Pattern Recognition Letters 34 (2013) 155–162

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A heuristic hierarchical clustering based on multiple similarity measurements

Chun-Zhong Li [a,*,1], Zong-Ben Xu [b,1], Tao Luo [b]

[a] Institute of Statistics and Applied Mathematics, Anhui University of Finance & Economics, BengBu 233030, China
[b] Institute for Information System Science, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 7 March 2012
Available online 16 October 2012

Communicated by S. Sarkar

Keywords:
Data mining
Agglomerative clustering
Heuristic
Blurring
Top-down
Structural nearest neighbor

ABSTRACT

Similarity is the core problem of clustering. Clustering algorithms that are based on a certain, fixed type of similarity are not sufficient to explore complicated structures. In this paper, a constructing method for multiple similarity is proposed to deal with complicated structures of data sets. Multiple similarity derives from the local modification of the initial similarity, based on the feedback information of elementary clusters. Combined with the proposed algorithm, the repeated modifications of local similarity measurement generate a hierarchical clustering result. Some synthetic and real data sets are employed to exhibit the superiority of the new clustering algorithm.

1. Introduction

Clustering is an important technology in data mining and pattern recognition, and it is the process of grouping a set of data into some meaningful groups based on similarity. In the grouping process the intra-group similarity is maximized and the inter-group similarity is minimized. The group formed is usually called a cluster. This technology is widely applied in text retrieval (Abolhassani and Mahdavi, 2009), image segmentation (Pichel et al., 2006; Mukherjee, 2002), image quantization (Schenders, 1997), and so on.

Similarity is of crucial importance for clustering. Each clustering algorithm pays much attention to similarity measurement. Most of the prevailing algorithms possess their own similarity definition or particular processing methods. The commonly-used similarity definitions are mainly based on L_2-distance, L_1-distance, Gaussian kernel function, etc. Each algorithm performs well on certain types of data sets or particular applications. The common character of the existing algorithms is that each of them has its own single and fixed similarity definition, which is not sufficient for data sets of complicated structural features.

Multiple similarity is necessary for clustering, for it enables the exploration of data sets of complicated structural features. In practical applications, there widely exist data sets with complicated structures. For an image shown in Fig. 1(a), the locations of pixels in color space are shown in Fig. 1(b), which exhibits a data set of complicated structures. In the color space, the structures are mixed and complicated, including density difference, connectedness, and direction, etc., which is difficult to explore with a single and fixed similarity definition.

A complicated structural data set may need various kinds of similarity measurements simultaneously. The research on multiple similarity is mainly in the fields of classification, such as multiple kernel learning (MKL) (Lanckriet et al., 2004; Sonnengurg et al., 2006). MKL is a supervised learning process which selects various kernels automatically. Compared with classical classification algorithm, such as SVM, the kernels in MKL are different though the kernel type is given. MKL takes the structures of data set into account, which improves its performance and generalization ability. Compared with MKL in classification, multiple similarity is difficult in clustering for there is no priors, and it is also important to obtain better clustering results.

In this paper, we propose a constructing method for multiple similarity in clustering. Elementary clusters emerge with initial similarity and proposed algorithm. Then, the structural information of the elementary clusters is gathered to information feedback, and the initial similarity is locally modified with the guidance of the feedback information, thus forming different similarity measurements suitable to redetermine the relations between elementary clusters. After that, new elementary clusters appear, and the structural information of them directs another modification of initial similarity to form new similarity measurements suitable for structural features of new elementary clusters. The process is repeated to generate a heuristic hierarchical clustering, until only one cluster remains.

The main contributions of multiple similarity mechanism contains: (i) Multiple similarity enables the identifying of outliers to be conducted simultaneously with the sewing up of elementary

* Corresponding author.
E-mail address: lichunzhongli@gmail.com (C.-Z. Li).
[1] Supported by the National 973 Project of China (Grant No. 2013CB329404).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.09.025

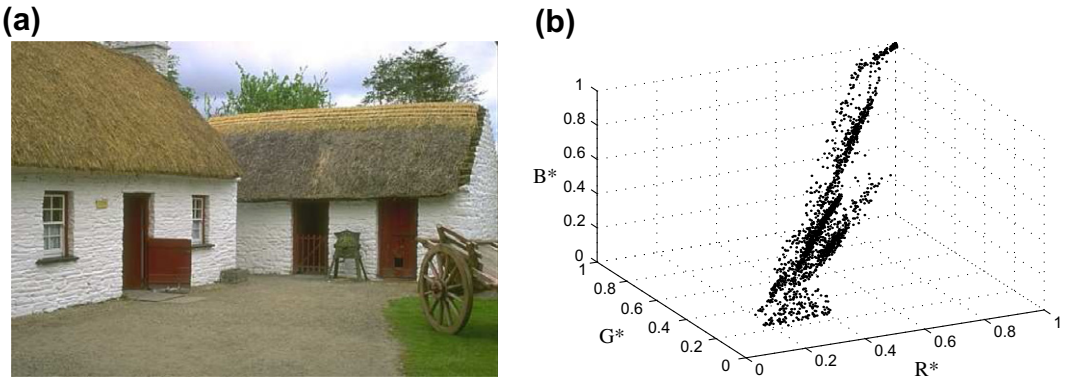

Fig. 1. An color image and locations of pixels in R*G*B* color space. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

clusters. (ii) Multiple similarity derives from the local revision of initial similarity in the heuristic process, which does not increase the computing complexity of clustering algorithm, $O(NlogN)$. (iii) Multiple similarity enlarges the application scope of the algorithm, and the clustering results are stable.

The remainder of the paper is as follows: Section 2 introduces three closely related algorithms; then the clustering framework is given in Section 3; following the clustering framework, multiple similarity mechanism is introduced in Sections 4–6 introduce the clustering algorithm and its analysis, respectively; experiments and conclusion are given in Section 7 and 8.

2. Iterative local centroid estimation

Among various clustering algorithms, iterative local centroid estimation is one of the typical approaches, the representatives of the approach include mean shift (MS) (Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu and Meer, 2002), clustering by scale space filtering (CSSF) (Leung et al., 2000), and Gaussian blurring mean shift (GBMS) (Carreira-Perpinan, 2004). The three algorithms are closely related with the new algorithm, and the new clustering framework is introduced in the following part.

$X = \{x_i\}_{i=1}^N$ is a data set and x_i is a d-dimensional row vector representing a pattern. By using the classical Gaussian kernel density estimation (Bishop, 1999), the distribution of the data set X can be represented as

$$p(x;\sigma) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(\sigma\sqrt{2\pi})^2}e^{-\frac{||x-x_i||^2}{2\sigma^2}}, \tag{1}$$

where σ is a scale parameter, and the derivation of $p(x;\sigma)$ is

$$\begin{aligned}\nabla_x p(x;\sigma) &= \frac{1}{\sigma^2 N}\sum_{i=1}^{N}\frac{(x_i-x)}{(\sigma\sqrt{2\pi})^2}e^{-\frac{||x-x_i||^2}{2\sigma^2}} \\ &= \frac{1}{2\pi\sigma^4}\left[\sum_{i=1}^{N}e^{-\frac{||x_i-x||^2}{2\sigma^2}}\right]\left[\frac{\sum_{i=1}^{N}x_i e^{-\frac{||x-x_i||^2}{2\sigma^2}}}{\sum_{i=1}^{N}e^{-\frac{||x-x_i||^2}{2\sigma^2}}} - x\right].\end{aligned} \tag{2}$$

In order to obtain the local maximums of $p(x;\sigma)$ and let them be centroids, $\nabla_x p(x;\sigma) = 0$ is needed. According to Eq. (2), the iterative numerical solution can be represented as

$$x(n+1) = \sum_{i=1}^{N}\frac{x_i e^{-\frac{||x(n)-x_i||^2}{2\sigma^2}}}{\sum_{i=1}^{N}e^{-\frac{||x(n)-x_i||^2}{2\sigma^2}}}. \tag{3}$$

Eq. (3) is called mean shift algorithm.

Clustering by scale space filtering (Leung et al., 2000) (CSSF) is different in approach but equally satisfactory in result with MS. Compared with MS, the biggest character of CSSF is that it is a hierarchical clustering with a series of scale parameters,

$$x(n+1) = \sum_{i=1}^{N}\frac{x_i e^{-\frac{||x(n)-x_i||^2}{2\sigma(t)^2}}}{\sum_{i=1}^{N}e^{-\frac{||x(n)-x_i||^2}{2\sigma(t)^2}}}, \tag{4}$$

where the parameter $\sigma(t)$ satisfies $\sigma(t) - \sigma(t-1) = 0.029$. MS can be viewed as a solution of CSSF with fixed σ, or CSSF can be viewed as nested MS algorithm. With certain σ, CSSF and Mean Shift are the same procedure to find local maximums along with gradient direction of Eq. (1).

Gaussian blurring mean shift (Carreira-Perpinan, 2004) (GBMS) is a revised version of MS. GBMS, compared with MS and CSSF, is also a procedure with limited convergence. The difference is that its filter and convolution data do not come from original data x but from $x(n)$, and the iterative scheme is represented as

$$x(n+1) = \sum_{i=1}^{N}\frac{x(n)e^{-\frac{||x(n)-x_i(n)||^2}{2\sigma^2}}}{\sum_{i=1}^{N}e^{-\frac{||x(n)-x_i(n)||^2}{2\sigma^2}}}. \tag{5}$$

Since Eq. (5) considers only the structures of updated data $x(n)$, it converges with much fewer steps compared with mean shift and CSSF. However, the computing complexity of its each iteration does not reduce.

3. A heuristic hierarchical clustering

In this section, a heuristic hierarchical clustering is introduced. Compared with the above-mentioned three algorithms, the clustering can accept various structural features, which can be applied in multiple similarity.

3.1. A heuristic hierarchical clustering framework

In Eqs. (3)–(5), the term $e^{-\frac{||x-x_i||^2}{2\sigma^2}}$ can be viewed as a similarity (relationship) measurement between two data points, $S(x,x_i;\sigma) = e^{-\frac{||x-x_i||^2}{2\sigma^2}}$.

The three mentioned algorithms are all based on iterative local centroid estimation, and the distribution of original data set affects clustering result directly or indirectly. From cognitive view of human (Santos and Marques, 2005), it is reasonable that the structure of original data set plays a decisive role in and has persistent effect on clustering results, so we proposed a new hierarchical algorithm – constant strength general gaussian blurring mean shift (Li and Xu, 2011)

$$x_i(n+1) \leftarrow \sum_j \frac{x_{(j|i)}(n)\cdot S(x_i, x_{(j|i)};\Theta)}{\sum_j S(x_i, x_{(j|i)};\Theta)}, \tag{6}$$

where $S(\cdot,\cdot)$ is a filtering template which can be modeled with general Gaussian kernel, and Θ is a set containing some parameters to determine filtering template S. $\langle j|i\rangle$ is the jth nearest neighbor of the ith data point. Compared with GBMS and MS, the convergence of Eq. (6) can be proved much easier. However, it's real contribution lies in its information processing mechanism itself, the procedure of the hierarchical clustering. The model can integrate various kinds of information with S, and it's computing complexity is approximately linear in application to large data set.

In consideration of the advantages of Eq. (6) and the aim to obtain multiple similarity measurements to mine the complicated structure of data set, we proposed a heuristic hierarchical clustering framework

$$x_i(n+1) \leftarrow \sum_j \frac{x_{\langle j|i\rangle}(n)\cdot S(x_i, x_{\langle j|i\rangle}; t, \Theta)}{\sum_j S(x_i, x_{\langle j|i\rangle}; t, \Theta)}, \quad n = 0, 1, \ldots, \tag{7}$$

and S derives from top-down procedure

$$S(x_i, x_{\langle j|i\rangle}; t, \Theta) \overset{top-down}{\leftarrow} S(x_i, x_{\langle j|i\rangle}; 0, \Theta), \quad t = 1, 2, \ldots. \tag{8}$$

Compared with Eq. (6), Eq. (7) has different similarity measurement form, $S(x_i, x_{\langle j|i\rangle}; t, \Theta)$, which is related to t. Once t is fixed, similarity measurement S is obtained, and the clustering procedure of Eq. (7) is similar to Eq. (6). Data points with same structure feature will shrink into their centroid as in iterative local centroid estimation. So the cluster C can be recognized as

$$C = \left\{x_i, x_j \mid ||x_i(n) - x_j(n)|| = 0; \quad i \neq j; \quad i, j = 1, \ldots, N\right\}, \tag{9}$$

and the stop condition of Eq. (7) is

$$\sum_{i=1}^{N} ||x_i(n+1) - x_i(n)||^2 = 0. \tag{10}$$

Each t corresponds to a complete iterative scheme of Eq. (7) and an elementary clustering result, and the elementary clusters provide information feedback to guide the local revision of initial similarity measurement, and finally help to form new similarity measurement, $S(x_i, x_{\langle j|i\rangle}; t, \Theta)$. The symbol "$\overset{top-down}{\leftarrow}$" in Eq. (8) represents the local revision of $S(x_i, x_{\langle j|i\rangle}; 0, \Theta)$. This operator locally revises initial similarity between some pairs of data points automatically rather than similarity between all pairs of data points. Following different t, we can obtain different multiple similarity measurements $S(x_i, x_{\langle j|i\rangle};\ t, \Theta)$. The following section introduces the way to obtain the multiple similarity measurement in the top-down procedure.

4. Multiple similarity measurement

Similarity definition depends on the features that users pay attention to and may be multiple in the same data set, especially for complicated structural data sets. With the new clustering framework, $S(x_i, x_{\langle j|i\rangle}; t)$ contains multiple kinds of similarity with the same step t. Different from multiple kernel learning, the number of kernels or revised regions does not need to be given.

4.1. Initial similarity definition

Without any information gathered for feedback, the cognitive features of human and application background play leading roles in clustering. For a common data set in feature space, the density difference feature and proximity feature are our primary concerns, though some special data points need more complicated cognitive features. The initial similarity is defined as

$$S\left(x_i, x_{\langle j|i\rangle}; 0, \{c, \alpha, k\}\right) = e^{-c\cdot||x_i - x_{\langle j|i\rangle}||} \cdot e^{-\alpha\frac{|\rho(x_i;k) - \rho(x_{\langle j|i\rangle};k)|}{\rho(x_i;k)+\rho(x_{\langle j|i\rangle};k)}}, \tag{11}$$

where 0 represents $t = 0$, and parameter set $\{c, \alpha, k\}$ contains three parameters to determine the filtering template S. $\rho(x_i; k)$ represents the average distance (Yousri and Kamel, 2009) between x_i and its nearest neighbors

$$\rho(x_i; k) = \frac{1}{k}\sum_{j=1}^{k} ||x_i - x_{\langle j|i\rangle}||. \tag{12}$$

Eq. (11) considers two structural features: density difference feature and proximity feature. The first term of Eq. (11) reflects proximity of data points, parameter c is a scale parameter, whose value controls the affecting region. The larger the value is, the smaller the affecting region is, and vice versa. The second term of Eq. (11) reflects the density difference feature of data set. The local density is measured with average distance between data point and its nearest neighborhood, and parameter c controls the density difference between two data points. Its value is directly proportional to identifying strength of difference, and $c = 0$ corresponds to the neglect of the density difference.

With heuristic clustering model, Eqs. (7) and (8), the initial similarity is revised locally with gathered information in top-down process.

4.2. Multiple similarity measurement with top-down process

With initial similarity definition $S(x_i, x_{\langle j|i\rangle}; 0, \Theta)$ and Eq. (7), many elementary clusters emerge, mainly based on local structures of data set. The emergence of these elementary clusters enrich the structural features of data set, which can be gathered to guide the following clustering procedure. For these gathered structural features, the large scale features are paid more attention, but the density difference should not be ignored. And the multiple similarity measurement has a specific model

$$S\left(x_i, x_{\langle j|i\rangle}; t, \{\alpha, k\}\right) = link(x_i, x_{\langle j|i\rangle}; k+t) \cdot e^{-\alpha\frac{|\rho(x_i;k) - \rho(x_{\langle j|i\rangle};k)|}{\rho(x_i;k)+\rho(x_{\langle j|i\rangle};k)}}, \tag{13}$$

where t is an integer greater than zero. $link(x_i, x_{\langle j|i\rangle})$ represents the linkage between x_i and its neighbor $x_{\langle j|i\rangle}$, which is modeled as

$$link(x_i, x_{\langle j|i\rangle}; k+t) = \frac{1}{2}\left[sgn\left(R(x_i, x_{\langle j|i\rangle}; k+t)\cdot(-1)^{sgn\left[N\cdot(k+t) - \sum_{l=1}^{N}\sum_{j=1}^{k+t} sgn R(x_l, x_{\langle j|l\rangle}; k+t)\right]}\right) + 1\right]. \tag{14}$$

In Eq. (14), $R(x_i, x_{\langle j|i\rangle})$ considers large scale structural feature with direction consistency measurement (DCM), $R(x_i, x_{\langle j|i\rangle})$ (Li et al., 2011). The DCM between x_i and its neighbor $x_{\langle j|i\rangle}$ is represented as

$$R(x_i, x_{\langle j|i\rangle}; k) = d\cdot\log\frac{k^2 - k}{k^2 - k + 1} + \log\left[\det\left(I + \left(D^T D\right) \setminus \left((x_{\langle j|i\rangle} - x_i)^T (x_{\langle j|i\rangle} - x_i)\right)\right)\right], \tag{15}$$

where $\setminus$ is left inverse of matrix operator and I is a unit matrix; D is a matrix, each column of which is the difference between each pair of neighbors of data point x_i. The basic idea of DCM is that it can measure the consistence with direction $x_{\langle j|i\rangle} - x_i$ and the directions, $\left\{x_{\langle j|i\rangle} - x_{\langle j'|i\rangle} | j \neq j'; j,\ j' = 1, \ldots k\right\}$. The more consistent the directions are, the much smaller the value is. The DCM has two advantages: (i) its value can be negative or positive. The negative value represents the consistency of the local directions, while the positive value represents the inconsistency. Meanwhile, the value can be used to detect outliers, which is introduced in next subsection. (ii) It can be

used to merge elementary clusters with similar manifold features. If the neighborhood has no distinct principal direction, DCM can degenerate into distance-based relationship only, e.g., k-nn.

$link(x_i, x_{\langle j|i\rangle})$ in Eq. (13) is to locally update $S(x_i, x_{\langle j|i\rangle}; t = 0)$ by $S(x_i, x_{\langle j|i\rangle}; t \neq 0)$. This update does not involve x_i and all its neighbors, but x_i and some of its neighbors. The main aim of Eq. (14) is to sew up different elementary clusters with same manifold feature and cut off the relation between misclassified two data points belonging to a same elementary cluster according to $R(x_i, x_{\langle j|i\rangle})$. Eq. (14) only contains two value, 1 and 0. 1 represents sewing up and 0 represents cutting off. $link(x_i, x_{\langle j|i\rangle})$ involves two cases: (i) If $R(x_i, x_{\langle j|i\rangle}) < 0$, the two data points belonging to different elementary clusters are linked; if $R(x_i, x_{\langle j|i\rangle}) > 0$, the relation between two data points in a same elementary clusters are cut off; (ii) If $R(x_i, x_{\langle j|i\rangle}) > 0$, for each x_i and $x_{\langle j|i\rangle}$, the data set has no local direction feature or manifold feature, and $link(x_i, x_{\langle j|i\rangle})$ is forced to be 1 with Eq. (14). In this case, the link is distance-based only.

In the heuristic clustering procedure, Eq. (7) and (8), the multiple similarity measurement $S(x_i, x_{\langle j|i\rangle}; t, \Theta)$ is only local revision of $S(x_i, x_{\langle j|i\rangle}; 0, \Theta)$ and not all pairs of data points between x_i and its neighbors are revised, but only a few related data points according to gathered structural feedback information. The procedure is to be described with the example shown in Fig. 2.

The data set shown in Fig. 2(a) is composed of two manifold clusters and a noise. In consistence with design and practical demand, the two manifold clusters should merge together in the hierarchical process, not the noise. It is difficult to conduct without any pretreatment of the data set, but it is existing and necessary in practical application as shown in Fig. 1(b).

With initial similarity, Eq. (11), the similarities between each pair of data points are obtained and shown as pixels in Fig. 2(b). The pixels that are painted with black represent that the similarity between two data points is greater than zero, and the pixels painted with white represent that the similarity is zero. The figure shows that the noise is misclassified into one cluster. With information feedback, there are two kinds of local revision of initial similarity. One is to change the similarity that is zero to be positive value, which is painted with red, and the other is to change the similarity that is greater than zero to be zero, which is boxed with blue rectangle.

From Fig. 2(b), it is obvious that the revision of the initial similarity is minor, only the pixels painted with red and blue.

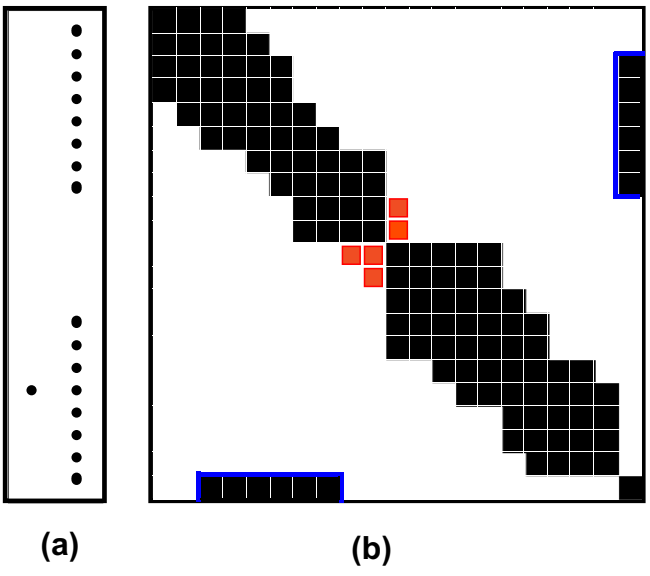

Fig. 2. An example.

The pixels painted with red represent the sewing up of data points, and pixels boxed with blue rectangle indicate the cutting off of the relationship between two data points. Moreover, the sewing up and cutting off are conducted simultaneously in the heuristic process.

5. Heuristic hierarchical clustering algorithm

Heuristic hierarchical clustering framework, Eq. (7) and (8), is a blurring algorithm. The blurring procedure of the framework is based on the similarity measurements, and the multiple similarity measurements derive from the heuristic procedure. Elementary clusters are generated in bottom-up procedure, and then the similarity is revised locally according to the cluster information with the top-down procedure. The heuristic hierarchical clustering procedure is shown in Fig. 3, in which the multiple similarity measurements are obtained.

From the algorithm flow, Fig. 3, many elementary clusters emerge with Eq. (7) based on the similarity Eq. (11). Some of these elementary clusters may have same structural features (e.g., density difference and manifold feature) and need to be re-clustered into a same one. With top-down procedure, the relationship between x_i and $x_{\langle j|i\rangle}$ is reconsidered. If two data points are in different elementary clusters and $link(x_i, x_{\langle j|i\rangle}) = 1$, they are regarded to have relation and are connected with red line as shown in Fig. 4. These red lines in the figure sew up the different elementary clusters that have same structural features. Clusters 3 and 4 are sewed up with redlines and shown in bottom-right of Fig. 4. Although some elementary clusters are linked with redlines in up-left of Fig. (4),

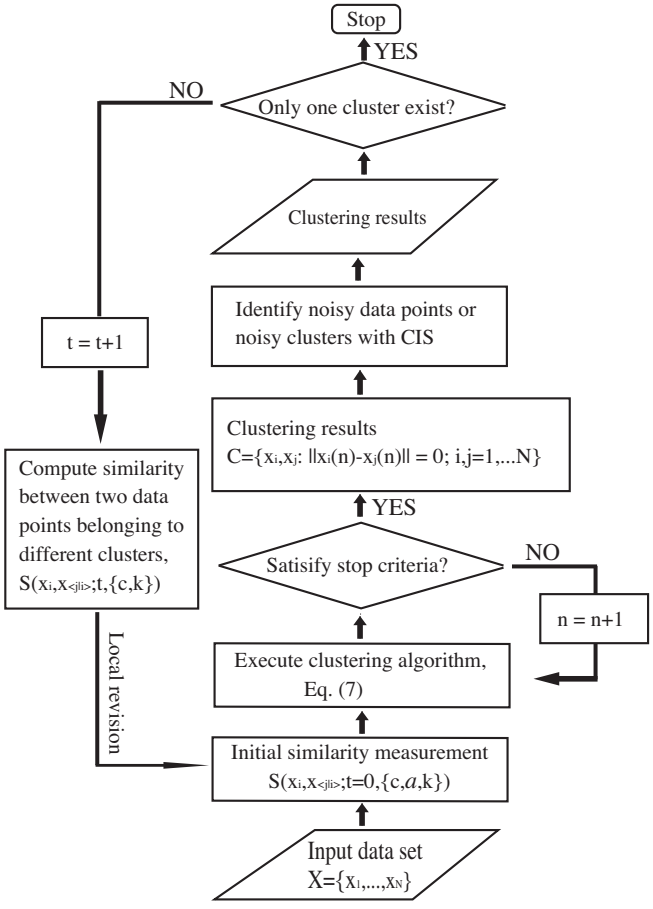

Fig. 3. The flow chart of the new clustering algorithm, which contains the construction of initial similarity, local modification of similarity with information feedback.

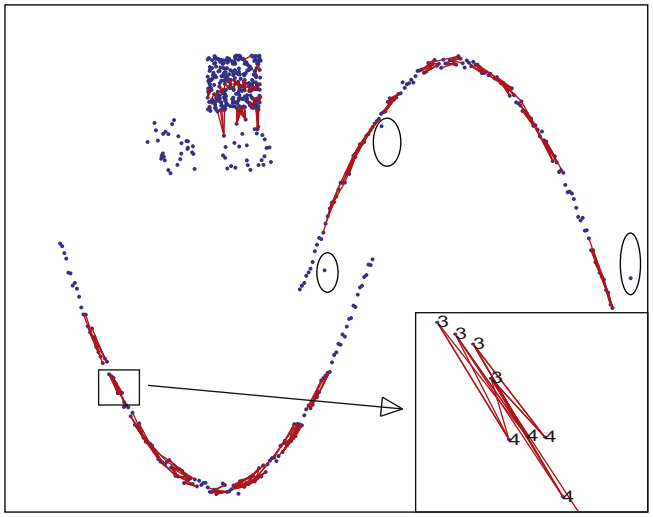

Fig. 4. Outlier identification and cluster reclassification. Redline ($link = 1$) connect elementary clusters, and outliers detected are labeled with ovals.

the linkage should be cutting off because they have different density features according to Eq. (13). Note that $link(x_i, x_{\langle j|i\rangle})$, Eq. (14), can degenerate into connection based on k-nn if the elementary clusters have no manifold structures.

The outliers (or noises) make the structural features difficult to recognize, so the detection of noises and outliers is needed in clustering process. In the new clustering algorithm shown in Fig. 3, the noises and outliers identification have a same criterion – cluster imbalance standard (CIS).

CIS is based on the intermediate clustering result generated by Eqs. (7) and (9), which shifts data points with similar structural features to centroid. $\#C_i$ represents the number of data points cluster C_i contains. In the new algorithm, the noises identification is based on imbalance of clusters. If $\#C_i$ is smaller compared with others, then cluster C_i is viewed as noisy cluster.

The effective clustering result is determined by the lifetime of the elementary clusters generated in the hierarchical clustering. With the increase of t, if the clustering result in a layer does not change within $[t_1, t_2]$, then the lifetime of the clustering result is $t_2 - t_1$. The clustering result with the longest lifetime is regarded as the effective clustering result.

6. Parameters and complexity of the clustering algorithm

The new clustering framework involves three parameters: c, α and k. Despite the involvement of three parameters, the robustness of clustering framework itself makes the parameters very stable and easy to execute.

Parameter c corresponds to the scale parameter in Gaussian kernel and controls the affecting region of a data point, especially for identifying noises with Eq. (11). The larger the value of c is, the smaller the influence field of a data point is, and more data points can be identified as noises. Note that the parameter is very stable, for it only affects elementary clustering result generated by the initial similarity, and the final perceptual result can not be affected because of top-down procedure. In order to generate meaningful elementary clusters, c should satisfy

$$c < \frac{N}{2 \cdot \sum_{i=1}^{N} ||x_i - x_{\langle 1|i\rangle}||}. \tag{16}$$

Parameter α is data-independent, which controls the strength of identifying density difference. In similarity measurements, Eq. (11) and (13), $diff = \frac{|A-B|}{A+B}$ is called difference degree between two constants, A and B. $\alpha > 50$ means that the relationship with $diff > 0.1$ is forced to be cut off. If the identifying strength needs to be loosened, it only needs to decrease α. α is set as 50 in this paper to cut off the relationship with $diff > 0.1$.

There are few efficient methods to select proper k because it is related to the size and distribution of data sets. Small k leads to emergence of more elementary clusters, and vice versa. Fortunately, the value of k is not crucial to get good clustering result, for it only affects the number of elementary clusters, but has little influence on final clustering result. Based on large amounts of experiments, the value is set within $[8, 30]$ for moderate-sized data sets in this paper. For large data sets, k is suggested to be 0.2–1% of the size of data set.

The computing complexity of the new clustering is $O(NlogN)$, and the complexity is not increased because the initial similarity is only locally revised in the top-down procedure.

7. Experiments

In this section, we have applied the new clustering algorithm to some synthetic data sets and real data sets to demonstrate its superiority, in which the synthetic data sets contain some noises and manifold structure. The clustering algorithms used for comparison contain classical hierarchical clustering, MS (Comaniciu and Meer, 2002), GBMS (Carreira-Perpinan, 2004) and others (such as single-linkage, complete-linkage, average-linkage, centroid-linkage, median-linkage and ward-linkage), the algorithms that can deal with manifold cluster and noises (such as NRSC (Li et al., 2007), Spectral-Ng (Ng et al., 2001), STSC (Zelnik and Perona, 2005) and Chameleon (Karypis et al., 1999)), clustering used in image segmentation (such as N-Cut (Shi and Malik, 2000) and MS (Comaniciu and Meer, 2002)).

7.1. Clustering results on synthetic data sets

The synthetic data set shown in Fig. 5 is used to exhibit the heuristic hierarchical clustering procedure. In the figure, the data set contains structural features such as density difference feature, proximity feature and local direction feature. With the increase of t, the generated elementary clusters merged together in sequence according to different structural features.

In Fig. 5 left column, the results are obtained in different level of the heuristic hierarchical clustering. From top to down, with $c = 50, \alpha = 50, k = 10, t$ varies and the results are shown as agglomerative hierarchical clustering results. Following the increase of t, elementary clusters merge together gradually from top to down, according to certain sequence. Density difference is first considered, and elementary clusters with small density difference merge together first. Then, elementary clusters with the same local direction merge together, and finally, elementary clusters with short distance merge together. If $k < 10$, only more elementary clusters will be obtained in the first several feedback processes, but the clustering result shown in top-left subfigure of Fig. 5 will also appear, and the following hierarchical results will not be affected.

For comparison, other algorithms were also experimented on the data set, and the clustering results from single linkage were more relatively satisfied as shown in Fig. 5 right column. Although single linkage is able to produce a hierarchical clustering results from top to down, it cannot reveal the local structures of data set, e.g. in the top-right subfigure of Fig. 5, the results ignore the density difference feature. Note that the single linkage algorithm is sensitive to noises and the results shown are specially processed to deal with noisy data points.

Another advantage of the new algorithm is its robustness to noises. We experimented the two data sets in massive noises,

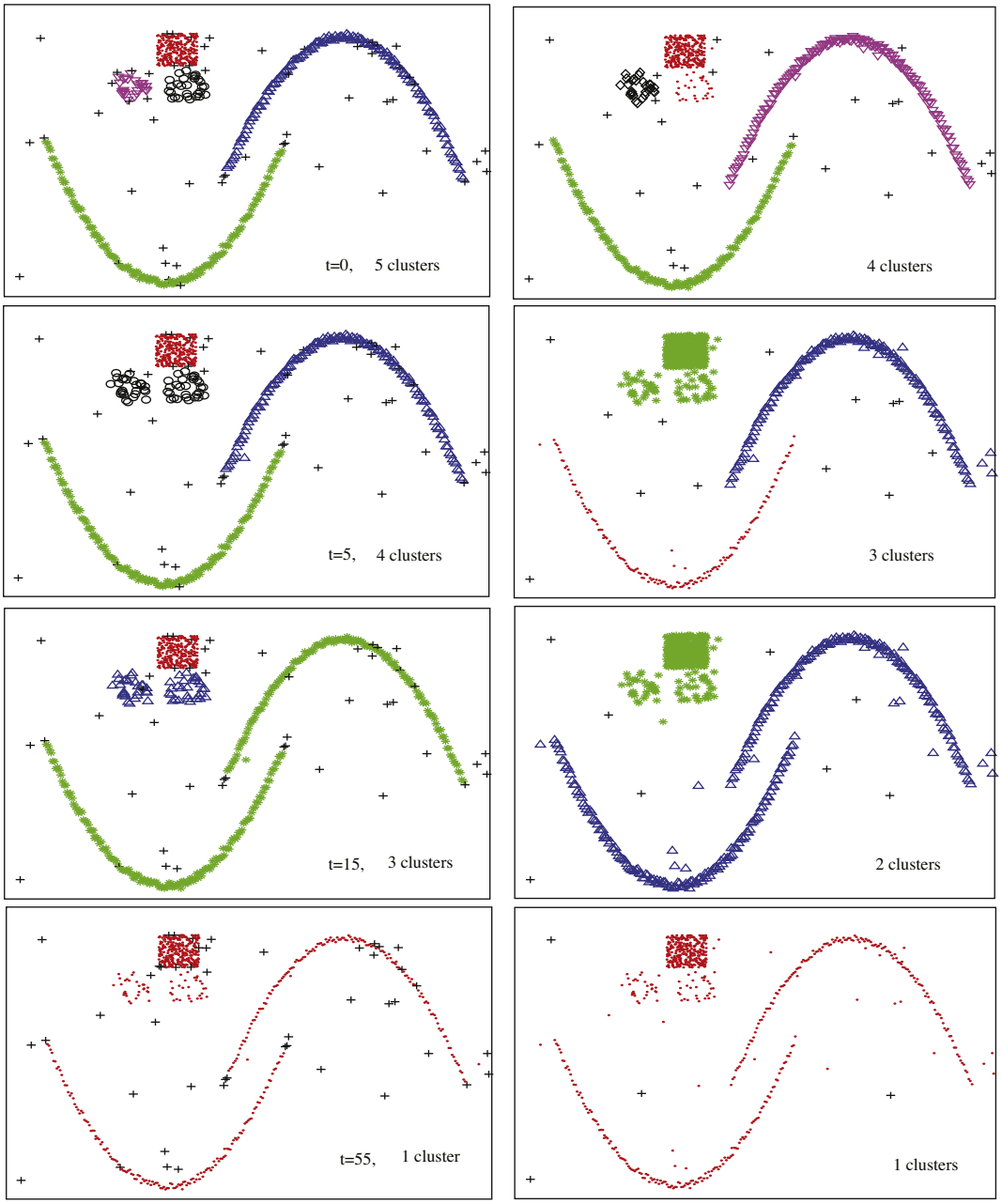

Fig. 5. The comparing results of new algorithm and single-linkage. Left column: clustering results of the new algorithm, $k = 10$, from top to down the value of t is 0, 5, 15 and 55, respectively; right column: the clustering results of single-linkage, from top to down the step number is 25, 15, 10 and 5, respectively.

and the clustering results are shown in Fig. 6. In Fig. 6(a), the massive noises pollute the structure of data set, even the native structure. With the top-down procedure, many elementary clusters appeared while lots of noises were detected, but finally the manifold structure of the data set was revealed. In Fig. 6(b), two Gaussian type clusters were recognized from massive noises, and in large scale, the two Gaussian type clusters could be recognized as one cluster.

7.2. Clustering results on real data sets

Some Benchmark data sets were used to test the new algorithm, in which Iris and Pendigits come from UCI Machine Learning Repository (Blake, 1998), and another data set are samples (two clusters, 0 and 1) from USPS handwriting data set.[2] For data sets Iris and USPS-01, 10% of uniformly distributed noises were added to them to test the algorithm.

The accuracy of clustering is measured with normalized mutual information (Strehl and Ghosh, 2003) (NMI for short), a measurement to compare results of different clustering solutions when the labels are known. This index lies between 0 and 1, and the higher the NMI is, the better the clustering solution is. The algorithms that are compared with on these benchmark data sets involve NRSC (Li et al., 2007), Spectral-Ng (Ng et al., 2001), STSC (Zelnik and Perona, 2005) and Chameleon (Karypis et al., 1999), and the clustering results on the data sets are shown in Fig. 7.

From Fig. 7, the new algorithm outperforms others overall. Among the algorithms used for comparison, NRSC and Chameleon have the advantage to deal with noises. However, the new algorithm performs better than them in dealing with noises, for the new algorithm puts emphasis on description of local structures and feedback mechanism in top-down procedure.

[2] http://www.gaussianprocess.org/gpml/data/.

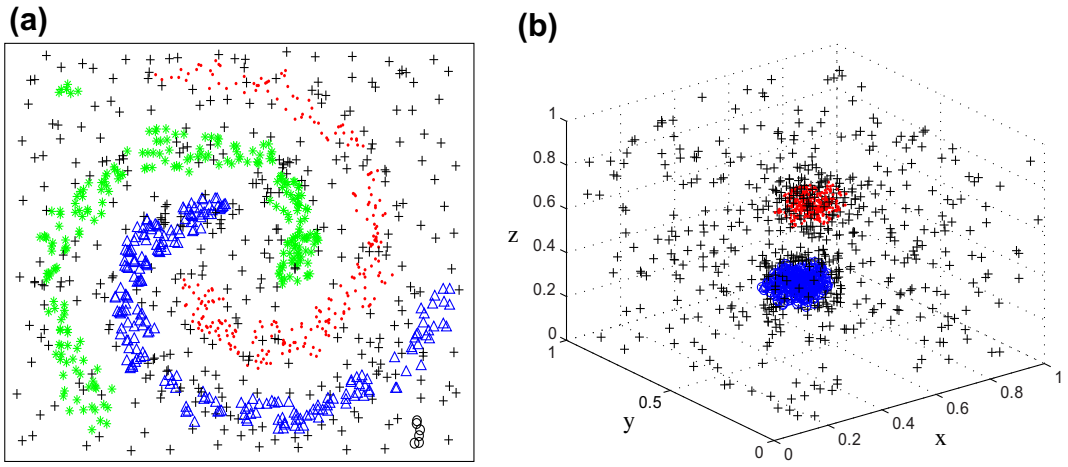

Fig. 6. Clustering results on manifold data set and data set sampled with mixed Gaussian which are polluted by lots of background noises. (a) Clustering result with $t = 3$. Parameters are set with $c = 60,\ \alpha = 50,\ k = 10$. (b) Clustering result with $t = 3$. Parameters are set with $c = 50,\ \alpha = 50,\ k = 10$.

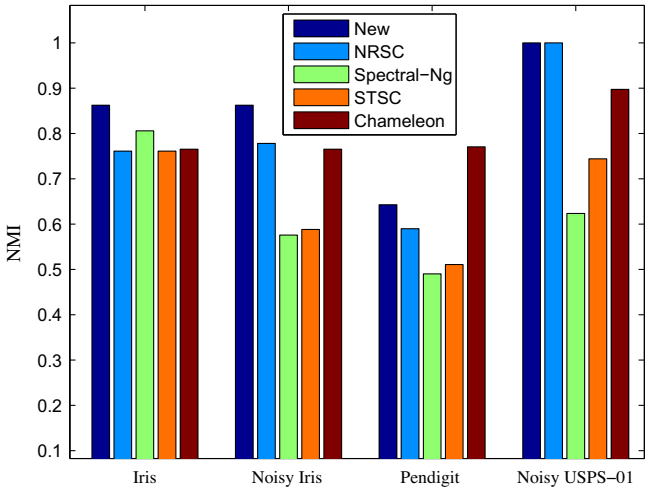

Fig. 7. NMI value of clustering results with different algorithms.

We now present an experiment with the new algorithm in image segmentation and the comparative results with three classical algorithms: K-means, N-Cut (Shi and Malik, 2000) and MS (Comaniciu and Meer, 2002). The image used in the experiment is shown in Fig. 1(a), and it is segmented by clustering the data sets in color space, Fig. 1(b).

Each result shown in Fig. 8 is selected with the best result with different clustering algorithms. The first row of Fig. 8 shows the results generated by the new algorithm: the elementary cluster, results with feedback, and results without feedback, respectively. The second row of Fig. 8 lists the comparative results with the other three algorithms. From the figure, the results with feedback are more organized than other results, both in comparison with new algorithm without feedback, or comparison with other algorithms.

8. Conclusion

Similarity measurement is one of the important factors in clustering algorithm design, but the definition and application of similarity measurement are limited in traditional algorithms.

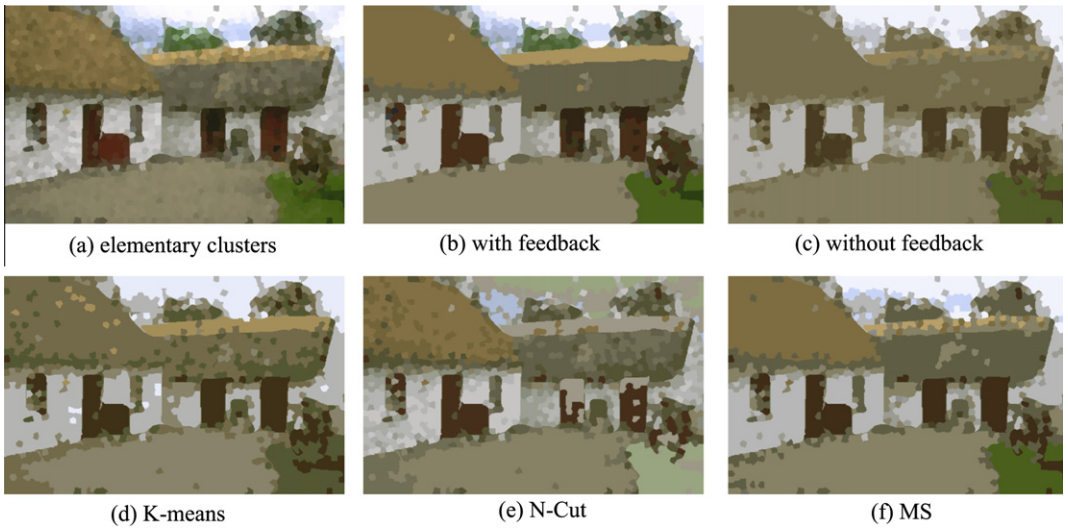

Fig. 8. Comparative results of image segmentation. (a) Elementary clusters generated before feedback; (b) results after feedback, $t = 7,\ k = 30,\ \alpha = 50,\ c = 50$; (c) results without feedback; (d) results with *K*-means, $k = 7$; (e) results with N-Cut, $k = 15$; (f) results with MS, $\sigma = 0.04$.

Generally, by merely considering the distance information between a pair of points, similarity measurement is only defined as a distance function, which inevitably will lose the orientation information between a pair of points. Furthermore, there is only one fixed similarity measurement for one clustering algorithm, and the similarity measurement can not be changed in the clustering process, which deprives the clustering algorithm of the possibility of further cluster identification in the clustering process. To further rationalize the definition and application of similarity measurement as well as the clustering process, the paper constructs a flexible similarity measurement, which can be specialized through information feedback in clustering process. Compared with traditional agglomerative clustering algorithms, the new algorithm emphasizes the collection of the dynamic information of clusters so as to adjust the similarity between the pairs of points automatically, which can best correspond to the structural information of the data set.

Combined with the multiple similarity, a new clustering framework is put forward, which is simple but powerful. The advantages of the new algorithm lie in two main points: the high computation speed and strong expansion ability. With strong expansion ability, many new similarity measurements can be added or learnt with it. Compared with traditional algorithms, the new one possesses obvious superiority, which is shown in the experiments.

References

Abolhassani, H., Mahdavi, M., 2009. Harmony *k*-means algorithm for document clustering. Data Min. Knowl. Disc. 18, 370–391.

Bishop, C.M., 1999. Pattern Recognition and Machine Learning, first ed. Addison-Wesley, Harlow, England.

Blake, Merz C.J., 1998. UCI repository of machine learning databases.

Cheng, Y., 1995. Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Machine Intell. 17 (8), 790–799.

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach towards feature space analysis. IEEE Trans. Pattern Anal. Machine Intell. 24 (5), 603–619.

Carreira-Perpinan, M.A., 2004. Fast nonparametric clustering with gaussian blurring mean-shift. In: Proc. 23rd Internat. Conf. Machine Learning, pp. 153–160.

Fukunaga, K., Hostetler, L.D., 1975. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans. Inform. Theory 21 (1), 32–40.

Karypis, G., Han, E.H.S., Kumar, V., 1999. Chameleon: hierarchical clustering using dynamic modeling. Computer 32 (8), 68–75.

Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L.E., Jordan, M.I., 2004. Learning the kernel matrix with semidefinite programming. J. Machine Learn. Res. 5, 22–72.

Leung, Y., Zhang, J.S., Xu, Z.B., 2000. Clustering by scale space filtering. IEEE Trans. Pattern Anal. Machine Intell. 22 (12), 1396–1410.

Li, C.Z., Xu, Z.B., 2011. Structure identification-based clustering according to density consistency. Math. Probl. Eng., 14 pages (Article ID. 890901).

Li, C.Z., Xu, Z.B., Yuan, Y.B., 2011. Dissimilarity based on direction information and its application. In: Proc. 4th Internat. Conf. Image and Signal Processing, vol. 1, pp. 139–143.

Li, Z.G., Liu, J.Z., Chen, S.F., Tang, X., 2007. Noise robust spectral clustering. In: Proc. 11th Internat. Conf. Computer Vision, pp. 1–8.

Mukherjee, J., 2002. MRF clustering for segmentation of color images. Pattern Recognition Lett. 23 (8), 917–929.

Ng, A.Y., Jordan, M.I., Weiss, Y. 2001. On spectral clustering analysis and algorithm. In: Advances in Neural Information Processing Systems, pp. 849–856.

Pichel, J.C., Singh, D.E., Rivera, F.F., 2006. Image segmentation based on merging of sub-optimal segmentations. Pattern Recognition Lett. 27 (10), 1105–1116.

Schenders, P., 1997. A comparison of clustering algorithms applied to color image quantization. Pattern Recognition Lett. 18 (11–13), 1379–1384.

Sonnengurg, S., Ratsch, G., Schafer, C., 2006. A general and efficient multiple kernel learning algorithm. Adv. Neural Inform. Process. Syst. 18, 1273–1280.

Shi, J.B., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Machine Intell. 22 (8), 888–905.

Strehl, A., Ghosh, J., 2003. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. J. Machine Learn. Res. 3 (3), 583–617.

Santos, J.M., Marqures, J. 2005. Human clustering on bi-dimensional data: An assessment. Technical Report 1, INEB-Instituto de Engenharia Biomdica.

Yousri, N.A., Kamel, M.S., 2009. A distance-relatedness dynamic model for clustering high dimensional data of arbitrary shapes and densities. Pattern Recognition 42 (7), 1193–1209.

Zelnik, M.L., Perona, P., 2005. Self-tuning spectral clustering. In: Advances in Neural Information Processing Systems, pp. 1601–1608.