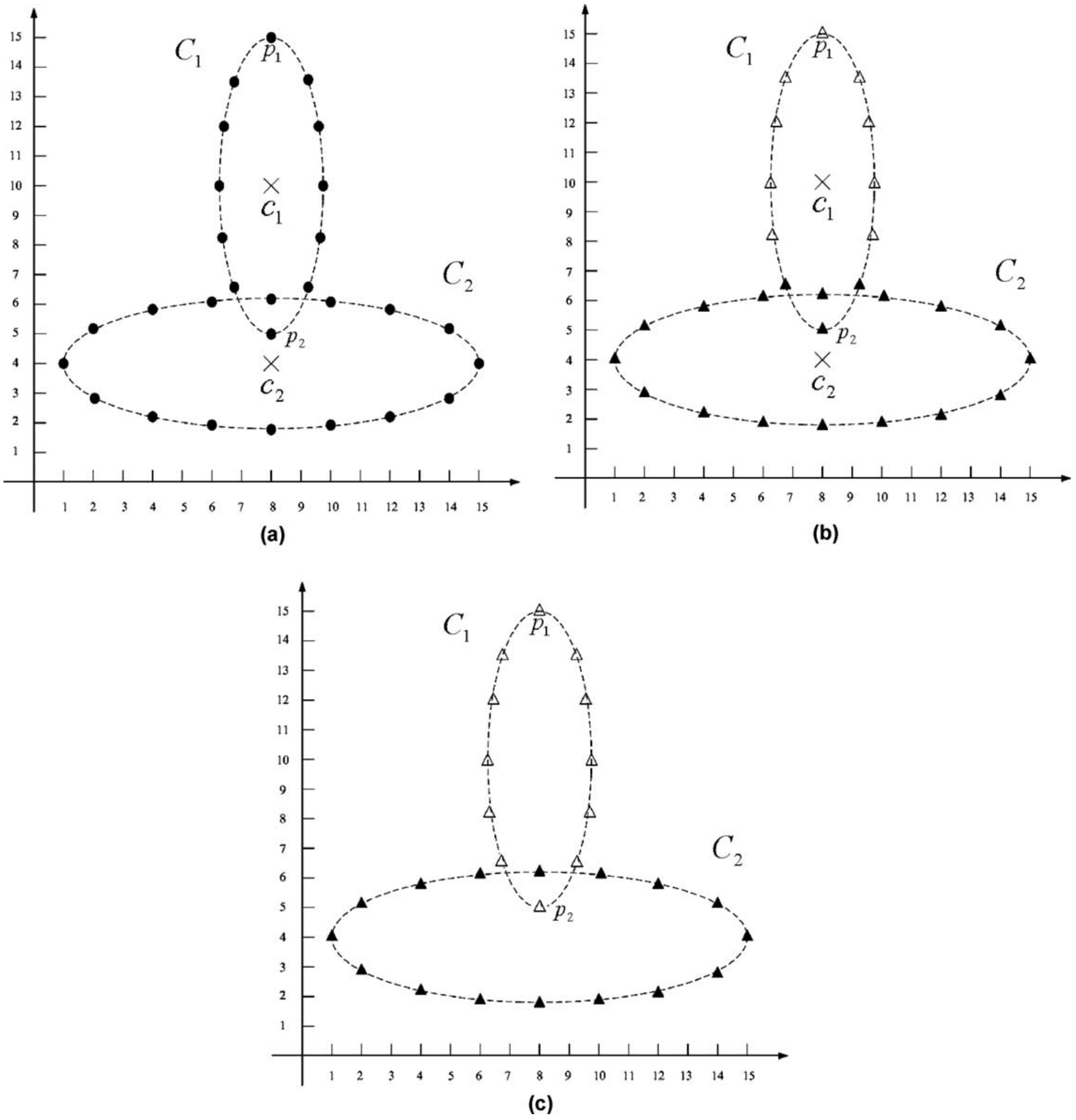

Fig. 1. One example to demonstrate the power of the PSK algorithm. (a) The given point symmetrical data set. (b) Two obtained clusters by running *K*-means algorithm on (a). (c) Two obtained clusters by running the PSK algorithm on (a).

posed two operators improve the previous DSL and OSL operators (Chung and Lin, in press). DSL operator and OSL operator are used to overcome the problems occurred in PSD measure (Su and Chou, 2001), and the problems are (1) lacking the distance difference symmetry property, and (2) leading to an unsatisfactory clustering result for the case of symmetrical inter-cluster.

An example is given to illustrate the first possible problem. In Fig. 2, there are four data points, namely the centroid c_k and the three data points p_i, p_j, and p_{j+1} at

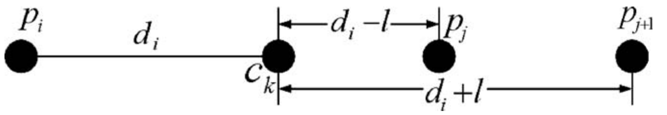

Fig. 2. An example for indicating the distance difference symmetry problem.

locations $c_k = (0,0)$, $p_i = (-d_i, 0)$, $p_j = (d_i - l, 0)$, and $p_{j+1} = (d_i + l, 0)$, respectively. It is known that the unit vectors of $\overrightarrow{p_ic_k}$, $\overrightarrow{c_kp_j}$, and $\overrightarrow{c_kp_{j+1}}$, i.e. $\frac{\overrightarrow{p_ic_k}}{\|\overrightarrow{p_ic_k}\|} = \frac{\overrightarrow{c_kp_j}}{\|\overrightarrow{c_kp_j}\|} = \frac{\overrightarrow{c_kp_{j+1}}}{\|\overrightarrow{c_kp_{j+1}}\|}$, are equivalent and the two related distance differences $\|\overline{p_ic_k} - \overline{c_kp_j}\|$ $(= l)$ and $\|\overline{p_ic_k} - \overline{c_kp_{j+1}}\|$ $(= l)$ are equivalent too. In Fig. 2, the most symmetrical point of p_i relative to the centroid c_k is the data point p_j or the data point p_{j+1}. By Eq. (1), we have

$$d_s(p_i, c_k) = \min\left\{\frac{\|(p_i - c_k) + (p_j - c_k)\|}{\|(p_i - c_k)\| + \|(p_j - c_k)\|}, \frac{\|(p_i - c_k) + (p_{j+1} - c_k)\|}{\|(p_i - c_k)\| + \|(p_{j+1} - c_k)\|}\right\} = \min\left\{\frac{l}{2d_i - l}, \frac{l}{2d_i + l}\right\} = \frac{l}{2d_i + l},$$

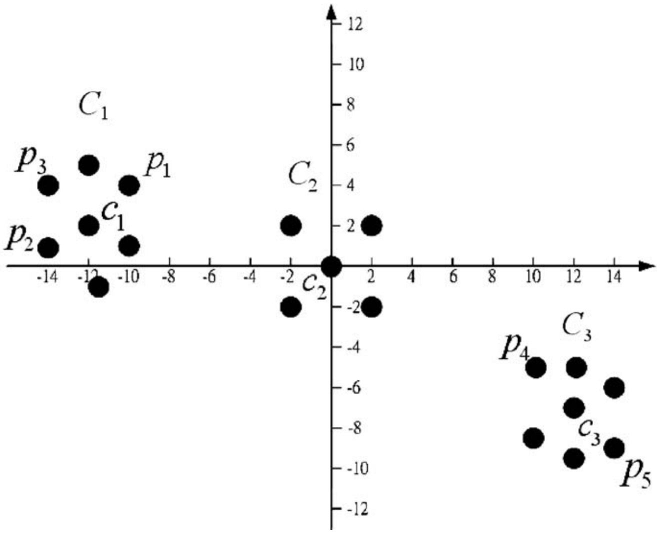

Fig. 3. An example of symmetrical intra-/inter-clusters.

so the data point p_{j+1} is selected as the most symmetrical point of p_i relative to the centroid c_k. This indicates that the PSD measure favors the far data point when we have more than two candidate data points and this may degrade the symmetrical robustness, and this is the first problem.

Next, a case of symmetrical inter-clusters (Chung and Lin, in press) is considered to explain why the PSD measure may generate an unsatisfactory clustering results. In Fig. 3, there are nineteen data points and the centroid of the first cluster C_1 is c_1; the centroid of the second cluster C_2 is c_2, and the centroid of the third cluster C_3 is c_3. Let $p_1=(-10,4)$, $p_2=(-14,1)$, $p_3=(-14,4)$, $p_4=(10,-5)$, $p_5=(14,-9)$, $c_1=(-12,2)$, $c_2=(0,0)$, and $c_3=(12,-7)$. We now consider the data point p_1, and by Eq. (1), it yields

$$\begin{aligned}d_s(p_1,c_1) &= \min_{1\leqslant i\leqslant 16, i\neq 1}\frac{\|(p_1-c_1)+(p_i-c_1)\|}{\|p_1-c_1\|+\|p_i-c_1\|}\\ &= \frac{\|(p_1-c_1)+(p_2-c_1)\|}{\|(p_1-c_1)\|+\|(p_2-c_1)\|}\\ &= \frac{1}{\sqrt{8}+\sqrt{5}}=0.20,\\ d_s(p_1,c_2) &= \min_{1\leqslant i\leqslant 16, i\neq 1}\frac{\|(p_1-c_2)+(p_i-c_2)\|}{\|p_1-c_2\|+\|p_i-c_2\|}\\ &= \frac{\|(p_1-c_2)+(p_4-c_2)\|}{\|(p_1-c_2)\|+\|(p_4-c_2)\|}\\ &= \frac{1}{\sqrt{116}+\sqrt{125}}=0.05,\end{aligned}$$

and

$$\begin{aligned}d_s(p_1,c_3) &= \min_{1\leqslant i\leqslant 16, i\neq 1}\frac{\|(p_1-c_3)+(p_i-c_3)\|}{\|p_1-c_3\|+\|p_i-c_3\|}\\ &= \frac{\|(p_1-c_3)+(p_5-c_3)\|}{\|(p_1-c_3)\|+\|(p_5-c_3)\|}\\ &= \frac{\sqrt{481}}{\sqrt{605}+\sqrt{8}}=0.80.\end{aligned}$$

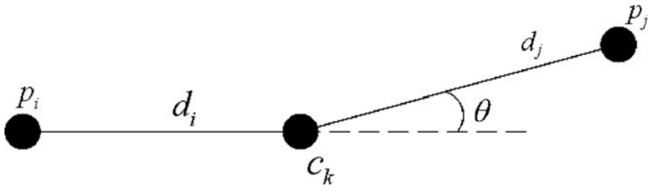

Fig. 4. Measure the distance difference symmetry.

From the above three PSD values, since $d_s(p_1,c_2)$ (=0.05) is the smallest, among the three concerning PSD values, the data point p_1 should be assigned to the cluster C_2, but it conflicts our visual inspection. From the data distribution of C_1, instead of assigning p_1 to C_2, it will be better to assign the data point p_1 to the cluster C_1. This is the second problem.

Due to the above two problems, Chung and Lin (in press) proposed the SSL operator which contains DSL operator and OSL operator to satisfy the distance difference symmetry property. In Fig. 4, c_k denotes the cluster centroid; p_i and p_j denote two related data points. Let $d_i=\overline{p_ic_k}$ and $d_j=\overline{p_jc_k}$, and then the distance similarity level (DSL) operator for measuring the distance difference symmetry between the distance $\overline{p_ic_k}$ and the distance $\overline{p_jc_k}$ is defined by

$$\mathrm{DSL}(p_i,c_k,p_j)=\begin{cases}1-\dfrac{|d_i-d_j|}{n\times d_i}, & \text{if } 0\leqslant\dfrac{d_j}{d_i}\leqslant n+1,\\ 0, & \text{otherwise,}\end{cases}$$

where the parameter n in the above equation is selected to be 1.

Besides the DSL operator, another operator used to proposed SSL operator, say orientation similarity level (OSL), was presented. Applying the projection concept (Hoffman and Kunze, 1961), the orientation similarity level between the two vectors, $v_i=\overrightarrow{p_ic_k}=(c_k-p_i)$ and $v_j=\overrightarrow{c_kp_j}=(p_j-c_k)$, is defined by

$$\mathrm{OSL}(p_i,c_k,p_j)=\frac{v_i\cdot v_j}{2\|v_i\|\|v_j\|}+0.5. \tag{4}$$

Now combining the DSL operator and OSL operator can obtain the symmetry similarity level (SSL) operator, and it is defined by

$$\mathrm{SSL}(p_i,c_k)=\max_{1\leqslant j\leqslant N}\sqrt{\frac{\mathrm{DSL}^2(p_i,c_k,p_j)+\mathrm{OSL}^2(p_i,c_k,p_j)}{2}}. \tag{5}$$

In their experiments (Chung and Lin, in press), the clustering result with SSL operator is better than PSD measure As shown in Fig. 5.

Examining the previous DSL operator in detail, one problem may happen. As shown in Fig. 6, the distance difference between the two data points d_i and d_j is constant, so intuitively the value of DSL for d_i or d_j relative to the centroid c_k must be the same. However, the above DSL operator violates this intuition since the value of DSL for point

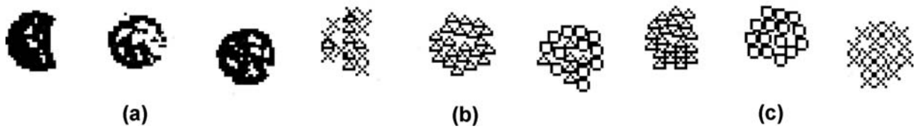

Fig. 5. Clustering performance comparison with SSL operator and PSD measure. (a) The data set contains three compact circles. Clustering result using PSD measure (b) and using SSL operator (c).

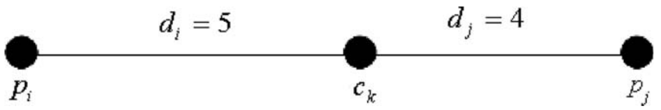

Fig. 6. An example to highlight the problem occurred in the previous DSL operator.

p_i relative to c_k is $1-\frac{1}{5}=0.8$, but the value of DSL for point p_j relative to c_k is $1-\frac{1}{4}=0.75$. In order to compensate the shortcoming and to satisfy the symmetry property, the DSL operator is modified into the following MDSL operator:

$$\mathrm{MDSL}(p_i,p_{ki},p_j)=\exp\left(-\frac{|d_i-d_j|}{\rho^2}\right), \tag{6}$$

where the value of ρ is specified by 5 empirically.

From Eq. (6), it is obvious that we have $0<\mathrm{MDSL}(p_i,p_{ki},p_j)\leqslant 1$ and the larger value of $\mathrm{MDSL}(p_i,p_{ki},p_j)$ is, the larger the distance similarity level between d_i $(=\overline{p_ip_{ki}})$ and d_j $(=\overline{p_jp_{ki}})$ is. When $d_i=d_j$, we have $\mathrm{MDSL}(p_i,p_{ki},p_j)=1$; it means that the distance $\overline{p_ip_{ki}}$ and the distance $\overline{p_jp_{ki}}$ have the highest MDSL values. The parameter ρ in Eq. (6) is used to control the tolerance of distance difference between two concerned data points, and empirically, the range of ρ can be selected from 5 to 10.

Besides the MDSL operator defined above, we continue using the original OSL operator, but for convenience, we change the symbol OSL to symbol MOSL, and it would be written as

$$\mathrm{MOSL}(p_i,p_{ki},p_j)=\frac{v_i\cdot v_j}{2\|v_i\|\|v_j\|}+0.5. \tag{7}$$

Combining the two operators, the MDSL and the MOSL, which are defined in Eqs. (6) and (7), respectively, our proposed MSSL operator to measure the symmetry similarity level between the two vectors, $\overrightarrow{p_ip_{ki}}$ and $\overrightarrow{p_{ki}p_j}$, is defined by

$$\mathrm{MSSL}(p_i,p_{ki},p_j)=\max_{1\leqslant j\leqslant N}\frac{\mathrm{MDSL}(p_i,p_{ki},p_j)+\mathrm{MOSL}(p_i,p_{ki},p_j)}{2} \tag{8}$$

for $1\leqslant k\leqslant K$ and $1\leqslant i\leqslant N$.

Because the values of $\mathrm{MDSL}(p_i,p_{ki},p_j)$ and $\mathrm{MOSL}(p_i,p_{ki},p_j)$ range from 0 to 1, it is easy to verify Eq. (8) that the value of $\mathrm{MSSL}(p_i,p_{ki},p_j)$ is also between 0 and 1. The larger $\mathrm{MSSL}(p_i,p_{ki},p_j)$ value is, the larger symmetry similarity level is. For the data point p_i relative to the projected point p_{ki} on the major axis of the corresponding cluster, our proposed MSSL operator is a good tool to find the most symmetrical data point p_j relative to p_{ki} where its $\mathrm{MSSL}(p_i,p_{ki},p_j)$ value is maximal among all the concerning data points.

4. The proposed line symmetry-based *K*-means algorithm

In this section, we present the proposed line symmetry-based *K*-means (LSK) algorithm which extends the previous PSK algorithm by Su and Chou from handling the point symmetrical data set to handling the point symmetrical data set, the line symmetrical data set, or both of them.

The proposed LSK algorithm adopts the conventional *K*-means algorithm as a preprocessing step, then utilizes the concept of a major axis and the proposed MSSL operator to measure the symmetry level of the concerning two data points. Our proposed LSK algorithm is shown below where Step 3 of the proposed algorithm constitutes the main contribution of this paper.

Step 1. (Select K initial cluster centroids)
Give N data points, we choose K data points randomly as the initial K cluster centroids.

Step 2. (Coarse-tuning by running the *K*-means algorithm)
Apply the *K*-means algorithm to update the selected K cluster centroids until the K cluster centroids are converged to fixed points or the terminating criteria is satisfied.

Step 3. (Fine-tuning)

Step 3.1. (Find the symmetrical line for each cluster)
As described in the first paragraph of Section 3, for each cluster, we use the moment-based approach to find out the relevant symmetrical line.

Step 3.2. (Prune impossible candidate symmetrical data points)
For each data point p_i, calculate the projected point p_{ki} on the relevant symmetrical line of cluster C_k, and then find out all possible candidate symmetrical data points p_j relative to each symmetrical line of the corresponding cluster such that $\mathrm{MDSL}(p_i,p_{ki},p_j)\geqslant\alpha$ (=0.6) and $\mathrm{MOSL}(p_i,p_{ki},p_j)\geqslant\beta$ (=0.97) are held, $1\leqslant i,j\leqslant N$ and $1\leqslant k\leqslant K$, where p_j belongs to the kth cluster already.

Step 3.3. (Search the most symmetrical data points among the candidates)
For data point p_i, find out the data point p_j relative to the symmetrical line of

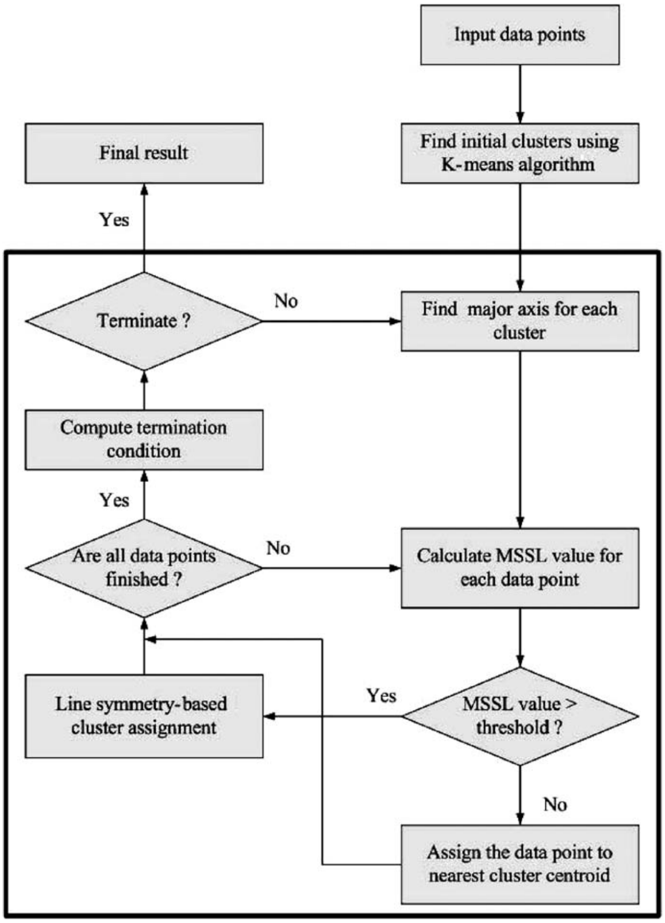
Input data points
Find initial clusters using K-means algorithm
Final result
Yes
Terminate ?
No
Find major axis for each cluster
Compute termination condition
Yes
Are all data points finished ?
No
Calculate MSSL value for each data point
Line symmetry-based cluster assignment
Yes
MSSL value > threshold ?
No
Assign the data point to nearest cluster centroid

Fig. 7. Flow chart for the proposed LSK algorithm.

cluster C_k such that the value of $\mathrm{MSSL}(p_i, p_{ki}, p_j)$ is the largest. If such a data point p_j does not exist since the relevant value of MSSL is still less than the threshold, then the data point p_i would be assigned to the cluster with the shortest Euclidean distance relative to the cluster centroid; otherwise the data point p_i is assigned to the kth cluster.

Step 4. (Update the centroid and the symmetrical line for each cluster)
After assigning all data points to the corresponding proper clusters, we then update the centroid and the symmetrical line for each cluster.

Step 5. (Continue or terminate)
If all the centroids are converged to fixed points or the number of iterations is larger than the allowable bound, stop the algorithm; otherwise go to Step 3.

After describing the above LSK algorithm, the following flow chart as shown in Fig. 7 is supplemented to make the relevant processes in the proposed LSK algorithm more clear.

5. Experimental results

In this section, several artificial and real data sets are used to demonstrate the feasibility and the extension capability of our proposed LSK algorithm. Experimental results reveal that our proposed LSK algorithm has encouraging results. Throughout the following experiments, the parameter ρ is selected to be five. In addition, the thresholds for MDSL and MOSL are selected to be 0.60 and 0.97, respectively.

Using the same experimental data set as in the PSK algorithm (Su and Chou, 2001), the given two crossed ellipsoidal shells data set with intra-symmetry property is shown in Fig. 8(a). Fig. 8(b) illustrates the clustering result by using the K-means algorithm, but the clustering result is unsatisfactory from the human visualization judgment. Fig. 8(c) illustrates the clustering result by using our proposed LSK algorithm and it has the same satisfactory clustering result as in the previous PSK algorithm.

Besides the data set with intra-symmetry property, the data set with both intra-symmetry property and inter-symmetry property used in (Chung and Lin, in press) is investigated in Fig. 9. The data set contains two compact ellipses and two crossed ellipsoidal shells as shown in Fig. 9(a). After running the K-means algorithm on Fig. 9(a), there are several misclassified data points as shown in Fig. 9(b)

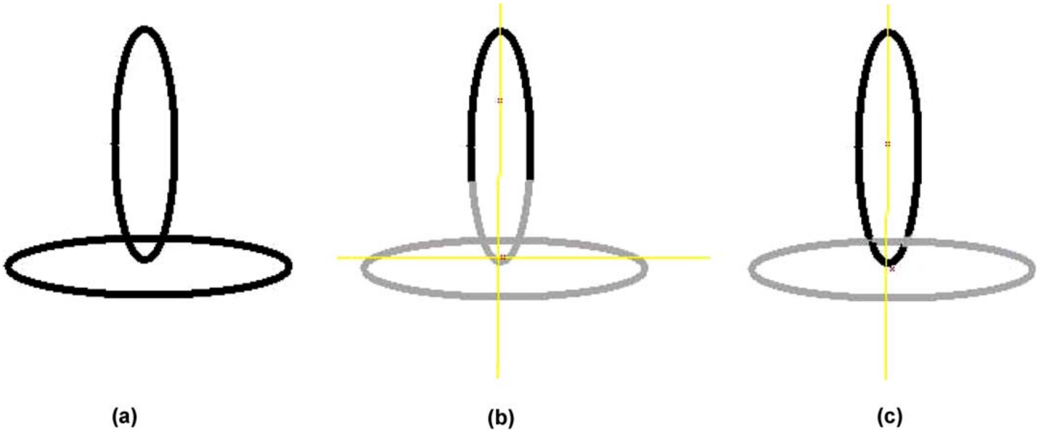

Fig. 8. Clustering results for data set with intra-symmetry property. (a) Given data points. Clustering result using K-means algorithm (b) and using our proposed LSK algorithm (c).

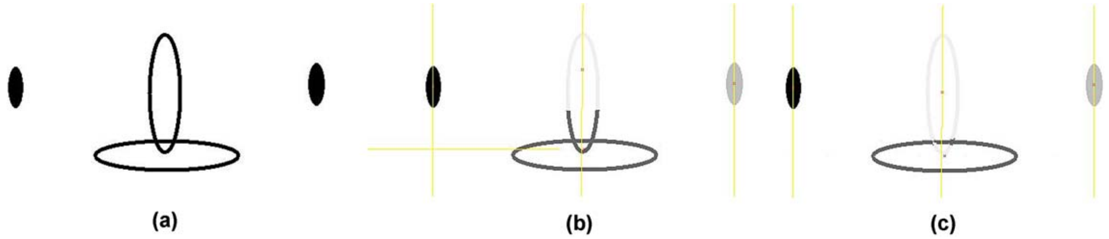

Fig. 9. Clustering results for data set with both intra- and inter-symmetry properties. (a) Given data points. Clustering result using *K*-means algorithm (b) and using our proposed LSK algorithm (c).

on the crossed ellipsoidal shells. Fig. 9(c) demonstrates the clustering result by using our proposed LSK algorithm and it has the same satisfactory clustering result as in the previous IPSK algorithm. According to the experimental results, it reveals that both the previous IPSK algorithm and our proposed LSK algorithm can cluster the data set with intra-symmetry property, the data set with inter-symmetry property, or both kinds of data sets quite well.

Experimental results in Figs. 8 and 9 have demonstrated that our proposed LSK algorithm can handle the clustering work for the data set with intra-symmetry property, the data set with inter-symmetry property, and the data set with both intra- and inter-symmetry properties. In what follows, two data sets are examined to confirm that our proposed LSK algorithm also can cluster data set with line symmetry property.

Most of natural scenes, such as leaves of plants, have the line symmetry property rather than the point symmetry property. Fig. 10(a) shows two real leaves of *Ficus microcapa* and *Wedelia trilobata* and they overlap a little each other. First the Sobel edge detector (Gonzalez and Wood, 2002) is used to obtain the edge pixels as the input data points which are shown in Fig. 10(b). After running the *K*-means algorithm, Fig. 10(c) demonstrates two temporary clusters and the corresponding two major axes which are obtained by using the moment-based technique. Furthermore, after running IPSK algorithm and our proposed LSK algorithm, the two clustering results are shown in Fig. 10(d) and (e), respectively. It is observed that IPSK algorithm cannot handle this case very well while our proposed LSK algorithm demonstrates a satisfactory clustering result.

Fig. 11(a) shows the two real leaves of *Erechtites valerianifolia* and they also overlap a little. Fig. 11(b) depicts all the data points of the two leaves. After running the *K*-means algorithm and using the moment-based technique, Fig. 11(c) demonstrates two temporary clusters and the corresponding two major axes. Furthermore, after running IPSK algorithm and our proposed LSK algorithm, the two clustering results are depicted in Fig. 11(d) and (e), respectively. It is obvious that IPSK algorithm cannot handle this

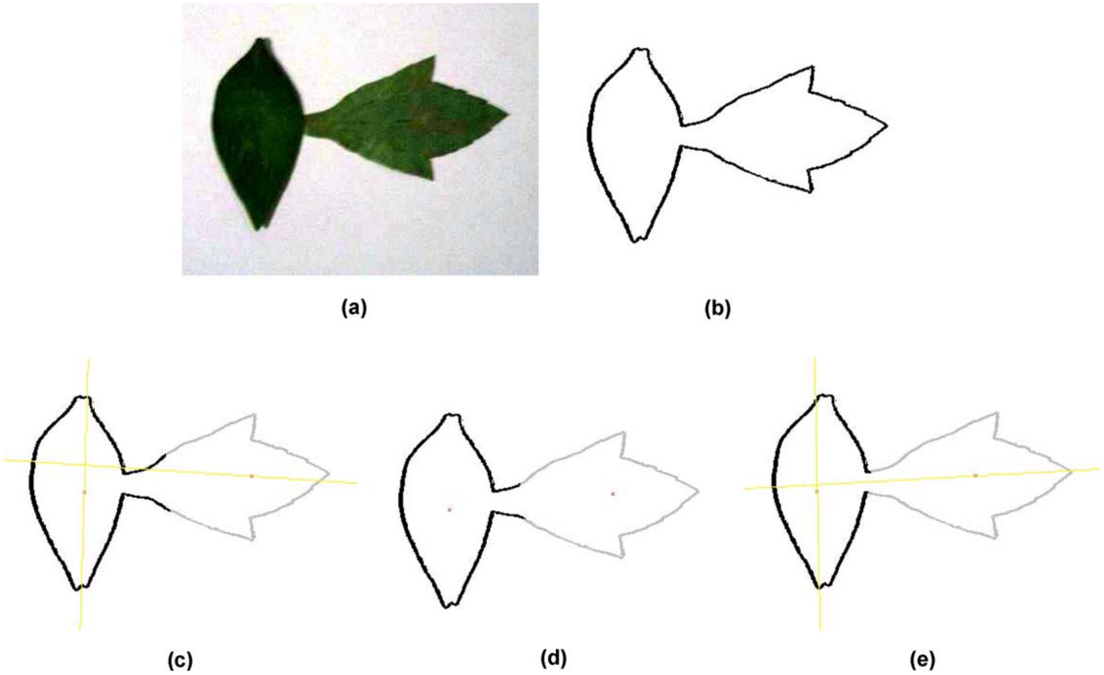

Fig. 10. Clustering result for the first data set with line symmetry property. (a) Two input leaves. (b) Edge pixels of leaves as input data points. Clustering result using *K*-means algorithm (c), using IPSK algorithm (d) and using our proposed LSK algorithm (e).

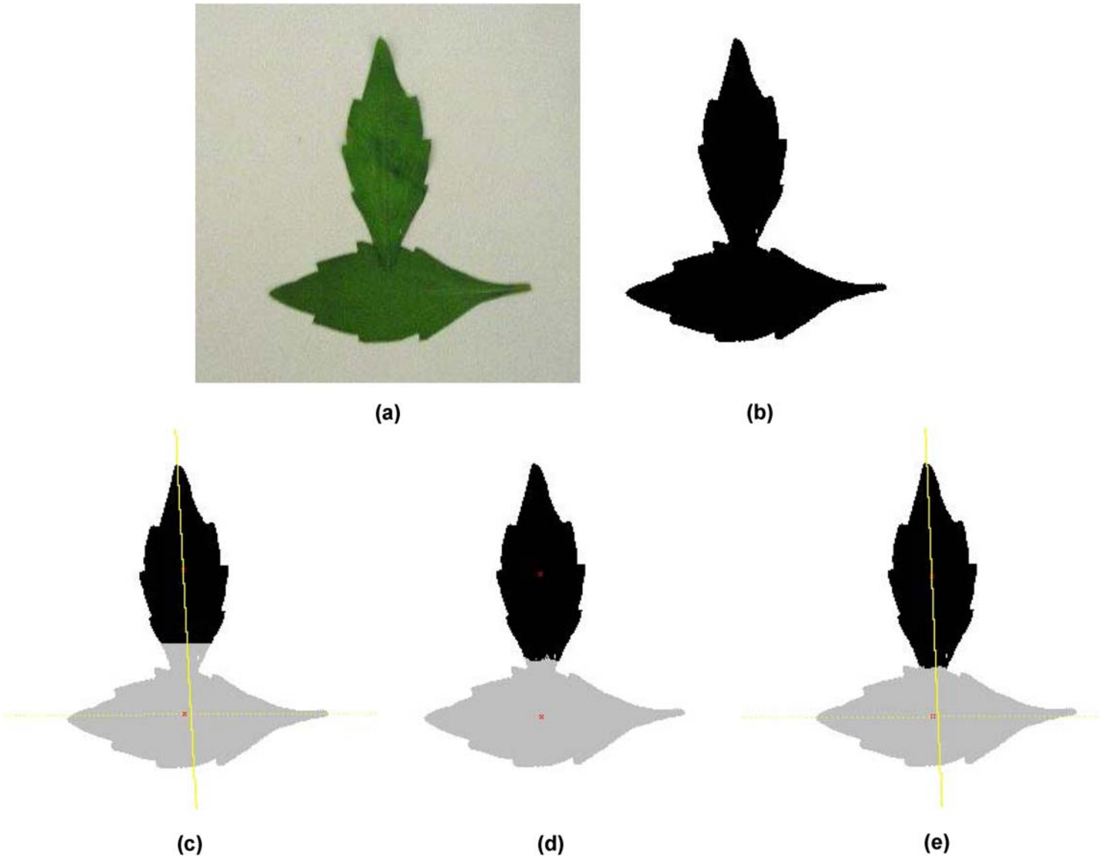

Fig. 11. Clustering result for the second data set with line symmetry property. (a) Two input leaves. (b) Leaves as input data points. Clustering result using *K*-means algorithm (c), using IPSK algorithm (d) and using our proposed LSK algorithm (e).

case very well. However, our proposed LSK algorithm illustrates a satisfactory clustering result.

6. Conclusions

In this paper, we have presented the line symmetry-based *K*-means algorithm. The proposed new clustering algorithm not only can cluster data sets with the property of line symmetry successfully, but also preserves the clustering advantages in the previous PSK algorithm and the previous IPSK algorithm. Under some real data sets, experimental results demonstrate that the feasibility of our proposed line-symmetry based *K*-means algorithm and the relevant experimental results are rather encouraging. Other than the clustering experiments using leaf example, it is an interesting future research topic to extend the results of this paper to face recognition.

References

Bajcsy, P., Ahuja, N., 1998. Location and density based hierarchical clustering using similarity analysis. IEEE Trans. Pattern Anal. Machine Intel. 20 (9), 1011–1015.

Chung, K.L., Lin, J.S., in press. Faster and more robust point symmetry-based *K*-means algorithm. Pattern Recognit., doi:10.1016/j.patcog.2005.09.015.

Fischer, B., Buhmann, J.M., 2003. Bagging for path based clustering. IEEE Trans. Pattern Anal. Machine Intel. 25 (11), 1411–1415.

Fred, L.N., Leitao, M.N., 2003. A new cluster isolation criterion based on dissimilarity increments. IEEE Trans. Pattern Anal. Machine Intel. 25 (8), 944–958.

Gersho, A., Gray, R.M., 1992. Vector Quantization and Signal Compression. Kluwer, Norwell, MA.

Gonzalez, R.C., Wood, R.E., 2002. Digital Image Processing, second ed. Prentice Hall, New Jersey.

Hartigan, J., 1975. Clustering Algorithms. Wiley, New York.

Hoffman, K., Kunze, R., 1961. Linear Algebra. Prentice Hall, New Jersey.

Hu, M.K., 1962. Visual pattern recognition by moment invariants. IEEE Trans. Inform. Theory 8 (2), 179–187.

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ.

Kanatani, K., 1997. Comments on "Symmetry as a continuous feature". IEEE Trans. Pattern Anal. Machine Intel. 19 (3), 246–247.

Ng, R.T., Han, J., 2002. CLARANS: A method for clustering objects for spatial data mining. IEEE Trans. Knowledge Data Eng. 14 (5), 1003–1016.

Sayood, K., 1996. Introduction to Data Compression. Morgan Kaufmann, San Francisco.

Su, M.C., Chou, C.H., 2001. A modified version of the *K*-means algorithm with a distance based on cluster symmetry. IEEE Trans. Pattern Anal. Machine Intel. 23 (6), 674–680.

Zabrodsky, H., Peleg, S., Avnir, D., 1995. Symmetry as a continuous feature. IEEE Trans. Pattern Anal. Machine Intel. 17 (12), 1154–1166.

Zhu, C., Po, L.M., 1998. Minimax partial distortion competitive learning for optimal codebook design. IEEE Trans. Image Process. 7 (10), 1400–1409.