2014 13th International Conference on Control, Automation, Robotics & Vision
Marina Bay Sands, Singapore, 10-12th December 2014 (ICARCV 2014)

Analysis on Local Optimum Existence Form of K-means-type

Zhang Chengning, Xia Qinhua, Zhao Fei
Ningbo Institute of Materials Technology&Engineering,
Chinese Academy of Sciences
Ningbo, China
{zhangchengning, xiaqinghua}@nimte.ac.cn

Zou Yuanyuan
Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China
Yangzhou Engineering Research Center, Shenyang Institute of Automation, Chinese Academy of Sciences, Yangzhou, China
yyzou@sia.cn

***Abstract*—With the hypothesis of Gaussian distribution of patterns, K-means and its extensions are good for clustering. As the representative of partitional clustering algorithm, K-means follows rules for running: numbers of clusters to be set, cluster initialization to be specified and certain objective function to be optimized. In general, FCM, ANN, EM share the identical idea with K-means in the beginning of running, and local optimum is the basic perspective of these K-means-type clustering methods. How numbers of clusters and cluster initialization affect local optimum existence is the query of this paper, the analysis will be given. In this paper, K-means-type algorithms are summarized, convergence proof will be shown, local optimum existence form is analyzed in detail and the classical probability expression of the existence is presented.**

***Keywords* – clustering, K-means, local optimum.**

I. INTRODUCTION

Clustering is the classic topic in pattern recognition[1–2]. Different from classifying, clustering is the technology of pattering labeling, rather than querying a decision boundary with labeled patterns[3]. Clustering is divided into two sub-branches according to methods: hierarchical based clustering[4] and partitional based clustering[5]. The former gives nested clusters in agglomerative mode or in divisive mode, single-link[6] and complete-link[7] are representatives of hierarchical clustering. The latter uses another methodology for clustering, it initializes numbers of clusters, specifies initial locations of clusters, and recursively check one objective function until that function converges. Basically, objective function utilizes some criterion, such as the sum of squared error[3]. Considering the partitional clustering procedure, one should notice that numbers of clusters and clusters initial locations which are pre-set parameters are empirical issues while facing an unknown pattern space. How to specify proper parameters brings K-means extensions.

The K-means[8] extension is developed in different ways. Some gives more exquisite mechanisms for numbers of clusters decision, and others are with additional heuristics methods. ISODATA[9] and FORGY[10] are early algorithms of K-means extensions. ISODATA employed cluster splitting and cluster merging while running, and that makes the algorithm good for several complicated clusters. Accompany with the development cost, more parameters are introduced, and that even more difficult for parameter initialization. FCM[11] is focusing the soft assignment to improve the clustering result. In K-means, each pattern is assigned to one cluster, and that could be simple in decision and the result could be also poor. FCM adopts a considered rule of clustering, and this method assigns each pattern to be a member of all clusters with different membership value. X-means[12] finds numbers of clusters by optimizing another criterion called Akaike Information Criterion and Bayesian Information Criterion, which are AIC and BIC rules respectively. EM[13] is a mixed model way to be used in clustering. This method regards patterns as Gaussian distribution, and with the hypothesis of multiple clusters existence, parameters are initialized followed the same way of K-means, then recursively converges. ANN[14] also needs topology of prototypes to be set. The winning neuron is confirmed where a neuron is the closest to one pattern. The neurons are updated when patterns are assigned in a feeding round, and that will be stopped as neurons are stabled. One could note that the ANN based clustering is followed the same mechanism as K-means.

Through the analysis of K-means extensions, it is difficult to decide the initial parameters for running algorithm. It is noticed that these parameters are crucial to K-means-type algorithms. Different numbers of clusters and cluster initializations lead to variant clustering results, and the query of this paper is to discuss the clustering result with different configurations of initializations. We firstly summaries three sub-branches of K-means-type to show the similarity of these algorithms. Then we give local convergence description of K-means-type both in aspects of numbers of clusters and cluster initialization. Finally, we analyze local optimum existence in detail and give a classical probability expression of existence.

International Science And Technology Cooperation Project Of Ningbo (Granted No.2013D10010)
*The Natural Science Foundation Of Jiangsu Province（Granted No.BK2011453）

II. ANALYSIS ON LOCAL OPTIMUM EXISTENCE

In this part, we spread the topic in three steps: summarization of similarity K-means-type representatives, description of local convergence and analysis of local optimum existence form.

A. Similarity of K-means-type representatives

The numerous extensions of K-means are difficult to be listed in this paper, but according to general procedure of K-means family, some regular steps are essential to make sure the algorithms running. For example, the numbers of clusters should be set, and the initial positions of clusters should be specified. As the initializations are commonly among the K-means-type algorithms, the consequent caused by the initializations should be studied. In this section, K-means-type algorithms are summarized with several representatives, such as fuzzy c-means, artificial neural network and expectation-maximization.

1) K-means

Squared error plays the role of an iterative criterion in partitional clustering. Methods were proposed in 1960s[8], the main idea depends on initialization and iteration. K-means algorithm is the representative of partitional clustering. Initialized points are regarded as the centers in a data set, then the algorithm iterates until the convergence is achieved. In formal description, there is a data set $X = \{x_1, x_2, .., x_N\}$, the initialized centers are represented as $C = \{C_1, C_2, ..., C_m\}$. Each pattern in X is assigned into centers with the rule of minimum distance, then updated centers are calculated by averaging the elements in each cluster. The updated centers are employed into an objective function which achieves asymptotical stability. The steps of K-means algorithm are as follows:

(1). Initialize the centers of clusters: confirm the initial number of clusters and the positions of the centers. Give the rule of convergence;

(2). Group patterns based on minimum distance rule;

(3). Update the centers of the clusters;

(4). Return (2) until the convergence is achieved.

2) FCM

Patterns are assigned into clusters with the rule of minimum distance in K-means algorithm, and that is called rigid decision rule of merging where a pattern is assigned into one cluster only. As the matter of fact, one pattern should not be that rigid to be assigned to one cluster. Following the fuzzy way of assigning a pattern, FCM[11] is the crucial variation of K-means algorithm. The membership is defined in the algorithm which describes the probability of evaluating one pattern to all existing clusters. Rather than 0 or 1, the value in the interval of 0 and 1 is given as the membership. Each pattern will not be assigned to a certain cluster in a round of iteration. Instead of that, a pattern will be given a membership value to each cluster and the membership keeps updating in the iteration. The ultimate assignment of each pattern will be given until the iteration is done. FCM algorithm is introduced formally in steps as follows:

(1). Select the fuzzification parameter m, number of clusters c, an arbitrary small positive threshold number ε. Define the initial matrix of prototype $M = [m_1, m_2, ..., m_c]$.

(2). Update the membership matrix U, and elements of U are updated as follows:

$$u_{ij}^{t+1} = 1 \Big/ \left(\sum_{l=1}^{c} (D_{lj} / D_{ij})^{1/1-m}\right),\ i = 1,2,...,c \ and\ j = 1,2,...,N$$

where, $D_{ij} = D(x_i, x_j)$, $m \in [1, \infty)$

(3). Update the prototype matrix:

$$m_i^{t+1} = \left(\sum_{j=1}^{N} u_{ij}^{t+1} x_j\right) \Big/ \left(\sum_{j=1}^{N} (u_{ij}^{t+1})^m\right),\ i = 1,2,...,c$$

(4). Repeat (2) and (3) until the convergence condition is achieved:

$$\left\| M^{t+1} - M^t \right\| < \varepsilon$$

3) ANN

In the point view of biology, clusters are seen as several neurons scatted in the space. Each cluster is constructed with a stimulated neuron and its surrounding neurons[15]. Neurons here are initially arranged with a specified topology, and patterns are then fed to the input layer of a neural network. A pattern will stimulate a neuron in the heap of neurons, and simulation is based on the distance between one pattern and one neuron. The winning neuron builds up a neighborhood with a specified radius, and the neurons within the neighborhood will also be stimulated with different level according to a defined function. The weights of stimulated neurons are updated with a learning expression. The learning mechanism updates weights of stimulated neurons by employing a learning rate parameter, which is a percentage. With different configuration of learning rate, the percentage changes its value from zero to one, zero represents null learning status and one represents complete learning, and other decimal fraction represents common learning. The larger the learning rate is, the more effects the input pattern will give to the stimulated neuron. Unstimulated neurons will have no chance to update their weights, or even be inhibited. That is to say, a null update or a negative update is added to the updating mechanism. This mechanism guarantees a new pattern will be correctly assigned into corresponding clusters. SOM [14](Self Organizing Map) or SOFM (Self Organizing Feature Map) shares the same philosophy to produce low dimension from high dimension. Though SO(F)M is not designed for clustering primitively, some algorithms are proposed using its idea. Steps are given as follows:

(1). Initialize the topology of prototypes $v_i\ i = 1,2,...,k$

(2). Feed a pattern x_j to the network for selecting the winning neuron which is closest to the pattern using the rule:

$$v_{win} = \min d(x_j, v_i)\ \ i = 1,2,...,k$$

(3). Update neurons:

$$v_i(t+1) = v_i(t) + \eta_i (x_j - v_i(t)) \quad i = 1, 2, \ldots, k$$

where, η_i is the learning rate which can either be specified according to some defined neighborhood or can be given as a exponential form:

$$\eta_i = \alpha(t)\exp(-d^2(x_j, v_i)/2\sigma^2(t))$$

where, $\alpha(t)$ and $\sigma(t)$ are monotonically decreasing learning rate and kernel width function respectively.

Repeat (2) and (3) until the neurons are keeping stable.

4) EM

Expectation Maximization[13] is developed with an elaborate consideration of multi-clusters distributed in the space. EM tries to solve the determination of parameter sets of entire clusters with the difficulty in which the distribution of parameters are unknown. The algorithm adopts the expectation of the parameters to evaluate the optimal distribution since the difficulty mentioned above.

Here, let $Q(\boldsymbol{\theta}' | \boldsymbol{\theta}) = E(\log f(\boldsymbol{x} | \boldsymbol{\theta}) | \boldsymbol{y}, \boldsymbol{\theta})$ be the expectation of patterns distribution, where, $\boldsymbol{\theta}$ is the current parameter set corresponding to respective clusters, $\boldsymbol{y}$ is the observed data in the space called incomplete data and $\boldsymbol{x}$ is the unobserved data in the space called complete data which can be observed through $\boldsymbol{y}$. $f(\boldsymbol{x} | \boldsymbol{\theta})$ represents the probability density of the complete data under the condition of current parameter $\boldsymbol{\theta}$. Steps are as follows in the iteration of EM:

(1). Initialize $\boldsymbol{\theta}^0$ to the data set.

(2). E-step: Compute the expectation of the complete data with logarithm form:

$$Q(\boldsymbol{\theta}' | \boldsymbol{\theta}) = E(\log f(\boldsymbol{x} | \boldsymbol{\theta}) | \boldsymbol{y}, \boldsymbol{\theta})$$

where, $f(\boldsymbol{x} | \boldsymbol{\theta}) = \log p(\boldsymbol{x} | \boldsymbol{\theta}) P(C_i)$

(3). M-step: Maximize Q to select the update parameter estimation:

$$\boldsymbol{\theta}^{t+1} = \max Q(\boldsymbol{\theta}^t, \boldsymbol{\theta})$$

(4). Repeat (2) and (3) until the convergence is achieved.

B. Local Convergence

In section A, K-means-type algorithms are summarized, and the similarity is also demonstrated. In order to study local optimum existence form, local convergence of K-means-type should be investigated. The convergence proof will be given at first, and then the local convergence description will be shown in virtue of numbers of clusters and clusters initial positions.

1) Convergence Proof

Let the criterion function of K-means-type be:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{N} d_{ij}(\boldsymbol{x}_j, \boldsymbol{c}_i) \quad (1)$$

After $t+1$ round of iterations, the set of patterns which have changed their labels can be described as $M = \{z_1^M, z_2^M, \ldots, z_m^M\}$. For the element z_k^M, suppose that z_k^M is from class S_i^t to class S_j^t due to clustering, and class S_i^{t+1} and class S_i^{t+1} is formed. Let the mean value of S_i^{t+1} and S_i^{t+1} be μ_i^{t+1} and μ_i^{t+1} respectively. Let the mean value of S_i^t and S_i^t be μ_i^t and μ_i^t respectively. Here, the relation of μ_i^{t+1} and μ_i^t can be expressed:

$$u_i^{t+1} = u_i^t - \frac{1}{n_i - 1}(z_k^M - u_i^t) \quad (2)$$

Also, we can get:

$$u_j^{t+1} = u_j^t + \frac{1}{n_j + 1}(z_k^M - u_j^t) \quad (3)$$

Thus, the objective function can be formulated:

$$\begin{aligned} J_i^t &= \sum_{x_l \in S_i^t} | x_l - u_i^t |^2 = \sum_{x_l \in S_i^{t+1}} | x_l - u_i^t |^2 + | z_k^M - u_i^t |^2 \\ &\geq \sum_{x_l \in S_i^{t+1}} | x_l - u_i^{t+1} |^2 + \sum_{x_l \in S_i^{t+1}} | \frac{z_k^M - u_i^t}{n_i - 1} |^2 + | z_k^M - u_i^t |^2 \quad (4) \\ &\geq J_i^{t+1} + \frac{n_i}{n_i - 1} | z_k^M - u_i^t |^2 \end{aligned}$$

And the result is:

$$J_i^t \geq J_i^{t+1} + \frac{n_i}{n_i - 1} | z_k^M - u_i^t |^2 \quad (5)$$

The same conclusion is :

$$J_j^t \geq J_j^{t+1} - \frac{n_j}{n_j + 1} | z_k^M - u_j^t |^2 \quad (6)$$

We have:

$$\begin{aligned} J_i^t + J_j^t &\geq J_i^{t+1} + J_j^{t+1} + \frac{n_i}{n_i - 1} | z_k^M - u_i^t |^2 - \frac{n_j}{n_j + 1} | z_k^M - u_j^t |^2 \quad (7) \\ &\doteq J_i^{t+1} + J_j^{t+1} + | z_k^M - u_i^t |^2 - | z_k^M - u_j^t |^2 \end{aligned}$$

For $z_k^M \in S_j^{t+1}$, the formula below can be got:

$$| z_k^M - u_i^t |^2 > | z_k^M - u_j^t |^2$$

Thus, we can have:

$$J_i^t + J_j^t \geq J_i^{t+1} + J_j^{t+1} + \Delta \geq J_i^{t+1} + J_j^{t+1} \quad (8)$$

Where: $\Delta = | z_k^M - u_i^t |^2 - | z_k^M - u_j^t |^2 \geq 0$

And the following result is obtained:

$$J^t = \sum_{i=1}^{C} J_i^t \geq \sum_{i=1}^{C} J_i^{t+1} = J^{t+1} \quad (9)$$

As $\{J^t\}$ is not empty and is bounded below $(J^t > 0)$, according to the greatest lower bound axiom:

$$L = \inf_t \{J^t\} < \infty \quad (10)$$

For $\forall \varepsilon > 0, \exists J^T$, such that:

$$J^T - \varepsilon < L \quad (11)$$

Otherwise, $L + \varepsilon$ is a bounded below of $\{J^t\}$, and contradiction occurs.

As $\{J^t\}$ is a decreasing sequence, with $\forall t > T$, we have:

$$\left| J^t - L \right| = J^t - L < J^T - L < \varepsilon \quad (12)$$

Thus:

$$\lim_{t \to \infty} J^t = L \quad (13)$$

Finally, K-means is convergent.

2) Local Convergence Description

K-means type is a convergent algorithm, and the convergence is local optimum not global. This conclusion can be described in two aspects: in the sense of numbers of clusters and in the sense of cluster initialization.

a) In the sense of numbers of clusters

In K-means type algorithms, numbers of clusters are difficult to be determined because of lack of prior knowledge, and the initialization of numbers of clusters is arbitrary. Supposing patterns are distributed in Gaussian, we will increase the numbers of clusters to demonstrate that the convergence can be achieved either. For ease of presentation, here we study the case that one class is divided into two sub-classes and the change of objective function in that case.

Supposing class j is divided into class k and class l, we have:

$$n_j m_j = n_k m_k + n_l m_l \tag{14}$$

Where, n_j is the number of patterns within class j, m_j is the mean value of class j, other symbols share the same meaning.

Thus:

$$n_k(m_j - m_k) + n_l(m_j - m_l) = 0 \tag{15}$$

Then: $(m_j - m_k)$ and $(m_j - m_l)$ are contrary signs.

Let $m_k < m_j < m_l$, here we have:

$$J_C^* = \sum_{n_k} \|x_i - m_j\|^2 + \sum_{n_l} \|x_i - m_j\|^2 \tag{16}$$

$$J_{\bar{C}}^* = \sum_{n_k} \|x_i - m_k\|^2 + \sum_{n_l} \|x_i - m_l\|^2 \tag{17}$$

Where, C and $\bar{C}$ are symbols of patterns before and after the division, J_C^* and $J_{\bar{C}}^*$ are corresponding cost function of C and $\bar{C}$.

Obviously, there is the result:

$$J_C^* > J_{\bar{C}}^* \tag{18}$$

Here, we can conclude that the more numbers of clusters are initialized, the less the cost is. Moreover, with the limit sight, if there are as many numbers of clusters as the numbers of patterns, the cost will be zero. That is to say, the arbitrary initialization of numbers of clusters is important to the cost. Without confirmed prior knowledge, the configuration of initialization is always disputable.

b) In the sense of cluster initialization

In this part, local convergence will be demonstrated by different configurations of cluster initialization. K-means type is starting with certain initialization of clusters in heuristic way, and the iteration is driven by data and that initial positions. Thus, the solution is local convergent. Here, we give a theorem to describe.

Theorem: In global optimization, there is a finite sequence $X = \{x_1, x_2, ..., x_N\}$, and the global optimal solution x_0. The neighborhood is: $B(x_0) = \{x \in \Omega \,\|\, x - x_0 | < \sigma, \sigma > 0\}$, and let $f: \Omega \in R^n \to R$, $f^* = \arg\min_{x \in \Omega}\{f(x)\} = f(x_0)$. The event is defined as: $E_i = | f(x_i) - f^* | < \varepsilon$, then we have:

$$P(\bigcup_{i=1}^{N} E_i) = c \in (0,1)$$

According to the theorem, it is noticed that the possibility value of global optimal solution is *c*, and that is not convergence in probability 1. Here, the element x_i of the finite sequence represents the final clustering centers of K-means type running. This theorem demonstrates that, with finite solutions of clustering, the global optimal solution is not obtained in possibility value 1. The following description is shown:

Errors which are produced by randomly selection of initialization follows the continuous uniform distribution model, the definition is:

$$p(E_i) = \begin{cases} \dfrac{1}{\beta - \alpha}, & E_i \in [\alpha, \beta] \\ 0, & else \end{cases} \tag{19}$$

Thus we have:

$$P(\bigcup_{i=1}^{N} E_i) = \sum_{i=1}^{N} \int_{\alpha}^{\varepsilon} p(E_i)dx = \frac{N(\varepsilon - \alpha)}{(\beta - \alpha)} = c \in (0,1) \tag{20}$$

Where, α and β are minimal error and maximal error respectively. It is noted that ε is crucial to the possibility expression.

C. Local Optimum Existence Form Analysis

The convergence proof and the local convergence description are presented in section A and section B. In this section, a detailed analysis of local optimum existence form will be given, and the m-class with n-center problem is the goal of this study. However, basic 2-class with 2 centers problem should be investigated firstly, for the ease of the analysis.

1) 2-class with 2-center problem

The investigation of local optimum existence form of K-means type in 2-class with 2-center case is find the possible existence forms. Due to this task, K-means type running mechanism should be reviewed, and then the impossible form also should be studied.

a) K-means-type mechanism

The running mechanism of K-means-type algorithm is important to understand the local optimum existence form. For each pattern in the data set, the algorithm assigns it to the closest center. Essentially, one pattern is assigned based on a perpendicular bisector of the pairwise centers. It is illustrated in figure 1.

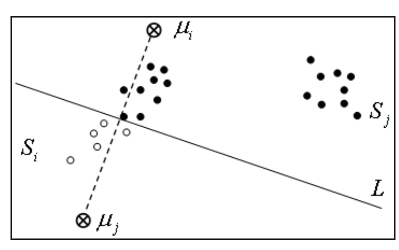
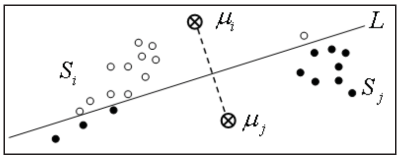
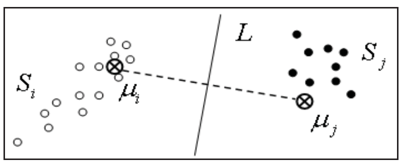


Figure 1. K-means-type running mechanism

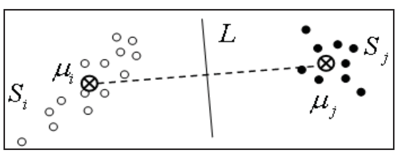
It is noticed that cluster centers are updated by patterns due to data driven style of K-means-types. However, no matter how the centers update, the decision bound is invariably the perpendicular bisector. During the iteration, if a high density region exists, a center is moving fast until the center is completely in that region. Meanwhile, as one center is moving close to one region, other regions are apart from the center, and some patterns will probably re-assign to other class center. This procedure is shown in figure 2.



(a) Initial phase



(b) Iterative phase



(c) Final phase

Figure 2. K-means type running procedure

b) local optimum existence form

In case of 2-class and 2-center, supposing patterns are followed Gaussian distribution, considering the mechanism of K-means-type algorithms, the local optimum is always that two classes are well clustered with the initial positions of centers. Meanwhile, in this part, local optimum forms are investigated in order to find other possible existence. That is to say, for Gaussian distributed patterns, forms of the local optimum that the decision line L goes through the two classes will be studied. For ease of analysis, we assume that homogenous ellipses are employed instead of Gaussian distributed patterns. Here, four cases of homogenous ellipses are analyzed under K-means-types mechanism, and the Gaussian distribution analysis is the fifth case.

Case 1: homogenous circles. The stable local optimum can be achieved in figure 3. If the perpendicular bisector formed by two centers is parallel to the line of two class: class S_i and class S_j.

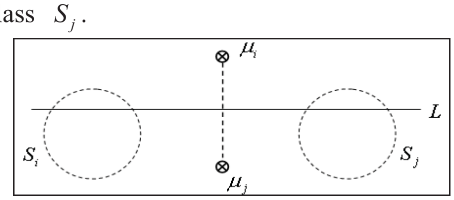


Figure 3. Local optimum for circular sheet of 2-class problem

Case 2: homogenous ellipses with identical variance. The stable local optimum can be achieved in figure 4. If the perpendicular bisector formed by two centers is parallel to the line of two class: class S_i and class S_j.

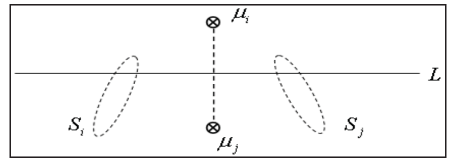


Figure 4. Local optimum for circular sheet of 2-class problem

Case 3: homogenous ellipses with different variance. The stable local optimum can be achieved in figure 5. If the perpendicular bisector formed by two centers is parallel to the line of two class: class S_i and class S_j.

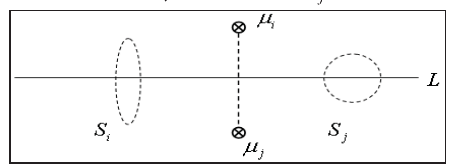


Figure 5. Local optimum for non-homoscedastic ellipse sheet of 2-class problem

Case 4: homogenous non-axis symmetric ellipse with different variance. The local optimum is not stable shown in figure 6.

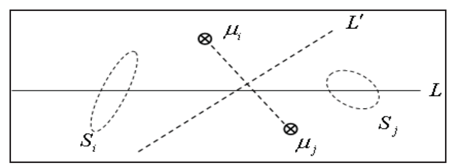


Figure 6. Local optimum for non-homoscedastic ellipse sheet of 2-class problem

Proof of Case 4: As illustrated in figure 6, two homogenous ellipses are in the feature space, the variances of them are different. The centroids of them are connected with line L. Here, we build an ellipse S_i' as the axial symmetry ellipse of S_i, and build an ellipse S_j' as the axial symmetry ellipse of S_j. The built ellipses are shown in solid line. Thus, the intersection regions between the ellipse and its symmetry one can be find.

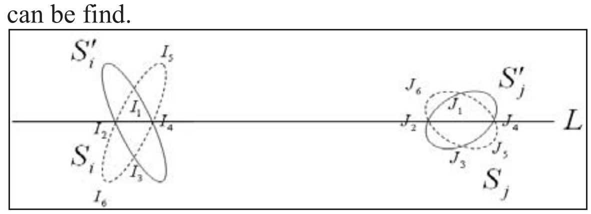


Figure 7. Symmetrical ellipse sheets forming

As shown in figure 7, region $I_1I_2I_3I_4$ and region $J_1J_2J_3J_4$ are symmetry regions with respect to line L. Thus, there exists the stable local optimum where the decision line L goes through the two classes. However, region

$I_1I_4I_5 \cup I_2I_3I_6$ and $J_3J_4J_5 \cup J_1J_2J_6$ are not axis symmetry region, which leads to the case that centers are bias to classes respectively.

Case 5: Gaussian distribution with different variance. This case is the original query in this discuss. Due to the random positions of samples, there exists an extremely low probability event of stable local optimum where the decision line L goes through two classes.

According to the analysis of five cases, we conclude that other possible local optimum forms exist in some possibility value, with elaborate hypothesis of axis symmetry homogenous ellipses. As for general Gaussian distributed patterns, the possibility of that local optimum is extremely low. Thus, in 2-class and 2-center case, the local optimum form that one decision line goes through two classes is almost in non-existent. Rather, two clusters are completely gathered with two centers.

2) m-class with n-center problem

The m-class with n-center problem is the extension of 2-class with 2-center where the local optimum existence form is analyzed using the conclusion of 2-class with 2-center case. That is to say, the 2-class with 2-center problem gives the result that the local optimum existence form is surely the completely gathered clusters rather than the case that there exists the decision line goes through the classes. It means that classes, under the mechanism of the K-means type algorithms, are not cut into sub-classes when the clustering is done. Based on this conclusion, the m-class and n-center problem may transform to 2-class and 2-center, and then all possible events of the local optimum existence forms can be analyzed in classical possibility model.

a) local optimum existence form

As for the m-class with n-center problem, the local optimum existence form research is based on the conclusion of 2-class with 2-center. Individual cluster keeps the completeness after the running of clustering algorithm. Though the initialization of numbers of clusters is arbitrary, the cluster completeness conclusion is guaranteed. To some extent, m-class with n-center problem can transform to 2-class with 2-center, and equally, 2-class with 2-center problem can transform to m-class with n-center. This will be shown in figure 8.

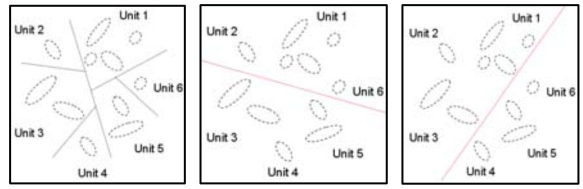


Figure 8. 2-class transformation for m-class

b) classical probability expression

Due to the cluster completeness conclusion and the arbitrary initialization of numbers of clusters, m-class with n-center problem can lead to various results that examples are shown in figure 9.

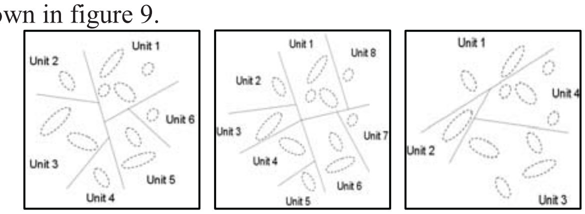


Figure 9. Several Units after clustering

Thus, different numbers of centers and different numbers of clusters can be investigated in the sense of mathematical permutations and combinations, and that is shown in table 1. In this table, the unit represents the groups that clusters are combined, that is based on the completeness rule.

In the case of m-cluster with n-center problem, supposing there exists k units. We have the conclusion:

$$\begin{cases} \sum_{i=1}^{k} T_i^{(k)} = m \\ \sum_{i=1}^{k} S_i^{(k)} = n \end{cases}, k \in [2, m-1] \tag{21}$$

Where, $T_i^{(k)}$ is the numbers of sub-classes within one unit. $S_i^{(k)}$ is the assigned numbers of centers within that unit. With k units, let the numbers of $\{T_i^{(k)}\}$ be $NT^{(k)}$ and let the assigned numbers of $\{S_i^{(k)}\}$ be $NS^{(k)}$. Thus, in $NT^{(k)}$ combinations of space separation, according to Bernoulli experiment, the possibility expression of the event of one center existence in each unit:

$$p = \sum_{k=2}^{m-1} \prod_{i=1}^{k} [p(T_i^k)]^{S_i^{(k)}} \tag{22}$$

Table 1 Configurations for m-class problem

Unit number \ Cluster Number	m					
2 Unit	C_m^1	C_m^2	C_m^3	C_m^4	$\dots$	$C_m^{m/2}$ or $C_m^{(m-1)/2}$
3 Unit	$C_m^1 C_{m-1}^1$	$C_m^{n_1} C_{m-n_1}^{n_2}$, where $n_1, n_2 \in Z$, and $m > n_1 + n_2$				
4 Unit	$C_m^1 C_{m-1}^1 C_{m-2}^1$	$C_m^{n_1} C_{m-n_1}^{n_2} C_{m-n_1-n_2}^{n_3}$, where $n_1, n_2, n_3 \in Z$ and $m > n_1 + n_2 + n_3$				
……						
m-1 Unit	$C_m^1 C_{m-1}^1 \dots C_2^1$	$C_m^{n_1} C_{m-n_1}^{n_2} \dots C_{m-(n_1+n_2+\dots+n_{m-2})}^{n_{m-2}}$, where $n_1, n_2, \dots, n_{m-2} \in Z$, and $m > \sum_{i=1}^{m-2} n_i$				

III. CONCLUSION

In this paper, we give the analysis of local optimum existence form of K-means type algorithms. The summarization of K-means-types is presented, and the problem caused by initialization of numbers of clusters is discussed in detail. The 2-class with 2-center local optimum existence form is analyzed to give a completeness rule. Based on this conclusion, m-cluster and n-center problem is consequently analyzed with possibility sense, and the classical possibility expression is given.

ACKNOWLEDGMENT

The author would like to thank Professor Xia Qinghua and Zou Yuanyuan for providing ideas of formulation and analysis.

REFERENCE

[1] G. J. Gan, et al., "Data Clustering: Theory, Algorithms, and Applications," 2007.

[2] B. S. Everitt, et al., Cluster Analysis, 2010.

[3] R. Xu, et al., Clustering. Hoboken, NJ: Wiley IEEE Press, 2009.

[4] M. G. H. Omran, et al., "An overview of clustering methods," Intelligent Data Analysis, vol. 11, pp. 583-605, 2007.

[5] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, pp. 651-666, 2010.

[6] P. H. A. Sneath, Numerical Taxonomy: W H Freeman & Co (Sd), 1973.

[7] B. King, "Step-wise clustering procedures," Journal of the American Statistical Association, vol. 62, pp. 86-101, 1967.

[8] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the 5th Berkeley Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[9] G. H. Ball and D. J. Hall, "ISODATA, a novel method of data analysis and pattern classification," Stanford Research Institute, Stanford,CA1965.

[10] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency vs interpretability of classifications," Biometrics, vol. 21, pp. 768-769, 1965.

[11] James C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981.

[12] D. Pelleg and A. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," in Proceedings of the 17th International Conference on Machine Learning, Stanford University, 2000, pp. 727 - 734.

[13] A. P. Dempster, et al., "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, pp. 1-38, 1977.

[14] T. Kohonen, "The self-organizing map," in Proceedings of the IEEE, 1990, pp. 1464-1480.

[15] S. L. Shieh and I. E. Liao, "A New Clustering Validity Index for Cluster Analysis Based on a Two-Level SOM," Ieice Transactions on Information and Systems, vol. E92D, pp. 1668-1674, Sep 2009.