

Chapter 2

The Uniform Effect of K-means Clustering

2.1 Introduction

This chapter studies the *uniform effect* of K-means clustering. As a well-known and widely used partitional clustering method, K-means has attracted great research interests for a very long time. Researchers have identified some data characteristics that may strongly impact the performance of K-means clustering, including the types and scales of data and attributes, the sparseness of data, and the noise and outliers in data [23]. However, further investigation is needed to unveil how data distributions can make impact on the performance of K-means clustering. Along this line, we provide an organized study of the effect of skewed data distributions on K-means clustering. The results can guide us for the better use of K-means. This is considered valuable, since K-means has been shown to perform as well as or better than a variety of other clustering techniques in text clustering, and has an appealing computational efficiency [17, 22, 29].

In this chapter, we first formally illustrate that K-means tends to produce clusters in relatively uniform sizes, even if the input data have varying true cluster sizes. Also, we show that some clustering validation measures, such as the entropy measure, may not capture the uniform effect of K-means, and thus provide misleading evaluations on the clustering results. To deal with this, the Coefficient of Variation (CV) [5] statistic is employed as a complement for cluster validation. That is, if the CV value of the cluster sizes has a significant change before and after the clustering, the clustering performance is considered questionable. However, the reverse is not true; that is, a minor change of the CV value does not necessarily indicate a good clustering result.

In addition, we have conducted extensive experiments on a number of real-world data sets, including text data, gene expression data, and UCI data, obtained from different application domains. Experimental results demonstrate that, for data with highly varying true cluster sizes (e.g. $CV > 1.0$), K-means tends to generate clusters in relatively uniform sizes ($CV < 1.0$). In contrast, for data sets with uniform true cluster sizes (e.g. $CV < 0.3$), K-means tends to generate clusters in varying sizes

($CV > 0.3$). In other words, for these two cases, the clustering performance of K-means is often poor.

The remainder of this chapter is organized as follows. Section 2.2 formally illustrates the uniform effect of K-means clustering. In Sect. 2.3, we illustrate the biased effect of the entropy measure. Section 2.4 shows experimental results. The related work is presented in Sect. 2.5, and we finally draw conclusions in Sect. 2.6.

2.2 The Uniform Effect of K-means Clustering

In this section, we mathematically formulate the fact that K-means clustering tends to produce clusters in uniform sizes, which is also called the *uniform effect* of K-means.

K-means is typically expressed by an objective function that depends on the proximities of the data points to the cluster centroids. Let $X = \{x_1, \dots, x_n\}$ be the data, and $m_l = \sum_{x \in C_l} \frac{x}{n_l}$ be the centroid of cluster C_l , $1 \leq l \leq k$, where n_l is the number of data objects in cluster C_l , and k is the number of clusters. The objective function of K-means clustering is then formulated as the sum of squared errors as follows:

$$F_k = \sum_{l=1}^k \sum_{x \in C_l} \|x - m_l\|^2. \quad (2.1)$$

Let $d(C_p, C_q) = \sum_{x_i \in C_p} \sum_{x_j \in C_q} \|x_i - x_j\|^2$. We have the sum of all pair-wise distances of data objects within k clusters as follows:

$$D_k = \sum_{i=1}^n \sum_{j=1}^n \|x_i - x_j\|^2 = \sum_{l=1}^k d(C_l, C_l) + 2 \sum_{1 \leq i < j \leq k} d(C_i, C_j). \quad (2.2)$$

Note that D_k is a constant for a given data set regardless of k . We use the subscript k for the convenience of the mathematical induction. Also, $n = \sum_{l=1}^k n_l$ is the total number of objects in the data.

2.2.1 Case I: Two Clusters

Here, we first illustrate the uniform effect of K-means when the number of clusters is only two. We have,

$$D_2 = \sum_{i=1}^n \sum_{j=1}^n \|x_i - x_j\|^2 = d(C_1, C_1) + d(C_2, C_2) + 2d(C_1, C_2).$$

In this case, D_2 is also a constant, and $n = n_1 + n_2$ is the total number of data objects. If we substitute m_l in Eq. (2.1) by $\sum_{x \in C_l} \frac{x}{n_l}$, we have

$$F_2 = \frac{1}{2n_1} \sum_{x_i, x_j \in C_1} \|x_i - x_j\|^2 + \frac{1}{2n_2} \sum_{x_i, x_j \in C_2} \|x_i - x_j\|^2 = \frac{1}{2} \sum_{l=1}^2 \frac{d(C_l, C_l)}{n_l}. \quad (2.3)$$

If we let

$$F_D^{(2)} = -n_1 n_2 \left[\frac{d(C_1, C_1)}{n_1^2} + \frac{d(C_2, C_2)}{n_2^2} - 2 \frac{d(C_1, C_2)}{n_1 n_2} \right],$$

we thus have

$$F_2 = -\frac{F_D^{(2)}}{2n} + \frac{D_2}{2n}. \quad (2.4)$$

Furthermore, we can show that

$$\frac{2d(C_1, C_2)}{n_1 n_2} = \frac{d(C_1, C_1)}{n_1^2} + \frac{d(C_2, C_2)}{n_2^2} + 2\|m_1 - m_2\|^2.$$

Therefore, we finally have

$$F_D^{(2)} = 2n_1 n_2 \|m_1 - m_2\|^2.$$

Equation (2.4) indicates that the minimization of the K-means objective function F_2 is equivalent to the maximization of the distance function $F_D^{(2)}$. As $F_D^{(2)} > 0$ when m_1 is not equal to m_2 , if we isolate the effect of $\|m_1 - m_2\|^2$, the maximization of $F_D^{(2)}$ implies the maximization of $n_1 n_2$, which leads to $n_1 = n_2 = n/2$.

Discussion. In the above analysis, we have isolated the effect of two components: $\|m_1 - m_2\|^2$ and $n_1 n_2$. For real-world data sets, the values of these two components are related to each other. Indeed, under certain circumstances, the goal of maximizing $n_1 n_2$ may contradict the goal of maximizing $\|m_1 - m_2\|^2$. Figure 2.1 illustrates such a scenario when $n_1 n_2$ is dominated by $\|m_1 - m_2\|^2$. In this example, we generate two true clusters, i.e. one `stick` cluster and one `circle` cluster, each of which contains 500 objects. If we apply K-means on these two data sets, we can have the clustering results in which 106 objects of the `stick` cluster are assigned to the `circle` cluster, as indicated by the green dots in the `stick` cluster. In this way, while the value of $n_1 n_2$ decreases a little bit, the value of $\|m_1 - m_2\|^2$ increases more significantly, which finally leads to the decrease of the overall objective function value. This implies that, K-means will increase the variation of true cluster sizes slightly in this scenario. However, it is hard to further clarify the relationship between these two components in theory, as this relationship is affected by many factors, such

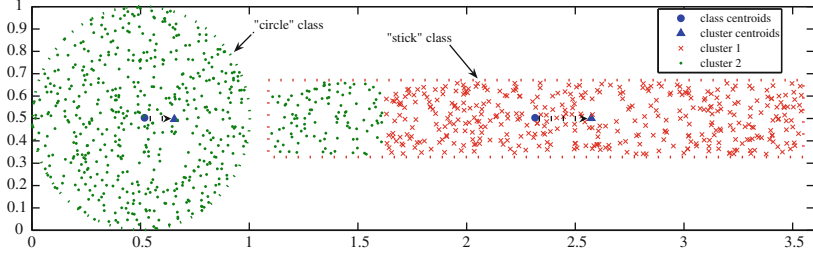


Fig. 2.1 Illustration of the violation of the uniform effect. © 2009 IEEE. Reprinted, with permission, from Ref. [25]

as the shapes of clusters and the densities of data. As a complement, we present an extensive experimental study in Sect. 2.4 to provide a better understanding to this.

2.2.2 Case II: Multiple Clusters

Here, we consider the case that the number of clusters is greater than two. If we substitute m_l , the centroid of cluster C_l , in Eq. (2.1) by $\sum_{x \in C_l} \frac{x}{n_l}$, we have

$$F_k = \sum_{l=1}^k \left(\frac{1}{2n_l} \sum_{x_i, x_j \in C_l} \|x_i - x_j\|^2 \right) = \frac{1}{2} \sum_{l=1}^k \frac{d(C_l, C_l)}{n_l}. \quad (2.5)$$

We then decompose F_k by using the two lemmas as follows:

Lemma 2.1

$$D_k = \sum_{l=1}^k \frac{n}{n_l} d(C_l, C_l) + 2 \sum_{1 \leq i < j \leq k} n_i n_j \|m_i - m_j\|^2. \quad (2.6)$$

Proof We use the mathematical induction.

When $k = 1$, by Eq. (2.2), the left hand side of Eq. (2.6) is $d(C_1, C_1)$. The right hand side of Eq. (2.6) is also equal to $d(C_1, C_1)$, as there is no cross-cluster item. As a result, Lemma 2.1 holds.

When $k = 2$, by Eq. (2.2), to prove Eq. (2.6) is equivalent to prove the following equation:

$$2d(C_1, C_2) = \frac{n_2}{n_1} d(C_1, C_1) + \frac{n_1}{n_2} d(C_2, C_2) + 2n_1 n_2 \|m_1 - m_2\|^2. \quad (2.7)$$

If we substitute $m_1 = \frac{\sum_{i=1}^{n_1} x_i}{n_1}$, $m_2 = \frac{\sum_{i=1}^{n_2} y_i}{n_2}$, and

$$\begin{aligned}
d(C_1, C_1) &= 2 \sum_{1 \leq i < j \leq n_1} \|x_i - x_j\|^2 = 2(n_1 - 1) \sum_{i=1}^{n_1} \|x_i\|^2 - 4 \sum_{1 \leq i < j \leq n_1} x_i x_j, \\
d(C_2, C_2) &= 2 \sum_{1 \leq i < j \leq n_2} \|y_i - y_j\|^2 = 2(n_2 - 1) \sum_{i=1}^{n_2} \|y_i\|^2 - 4 \sum_{1 \leq i < j \leq n_2} y_i y_j, \\
d(C_1, C_2) &= \sum_{1 \leq i \leq n_1} \sum_{1 \leq j \leq n_2} \|x_i - y_j\|^2 = 2n_2 \sum_{i=1}^{n_1} \|x_i\|^2 + 2n_1 \sum_{i=1}^{n_2} \|y_i\|^2 \\
&\quad - 4 \sum_{1 \leq i \leq n_1} \sum_{1 \leq j \leq n_2} x_i y_j
\end{aligned}$$

into Eq. (2.7), we can show that the left hand side will be equal to the right hand side. Therefore, Lemma 2.1 also holds for $k = 2$.

Now we assume that Lemma 2.1 also holds when the cluster number is $k - 1$. Then for the case that the cluster number is k , we first define $D_{k-1}^{(i)}$ as the sum of squared pair-wise distances between data objects within $k - 1$ clusters selected from the total k clusters excluding cluster i . It is trivial to note that $D_{k-1}^{(i)} < D_k$, and they have relationship as follows:

$$D_k = D_{k-1}^{(p)} + d(C_p, C_p) + 2 \sum_{1 \leq j \leq k, j \neq p} d(C_p, C_j). \quad (2.8)$$

Note that Eq. (2.8) holds for any $p = 1, 2, \dots, k$. So actually we have k equations. We sum up these k equations and get

$$kD_k = \sum_{p=1}^k D_{k-1}^{(p)} + \sum_{p=1}^k d(C_p, C_p) + 4 \sum_{1 \leq i < j \leq k} d(C_i, C_j). \quad (2.9)$$

As Eq. (2.6) holds for the case that the cluster number is $k - 1$, we have

$$D_{k-1}^{(p)} = \sum_{1 \leq l \leq k, l \neq p} \left[\frac{n - n_p}{n_l} d(C_l, C_l) \right] + 2 \sum_{1 \leq i < j \leq k}^{i, j \neq p} [n_i n_j \|m_i - m_j\|^2].$$

So the first part of the right hand side of Eq. (2.9) is

$$\begin{aligned}
\sum_{p=1}^k D_{k-1}^{(p)} &= (k - 2) \left(\sum_{l=1}^k \frac{n}{n_l} d(C_l, C_l) + 2 \sum_{1 \leq i < j \leq k} n_i n_j \|m_i - m_j\|^2 \right) \\
&\quad + \sum_{l=1}^k d(C_l, C_l).
\end{aligned} \quad (2.10)$$

Accordingly, we can further transform Eq.(2.9) into

$$\begin{aligned}
 kD_k = (k-2) & \left(\sum_{l=1}^k \left[\frac{n}{n_l} d(C_l, C_l) \right] + 2 \sum_{1 \leq i < j \leq k} [n_i n_j \|m_i - m_j\|^2] \right) \\
 & + 2 \left[\sum_{l=1}^k d(C_l, C_l) + 2 \sum_{1 \leq i < j \leq k} d(C_i, C_j) \right]. \quad (2.11)
 \end{aligned}$$

According to Eq.(2.2), we know that the second part of the right hand side of Eq.(2.11) is exactly $2D_k$. So we can finally have

$$D_k = \sum_{l=1}^k \frac{n}{n_l} d(C_l, C_l) + 2 \sum_{1 \leq i < j \leq k} n_i n_j \|m_i - m_j\|^2,$$

which implies that Lemma 2.1 also holds for the case that the cluster number is k . We complete the proof. \square

Lemma 2.2 *Let*

$$F_D^{(k)} = D_k - 2nF_k. \quad (2.12)$$

Then

$$F_D^{(k)} = 2 \sum_{1 \leq i < j \leq k} [n_i n_j \|m_i - m_j\|^2]. \quad (2.13)$$

Proof If we substitute F_k in Eq.(2.5) and D_k in Eq.(2.6) into Eq.(2.12), we can know that Eq.(2.13) is true. \square

Discussion. By Eq.(2.12), we know that the minimization of the K-means objective function F_k is equivalent to the maximization of the distance function $F_D^{(k)}$, where both D_k and n are constants for a given data set. For $F_D^{(k)}$ in Eq.(2.13), if we assume for all $1 \leq i < j \leq k$, $\|m_i - m_j\|^2$ are the same, i.e. all the pair-wise distances between two centroids are the same, then it is easy to show that the maximization of $F_D^{(k)}$ is equivalent to the uniform distribution of n_i , i.e. $n_1 = n_2 = \dots = n_k = n/k$. Note that we have isolated the effect of two components: $\|m_i - m_j\|^2$ and $n_i n_j$ here to simplify the discussion. For real-world data sets, however, these two components are interactive.

2.3 The Relationship between K-means Clustering and the Entropy Measure

In this section, we study the relationship between K-means clustering and a widely used clustering validation measure: Entropy (E).

2.3.1 The Entropy Measure

Generally speaking, there are two types of clustering validation techniques [10, 13], which are based on external and internal criteria, respectively. Entropy is an external validation measure using the class labels of data as external information. It has been widely used for a number of K-means clustering applications [22, 29].

Entropy measures the purity of the clusters with respect to the given class labels. Thus, if each cluster consists of objects with a single class label, the entropy value is 0. However, as the class labels of objects in a cluster become more diverse, the entropy value increases.

To compute the entropy of a set of clusters, we first calculate the class distribution of the objects in each cluster. That is, for each cluster j , we compute p_{ij} , the probability of assigning an object of class i to cluster j . Given this class distribution, the entropy of cluster j is calculated as

$$E_j = - \sum_i p_{ij} \log(p_{ij}),$$

where the sum is taken over all classes. The total entropy for a set of clusters is computed as the weighted sum of the entropies of all clusters:

$$E = \sum_{j=1}^m \frac{n_j}{n} E_j,$$

where n_j is the size of cluster j , m is the number of clusters, and n is the total number of data objects.

2.3.2 The Coefficient of Variation Measure

Before we describe the relationship between the entropy measure and K-means clustering, we first introduce the Coefficient of Variation (CV) statistic [5], which is a measure of the data dispersion. CV is defined as the ratio of the standard deviation to the mean. Given a set of data objects $X = \{x_1, x_2, \dots, x_n\}$, we have

$$CV = \frac{s}{\bar{x}}, \quad (2.14)$$

where

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \text{ and } s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}.$$

CV is a dimensionless number that allows comparing the variations of populations that have significantly different mean values. In general, the larger the CV value, the greater the variation in the data.

Recall that K-means clustering has a uniform effect (Sect. 2.2). CV can serve as a good indicator for the detection of the uniform effect. That is, if the CV value of the cluster sizes has a significant change after K-means clustering, we know that the uniform effect exists, and the clustering quality tends to be poor. However, it does not necessarily indicate a good clustering performance if the CV value of the cluster sizes only has a minor change after the clustering.

2.3.3 The Limitation of the Entropy Measure

In practice, we have observed that the entropy measure tends to favor clustering algorithms, such as K-means, which produce clusters with relatively uniform sizes. We call this the *biased effect* of the entropy measure. To illustrate this, we create a sample data set shown in Table 2.1. This data set consists of 42 documents belonging to five classes, i.e. five true clusters, whose CV value is 1.119.

For this data set, assume we have two clustering results generated by different clustering algorithms, as shown in Table 2.2. In the table, we can observe that the first clustering result has five clusters with relatively uniform sizes. This is also indicated by the CV value of 0.421. In contrast, for the second clustering result, the CV value of the cluster sizes is 1.201, which indicates a severe imbalance. According to the entropy measure, clustering result *I* is better than clustering result *II*. This is due to the fact that the entropy measure penalizes a large impure cluster more just as the first cluster in clustering *I*. However, if we look at the five true clusters carefully, we can find that the second clustering result is much closer to the true cluster distribution, and the first clustering result is actually far away from the true cluster distribution. This is also reflected by the CV values; that is, the CV value (1.201) of five cluster sizes in the second clustering result is much closer to the CV value (1.119) of five true cluster sizes.

In summary, this example illustrates that the entropy measure tends to favor K-means which produces clusters in relatively uniform sizes. This effect becomes even more significant in the situation that the data have highly imbalanced true clusters. In other words, if the entropy measure is used for validating K-means clustering, the validation result can be misleading.

Table 2.1 A document data set

1: Sports, Sports	24 objects
2: Entertainment, Entertainment	2 objects
3: Foreign, Foreign, Foreign, Foreign, Foreign	5 objects
4: Metro, Metro, Metro, Metro, Metro, Metro, Metro, Metro, Metro, Metro	10 objects
5: Politics	1 object
$CV = 1.119$	

Table 2.2 Two clustering results

Clustering <i>I</i>	1: Sports Sports Sports Sports Sports Sports Sports Sports	$CV = 0.421$
	2: Sports Sports Sports Sports Sports Sports Sports Sports	$E = 0.247$
	3: Sports Sports Sports Sports Sports Sports Sports Sports	
	4: Metro Metro Metro Metro Metro Metro Metro Metro Metro	
	5: Entertainment Entertainment Foreign Foreign Foreign Foreign Foreign Politics	
Clustering <i>II</i>	1: Sports Sports Sports Sports Sports Sports Sports Sports Sports Sports Sports Sports Sports Sports Sports Foreign	$CV = 1.201$
	2: Entertainment Entertainment	$E = 0.259$
	3: Foreign Foreign Foreign	
	4: Metro Metro Metro Metro Metro Metro Metro Metro Metro Metro Foreign	
	5: Politics	

2.4 Experimental Results

In this section, we conduct experiments on a number of real-world data sets to show the uniform effect of K-means clustering and the bias effect of the entropy measure.

2.4.1 Experimental Setup

We first introduce the experimental setup, including the clustering tools and the data information.

Clustering Tools. In our experiments, we used the CLUTO implementation of K-means.¹ As the Euclidean notion of proximity is not very effective for K-means

¹ <http://glaros.dtc.umn.edu/gkhome/views/cluto>

Table 2.3 Some notations used in experiments

CV_0 :	The CV value of the true cluster sizes
CV_1 :	The CV value of the resulting cluster sizes
DCV :	$CV_0 - CV_1$
\bar{S} :	The average cluster sizes
STD_O :	The standard deviation of the true cluster sizes
STD_1 :	The standard deviation of the resulting cluster sizes
E :	The entropy measure

Table 2.4 Some characteristics of experimental data sets

Data	Source	#object	#feature	#class	MinClassSize	MaxClassSize	CV_0
<i>Document data</i>							
fbis	TREC	2463	2000	17	38	506	0.961
hitech	TREC	2301	126373	6	116	603	0.495
sports	TREC	8580	126373	7	122	3412	1.022
tr23	TREC	204	5832	6	6	91	0.935
tr45	TREC	690	8261	10	14	160	0.669
la2	TREC	3075	31472	6	248	905	0.516
ohscal	OHSUMED-233445	11162	11465	10	709	1621	0.266
re0	Reuters-21578	1504	2886	13	11	608	1.502
re1	Reuters-21578	1657	3758	25	10	371	1.385
k1a	WebACE	2340	21839	20	9	494	1.004
k1b	WebACE	2340	21839	6	60	1389	1.316
wap	WebACE	1560	8460	20	5	341	1.040
<i>Biomedical data</i>							
LungCancer	KRBDSR	203	12600	5	6	139	1.363
Leukemia	KRBDSR	325	12558	7	15	79	0.584
<i>UCI data</i>							
ecoli	UCI	336	7	8	2	143	1.160
page-blocks	UCI	5473	10	5	28	4913	1.953
pendigits	UCI	10992	16	10	1055	1144	0.042
letter	UCI	20000	16	26	734	813	0.030

clustering on high-dimensional data, the cosine similarity is used in the objective function of K-means. Some notations used in the experiments are shown in Table 2.3.

Experimental Data. We used a number of real-world data sets that were obtained from different application domains. Some characteristics of these data sets are shown in Table 2.4.

Document Data. The `fbis` data set was obtained from the Foreign Broadcast Information Service data of the TREC-5 collection.² The `hitech` and `sports` data sets were derived from the San Jose Mercury newspaper articles that were distributed as part of the TREC collection (TIPSTER Vol. 3). The `hitech` data

² <http://trec.nist.gov>

set contains documents about computers, electronics, health, medical, research, and technology, and the `sports` data set contains documents about baseball, basketball, bicycling, boxing, football, golfing, and hockey. Data sets `tr23` and `tr45` were derived from the TREC-5, TREC-6, and TREC-7 collections. The `la2` data set is part of the TREC-5 collection and contains news articles from the Los Angeles Times. The `ohscal` data set was obtained from the OHSUMED collection [12], which contains documents from the antibodies, carcinoma, DNA, in-vitro, molecular sequence data, pregnancy, prognosis, receptors, risk factors, and tomography categories. The data sets `re0` and `re1` were from Reuters-21578 text categorization test collection Distribution 1.0.³ The data sets `k1a` and `k1b` contain exactly the same set of documents but differ in how the documents were assigned to different classes. In particular, `k1a` contains a finer-grain categorization than that contained by `k1b`. The data set `wap` was obtained from the WebACE project (WAP) [11]; each document corresponds to a web page listed in the subject hierarchy of Yahoo!. For all these data sets, we used a stop-list to remove common words, and the words were stemmed using Porter's suffix-stripping algorithm [20].

Biomedical Data. `LungCancer` [1] and `Leukemia` [26] data sets were obtained from Kent Ridge Biomedical Data Set Repository (KRBDSR), which is an online repository of high-dimensional biomedical data.⁴ The `LungCancer` data set consists of samples of lung adenocarcinomas, squamous cell lung carcinomas, pulmonary carcinoid, small-cell lung carcinomas, and normal lung described by 12600 genes. The `Leukemia` data set contains six subtypes of pediatric acute lymphoblastic leukemia samples and one group samples that do not fit in any of the above six subtypes, and each sample is described by 12558 genes.

UCI Data. In addition to the high-dimensional data above, we also used some UCI data sets in lower dimensionality.⁵ The `ecoli` data set is about the information of cellular localization sites of proteins. The `page-blocks` data set contains the information of five type blocks of the page layout of a document that is detected by a segmentation process. The `pendigits` and `letter` data sets contain the information of handwritings. The `pendigits` data set includes the number information of 0–9, while the `letter` data set contains the letter information of A–Z.

Note that for each data set in Table 2.4, the experiment was conducted ten times to void the randomness, and the average value is presented.

2.4.2 The Evidence of the Uniform Effect of K-means

Here, we illustrate the uniform effect of K-means clustering. In the experiment, we first used CLUTO with default settings to cluster the input data sets, and then

³ <http://www.research.att.com/~lewis>.

⁴ <http://sdmc.i2r.a-star.edu.sg/rp/>

⁵ <http://www.ics.uci.edu/~mllearn/MLRepository.html>

Table 2.5 Experimental results on real-world data sets

Data	\bar{S}	STD_0	STD_1	CV_0	CV_1	DCV	E
fbis	145	139	80	0.96	0.55	0.41	0.345
hitech	384	190	140	0.50	0.37	0.13	0.630
k1a	117	117	57	1.00	0.49	0.51	0.342
k1b	390	513	254	1.32	0.65	0.66	0.153
la2	513	264	193	0.52	0.38	0.14	0.401
ohscal	1116	297	489	0.27	0.44	-0.17	0.558
re0	116	174	45	1.50	0.39	1.11	0.374
re1	66	92	22	1.39	0.32	1.06	0.302
sports	1226	1253	516	1.02	0.42	0.60	0.190
tr23	34	32	14	0.93	0.42	0.51	0.418
tr45	69	46	30	0.67	0.44	0.23	0.329
wap	78	81	39	1.04	0.49	0.55	0.313
LungCancer	41	55	26	1.36	0.63	0.73	0.332
Leukemia	46	27	17	0.58	0.37	0.21	0.511
ecoli	42	49	21	1.16	0.50	0.66	0.326
page-blocks	1095	2138	1029	1.95	0.94	1.01	0.146
letter	769	23	440	0.03	0.57	-0.54	0.683
pendigits	1099	46	628	0.04	0.57	-0.53	0.394
Min	34	23	14	0.03	0.33	-0.54	0.146
Max	1226	2138	1029	1.95	0.94	1.11	0.683

computed the CV values of the cluster sizes. The number of clusters K was set to the true cluster number for the purpose of comparison.

Table 2.5 shows the experimental results on real-world data sets. As can be seen, for 15 data sets with relatively large CV_0 values, K-means tends to reduce the variation of the cluster sizes in the clustering results, as indicated by the smaller CV_1 values. This means that the uniform effect of K-means exists for data sets with highly imbalanced true clusters. Indeed, if we look at Eq. (2.13) in Sect. 2.2, this result implies that the factor $\|m_i - m_j\|^2$ is dominated by the factor $n_i n_j$.

For data sets `ohscal`, `letter`, and `pendigits` with very small CV_0 values, however, K-means increases the variation of the cluster sizes slightly, as indicated by the corresponding CV_1 values. This implies that the uniform effect of K-means is not significant for data sets with true clusters in relatively uniform sizes. Indeed, according to Eq. (2.13) in Sect. 2.2, this result indicates that the factor $n_i n_j$ is dominated by the factor $\|m_i - m_j\|^2$.

2.4.3 The Quantitative Analysis of the Uniform Effect

In this experiment, we attempt to get a quantitative understanding about the uniform effect of K-means on the clustering results.

Fig. 2.2 The linear relationship between DCV and CV_0 . © 2009 IEEE. Reprinted, with permission, from Ref. [25]

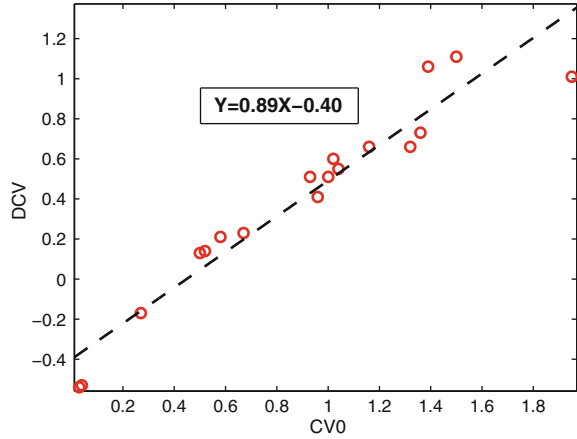


Fig. 2.3 The relationship between CV_1 and CV_0 . © 2009 IEEE. Reprinted, with permission, from Ref. [25]

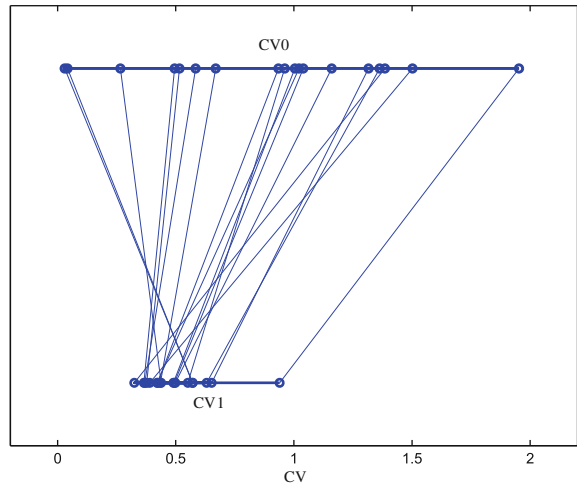
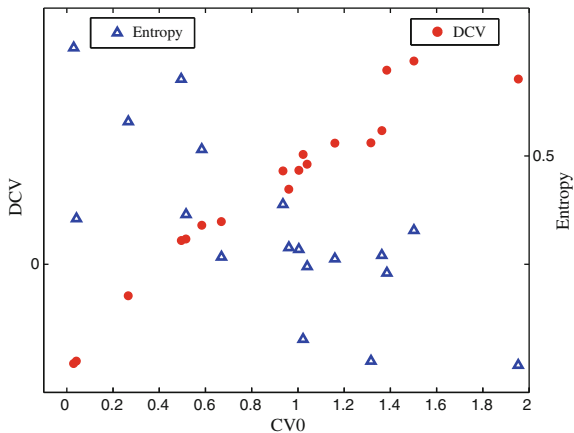


Figure 2.2 shows the relationship between DCV and CV_0 , in which all the points (CV_0 , DCV) are fitted into a linear line: $y = 0.89x - 0.40$. Apparently, the DCV value increases with the increase of the CV_0 value, and $y = 0$ when $x = 0.45$ in the linear fitting line. This indicates that if $CV_0 > 0.45$, K-means clustering tends to have $CV_1 < CV_0$. Otherwise, $CV_1 > CV_0$. In other words, 0.45 is the empirical threshold to invoke the uniform effect of K-means.

Figure 2.3 shows the relationship between CV_0 and CV_1 for all the experimental data sets listed in Table 2.4. Note that there is a link between CV_0 and CV_1 for every data set. An interesting observation is that, while the range of CV_0 is between 0.03 and 1.95, the range of CV_1 is restricted into a much narrower range from 0.33 to 0.94. So we have the empirical value range of CV_1 : $[0.3, 1]$.

Fig. 2.4 Illustration of the biased effect of the entropy measure. © 2009 IEEE. Reprinted, with permission, from Ref. [25]



2.4.4 The Evidence of the Biased Effect of the Entropy Measure

In this subsection, we present the biased effect of the entropy measure on the clustering results of K-means. Figure 2.4 shows the entropy values of the clustering results of all 18 data sets. A general trend is that while the difference of the cluster size distributions before and after clustering increases with the increase of CV_0 , the entropy value tends to decrease. In other words, there is a disagreement between DCV and the entropy measure on evaluating the clustering quality. Entropy indicates a higher quality as CV_0 increases, but DCV denies this by showing that the distribution of the clustering result is getting farther away from the true distribution. This observation well agrees with our analysis in Sect. 2.3 that entropy has a biased effect on K-means.

To further illustrate the biased effect of entropy, we also generated two groups of synthetic data sets. These data sets have wide ranges of distributions of true cluster sizes. The first group of synthetic data sets was derived from the `pendigits` data set. We applied the following sampling strategy: (1) The original data set was first sampled to get a sample of 10 classes, each of which contains 1000, 100, 100, \dots , 100 objects, respectively; (2) To get data sets with decreasing CV_0 values, the size of the largest class was gradually reduced from 1000 to 100; (3) To get data sets with increasing CV_0 values, the sizes of the remaining nine classes were gradually reduced from 100 to 30. A similar sampling strategy was also applied to the `letter` data set for generating the second group of synthetic data sets. Note that we repeated sampling a data set ten times, and output the average evaluation of the clustering results.

Figures 2.5 and 2.6 show the clustering results evaluated by the entropy measure and DCV , respectively. A similar trend can be observed; that is, the entropy value decreases as the CV_0 value increases. This further justifies the existence of the biased effect of the entropy measure on the clustering result of K-means.

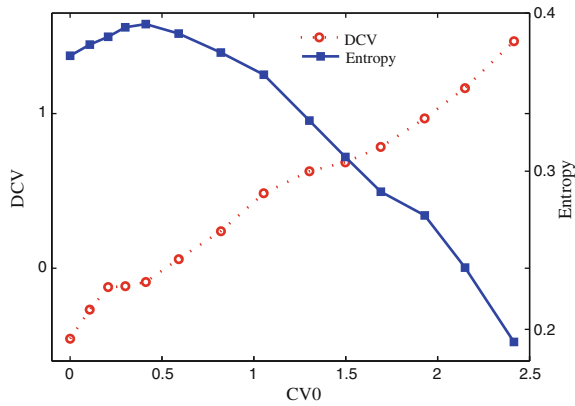


Fig. 2.5 The biased effect of entropy: synthetic data from `pendigits`. © 2009 IEEE. Reprinted, with permission, from Ref. [25]

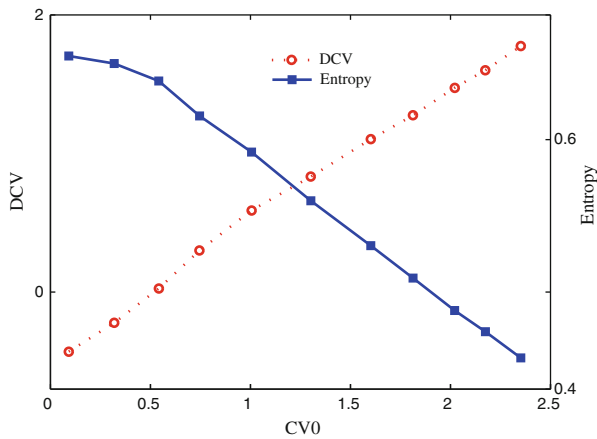


Fig. 2.6 The biased effect of entropy: synthetic data from `letter`. © 2009 IEEE. Reprinted, with permission, from Ref. [25]

2.4.5 The Hazard of the Biased Effect

Having the biased effect, it is very dangerous to use the entropy measure for the validation of K-means. To illustrate this, we selected five data sets with high CV_0 values, i.e. `re0`, `re1`, `wap`, `ecoli`, and `k1a`, for experiments. We did K-means clustering on these data sets, and labeled each cluster by the label of the members in majority. We found that many true clusters were disappeared in the clustering results. Figure 2.7 shows the percentage of the disappeared true clusters in the clustering results. As can be seen, every data set has a significant number of true clusters disappeared. For the `re0` data set ($CV_0 = 1.502$), even more than 60 % true clusters

Fig. 2.7 The percentage of the disappeared true clusters in highly imbalanced data. © 2009 IEEE. Reprinted, with permission, from Ref. [25]

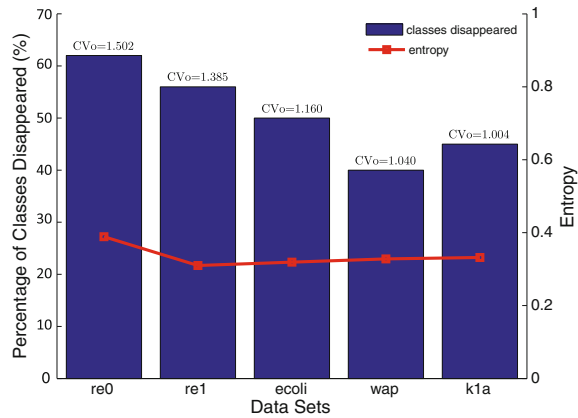
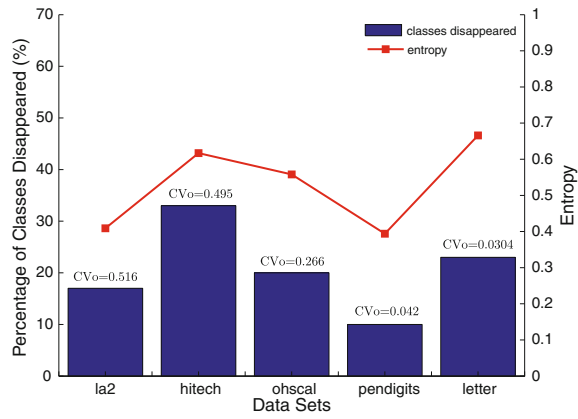


Fig. 2.8 The percentage of the disappeared true clusters in relatively balanced data. © 2009 IEEE. Reprinted, with permission, from Ref. [25]



disappear after K-means clustering! In sharp contrast, as shown in Fig. 2.7, very low entropy values were achieved for these five data sets, which imply that the performance of K-means clustering is “excellent”. This experiment clearly illustrates the hazard of the biased effect of entropy.

For the purpose of comparison, we also conducted a similar experiment on five data sets with low CV_0 values. Figure 2.8 shows the percentage of the disappeared true clusters. An interesting observation is that, compared to the results of data sets with high CV_0 values, the percentages of the disappeared true clusters become much smaller, and the entropy values increase. In other words, the entropy measure is more reliable for data sets with relatively uniform true cluster sizes.

2.5 Related Work

People have investigated K-means clustering from various perspectives. Many data factors, which may strongly affect the performance of K-means clustering, have been identified and addressed. In the following, we highlight some research results which are most related to the main theme of this chapter.

First, people have studied the impact of high dimensionality on the performance of K-means clustering, and found that the traditional Euclidean notion of proximity is not very effective for K-means clustering on real-world high-dimensional data, such as gene expression data and document data. To meet this challenge, one research direction is to make use of dimension reduction techniques, such as Multidimensional Scaling (MDS) [2], Principal Components Analysis (PCA) [15], and Singular Value Decomposition (SVD) [6]. Also, several feature transformation techniques have been proposed for high-dimensional document data, such as Latent Semantic Indexing (LSI), Random Projection (RP), and Independent Component Analysis (ICA). In addition, feature selection techniques have been widely used, and a detailed discussion and comparison of these techniques have been provided by Tang et al. [24]. Another direction for this problem is to redefine the notions of proximity, e.g. by the Shared Nearest Neighbors (SNN) similarity introduced by Jarvis and Patrick [14]. Finally, some other similarity measures, e.g. the cosine measure, have also shown appealing effects on clustering document data [29].

Second, it has been recognized that K-means has difficulty in detecting the “natural” clusters with non-globular shapes [13, 23]. To address this, one research direction is to modify the K-means clustering algorithm. For instance, Guha et al. [9] proposed the CURE method which makes use of multiple representative points to get the shape information of the “natural” clusters. Another research direction is to use some non-prototype-based clustering methods which usually perform better than K-means on data in non-globular or irregular shapes [23].

Third, outliers and noise in the data can also degrade the performance of clustering algorithms [16, 27, 30], especially for prototype-based algorithms such as K-means. To deal with this, one research direction is to incorporate some outlier removal techniques before conducting K-means clustering. For instance, a simple method of detecting outliers is based on the distance measure [16]. Breunig et al. [4] proposed a density based method using the Local Outlier Factor (LOF) for the purpose of identifying outliers in data with varying densities. There are also some other clustering based methods to detect outliers as small and remote clusters [21], or objects that are farthest from their corresponding cluster centroids [18]. Another research direction is to handle outliers during the clustering process. There have been several techniques designed for such purpose. For example, DBSCAN automatically classifies low-density points as noise points and removes them from the clustering process [8]. Also, SNN density-based clustering [7] and CURE [9] explicitly deal with noise and outliers during the clustering process.

Fourth, many clustering algorithms that work well for small or medium-size data are unable to handle large-scale data. Along this line, a discussion of scaling

K-means clustering to large-scale data was provided by Bradley et al. [3]. A broader discussion of specific clustering techniques can be found in [19]. Some representative techniques include CURE [9], BIRCH [28], and so on.

Finally, some researchers have identified some other factors, such as the types of attributes and data sets, that may impact the performance of K-means clustering. However, in this chapter, we focused on understanding the uniform effect of K-means and the biased effect of the entropy measure, which have not been systematically studied in the literature.

2.6 Concluding Remarks

In this chapter, we present an organized study on K-means clustering and cluster validation measures from a data distribution perspective. We first theoretically illustrate that K-means clustering tends to produce clusters with uniform sizes. We then point out that the widely adopted validation measure entropy has a biased effect and therefore cannot detect the uniform effect of K-means. Extensive experiments on a number of real-world data sets clearly illustrate the uniform effect of K-means and the biased effect of the entropy measure, via the help of the Coefficient of Variation statistic. Most importantly, we unveil the danger induced by the combined use of K-means and the entropy measure. That is, many true clusters will become unidentifiable when applying K-means for highly imbalanced data, but this situation is often disguised by the low values of the entropy measure.

References

1. Bhattacharjee, A., Richards, W., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E., Lander, E., Wong, W., Johnson, B., Golub, T., Sugarbaker, D., Meyerson, M.: Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. In: Proceedings of the National Academy of Sciences of the United States of America, **98**, 13790–13795 (2001)
2. Borg, I., Groenen, P.: Modern Multidimensional Scaling—Theory and Applications. Springer Verlag, New York (1997)
3. Bradley, P., Fayyad, U., Reina, C.: Scaling clustering algorithms to large databases. In: Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 9–15 (1998)
4. Breunig, M., Kriegel, H.P., Ng, R., Sander, J.: Lof: identifying density based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 427–438 (2000)
5. DeGroot, M., Schervish, M.: Probability and Statistics, 3rd edn. Addison Wesley, Upper Saddle River (2001)
6. Demmel, J.: Applied numerical linear algebra. Soc. Ind. App. Math. **32**, 206–216 (1997)
7. Ertoz, L., Steinbach, M., Kumar, V.: A new shared nearest neighbor clustering algorithm and its applications. In: Proceedings of the Workshop on Clustering High Dimensional Data and its Applications at the 2nd SIAM International Conference on Data Mining (2002)

8. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the 2th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 226–231 (1996)
9. Guha, S., Rastogi, R., Shim, K.: Cure: an efficient clustering algorithm for large databases. In: *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp. 73–84 (1998)
10. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Cluster validity methods: Part i. *SIGMOD Record* **31**(2), 40–45 (2002)
11. Han, E.H., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., Moore, J.: Webace: a web agent for document categorization and exploration. In: *Proceedings of the 2nd International Conference on Autonomous Agents* (1998)
12. Hersh, W., Buckley, C., Leone, T.J., Hickam, D.: Ohsumed: an interactive retrieval evaluation and new large test collection for research. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in, Information Retrieval*, pp. 192–201 (1994)
13. Jain, A., Dubes, R.: *Algorithms for Clustering Data*. Prentice Hall, Englewood cliff (1998)
14. Jarvis, R., Patrick, E.: Clusering using a similarity measure based on shared nearest neighbors. *IEEE Trans. Comput.* **C-22**(11), 1025–1034 (1973)
15. Jolliffe, I.: *Principal Component Analysis*, 2nd edn. Springer Verlag, New York (2002)
16. Knorr, E., Ng, R., Tucakov, V.: Distance-based outliers: algorithms and applications. *VLDB J.* **8**, 237–253 (2000)
17. Krishna, K., Narasimha Murty, M.: Genetic k-means algorithm. *IEEE Trans. Syst. Man Cybern. Part B* **29**(3), 433–439 (1999)
18. Larsen, B., Aone, C.: Fast and effective text mining using linear-time document clustering. In: *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 16–22 (1999)
19. Murtagh, F.: Clustering massive data sets. In: Abello, J., Pardalos, P.M., Resende, M.G. (eds.) *Handbook of Massive Data Sets*, pp. 501–543. Kluwer Academic Publishers, Norwell (2002)
20. Porter, M.F.: An algorithm for suffix stripping. *Program* **14**(3), 130–137 (1980)
21. Portnoy, L., Eskin, E., Stolfo, S.: Intrusion detection with unlabeled data using clustering. In: *Proceedings of the 2001 ACM CSS Workshop on Data Mining Applied to, Security (DMSA-2001)* (2001)
22. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: *Proceedings of the KDD Workshop on Text Mining* (2000)
23. Tan, P.N., Steinbach, M., Kumar, V.: *Introduction to Data Mining*. Addison-Wesley, Upper Saddle River (2005)
24. Tang, B., Shepherd, M., Heywood, M., Luo, X.: Comparing dimension reduction techniques for document clustering. In: *Proceedings of the Canadian Conference on, Artificial Intelligence*, pp. 292–296 (2005)
25. Xiong, H., Wu, J., Chen, J.: K-means clustering versus validation measures: a data-distribution perspective. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **39**(2), 318–331 (2009)
26. Yeoh, E.J.: Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1**, 133–143 (2002)
27. Zhang, J.S., Leung, Y.W.: Robust clustering by pruning outliers. *IEEE Trans. Syst. Man Cybern. Part B* **33**(6), 983–998 (2003)
28. Zhang, T., Ramakrishnan, R., M.Livny: Birch: an efficient data clustering mehtod for very large databases. In: *Proceedings of 1996 ACM SIGMOD International Conference on Management of Data*, pp. 103–114 (1996)
29. Zhao, Y., Karypis, G.: Criterion functions for document clustering: experiments and analysis. *Mach. Learn.* **55**(3), 311–331 (2004)
30. Zhou, A., Cao, F., Yan, Y., Sha, C., He, X.: Distributed data stream clustering: a fast em-based approach. In: *Proceedings of the 23rd International Conference on Data, Engineering*, pp. 736–745 (2007)

Advances in K-means Clustering

A Data Mining Thinking

Wu, J.

2012, XVI, 180 p., Hardcover

ISBN: 978-3-642-29806-6