K-means Clustering Versus Validation Measures: A Data Distribution Perspective[1]

Hui Xiong, *Senior Member, IEEE,* Junjie Wu, and Jian Chen, *Fellow, IEEE*

Abstract

K-means is a well-known and widely used partitional clustering method. While there are considerable research efforts to characterize the key features of the K-means clustering algorithm, further investigation is needed to understand how data distributions can have impact on the performance of K-means clustering. To that end, in this paper, we provide a formal and organized study of the effect of skewed data distributions on K-means clustering. Along this line, we first formally illustrate that K-means tends to produce clusters of relatively uniform size, even if input data have varied "true" cluster sizes. Also, we show that some clustering validation measures, such as the entropy measure, may not capture this uniform effect and provide misleading information on the clustering performance. Viewed in this light, we provide the Coefficient of Variation (CV) as a necessary criterion to validate the clustering results. Our findings reveal that K-means tends to produce clusters in which the variations of cluster sizes, as measured by CV, are in a range of about 0.3 to 1.0. Specifically, for data sets with large variation in "true" cluster sizes (e.g. CV $>$ 1.0), K-means reduces variation in resultant cluster sizes to less than 1.0. In contrast, for data sets with small variation in "true" cluster sizes (e.g. CV $<$ 0.3), K-means increases variation in resultant cluster sizes to greater than 0.3. In other words, for the above two cases, K-means produces the clustering results which are away from the "true" cluster distributions.

Index Terms

K-means Clustering, Clustering Validation, Coefficient of Variation (CV), Entropy, F-measure

[1]A preliminary version of this work has been published as a six-page poster paper in ACM SIGKDD 2006 [41].

Hui Xiong is with the Management Science and Information Systems Department, Rutgers Business School, Rutgers University, E-mail: hxiong@rutgers.edu.

Junjie Wu is with the School of Economics and Management, Beihang Universtiy, E-mail: wujj@buaa.edu.cn

Jian Chen is with the School of Economics and Management, Tsinghua University, E-mail: chenj@sem.tsinghua.edu.cn

I. INTRODUCTION

Cluster analysis [17] provides insight into the data by dividing the objects into groups (clusters) of objects, such that objects in a cluster are more similar to each other than to objects in other clusters. As a well-known and widely used partitional clustering method, K-means [30] has attracted great interest in the literature. There are considerable research efforts to characterize the key features of the K-means clustering algorithms. Indeed, people have identified some data characteristics that may strongly affect the K-means clustering analysis including high dimensionality, the size of the data, the sparseness of the data, noise and outliers in the data, types of attributes and data sets, and scales of attributes [38]. However, further investigation is needed to understand how data distributions can have the impact on the performance of K-means clustering. Along this line, we provide a formal and organized study of the effect of skewed data distributions on K-means clustering. The understanding from this organized study can guide us for the better use of K-means. This is noteworthy since, for document data, K-means has been shown to perform as well as or better than a variety of other clustering techniques and has an appealing computational efficiency [24], [37], [45].

In this paper, we first formally illustrate that K-means tends to produce clusters of relatively uniform sizes, even if input data have varied "true" cluster sizes. Also, we show that some clustering validation measures, such as the entropy measure, may not capture this uniform effect and provide misleading information on the clustering performance. Viewed in this light, we provide the Coefficient of Variation (CV) [9] as a necessary criterion to validate the clustering results. In other words, if the CV values of cluster sizes have a significant change after the clustering process, we know that the clustering performance is poor. However, it does not necessarily indicate a good clustering performance if the CV values of cluster sizes have a minor change after the clustering process. Note that the CV, described in more detail later (Section III A), is a measure of dispersion of a data distribution and is a dimensionless number that allows comparison of the variation of populations that have significantly different mean values. In general, the larger the CV value is, the greater the variability is in the data.

In addition, we have conducted extensive experiments on a number of real-world data sets from different application domains including text document data sets, gene expression data sets, and UCI data sets. Indeed, our experimental results also show that, for data sets with large variation in "true" cluster sizes (e.g. CV > 1.0), K-means reduces variation in resultant cluster sizes to less than 1.0. In contrast, for data sets with small variation in"true" cluster sizes (e.g. CV < 0.3), K-means increases variation slightly in resultant cluster sizes to greater than 0.3. In other words, for these two cases, K-means produces the clustering results which are away from the "true" cluster distributions.

Outline: The remainder of this paper is organized as follows. Section II illustrates the effect of skewed data distributions on K-means clustering. In Section III, we introduce three external clustering validation measures. Section IV shows experimental results. The related work is described in Section V. Finally, we draw conclusions and suggest future work in Section VI.

II. The Effect of K-means Clustering on the Distribution of the Cluster Sizes

K-means [30] is a prototype-based, simple partitional clustering technique which attempts to find k non-overlapping clusters. These clusters are represented by their centroids (a cluster centroid is typically the mean of the points in the cluster). The clustering process of K-means is as follows. First, k initial centroids are selected, where k is specified by the user and indicates the desired number of clusters. Every point in the data is then assigned to the closest centroid, and each collection of points assigned to a centroid forms a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. This process is repeated until no point changes clusters.

A. CASE I: The Number of Clusters is Two

Typically, K-means is expressed by an objective function that depends on the proximities of the data points to one another or to the cluster centroids. In the following, we illustrate the effect of K-means clustering on the distribution of the cluster sizes when the number of clusters is two.

Objective Function: Sum of Squared Errors (SSE)

Let $X = \{x_1, \ldots, x_n\}$ be the data, $m_l = \sum_{x \in C_l} \frac{x}{n_l}$ be the centroid of cluster C_l, n_l be the number of data objects in the cluster C_l, and k be the number of clusters ($1 \leq l \leq k$). Then, an objective function of K-means clustering is the sum of squared error as follows.

$$F_k = \sum_{l=1}^{k} \sum_{x \in C_l} \parallel x - m_l \parallel^2 \tag{1}$$

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

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

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

To simplify the discussion, we first consider the case that $k = 2$, then

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

In this case, D_2 is also a constant and $n = n_1 + n_2$ is the total number of objects in the data. If we substitute the definition of m_l to Equation (1), we have

$$\begin{aligned} F_2 &= \frac{1}{2n_1} \sum_{x_i, x_j \in C_1} \parallel x_i - x_j \parallel^2 + \frac{1}{2n_2} \sum_{x_i, x_j \in C_2} \parallel x_i - x_j \parallel^2 \\ &= \frac{1}{2} \sum_{l=1}^{2} \frac{d(C_l, C_l)}{n_l} \end{aligned} \tag{3}$$

Let

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

we have

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

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,

$$F_D^{(2)} = 2n_1 n_2 \| m_1 - m_2 \|^2 > 0$$

In other words, the minimization of the K-means objective function F_2 is equivalent to the maximization of the distance function $F_D^{(2)}$. Since $F_D^{(2)} > 0$, 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 $n_1 = n_2 = n/2$.

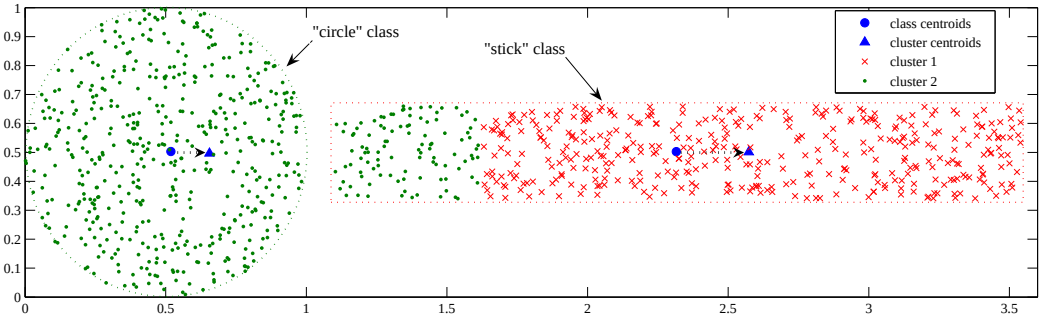


Fig. 1. An Illustrative Example of Potential Violations of the Uniform Effect.

Discussions. 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 $n_1 n_2$ can be contradicted by the goal of maximizing $\| m_1 - m_2 \|^2$. For instance, Figure 1 illustrates a scenario such that $n_1 n_2$ is dominated by $\| m_1 - m_2 \|^2$. In this example, we simulated 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 the 106 objects in the `stick` cluster are assigned to the `circle` cluster, as indicated by the dots in the `stick` cluster. In this way, while $n_1 n_2$ is decreased a little bit, the value of $\| m_1 - m_2 \|^2$

increases more significantly. As a result, the overall objective function value is decreased. Thus, in this scenario, K-means will increase the variation of "true" cluster sizes slightly. However, it is hard to have a further theoretical analysis to clarify the relationship between these two components, since this relationship is affected by many factors, such as cluster shapes and the density in the data. Instead, we present an extensive empirical study in Section IV to provide a better understanding on this.

B. CASE II: The Number of Clusters > 2

Here, we consider the case that the number of clusters is greater than two. In this case, we also use sum of squared errors (SSE) as the objective function. To make the discussion consistent with the case that the number of clusters is two, we still use the same notations as Section II-A for m_l, F_k, D_k, and $d(C_p, C_q)$. First, if we substitute m_l, the centroid of cluster C_l, to Equation (1), we have

$$F_k = \sum_{l=1}^{k} (\frac{1}{2n_l} \sum_{x_i, x_j \in C_l} \parallel x_i - x_j \parallel^2) = \frac{1}{2} \sum_{l=1}^{k} \frac{d(C_l, C_l)}{n_l} \tag{4}$$

Proposition 1: For D_k in Equation (2), we have

$$D_k = \sum_{l=1}^{k} [\frac{n}{n_l} d(C_l, C_l)] + 2 \sum_{1 \le i < j \le k} [n_i n_j \parallel m_i - m_j \parallel^2] \tag{5}$$

Proof: We prove this by mathematical induction.

For $k = 1$, by Equation (2), the left hand side of Equation (5) is $d(C_1, C_1)$. Also, the right hand side of Equation (5) is equal to $d(C_1, C_1)$, since there is no cross-cluster item. As a result, Proposition 1 holds when $k = 1$.

For $k = 2$, by Equation (2), to prove Equation (5) is equivalent to prove the following Equation.

$$2d(C_1, C_2) = \frac{n_2}{n_1} d(C_1, C_1) + \frac{n_1}{n_2} d(C_2, C_2) + 2n_1 n_2 \parallel m_1 - m_2 \parallel^2 \tag{6}$$

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

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

into Equation (6), we can show that the left hand side will be equal to the right hand side. Therefore, Proposition 1 also holds for $k = 2$.

Now we assume that Proposition 1 also holds for the case that 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 total k clusters without cluster i. In other words, if we disregard the data objects in cluster i ($i = 1, 2, \cdots, k$), then the sum of squared pair-wise distances between the rest data objects in the rest $k - 1$ clusters is exactly the value of $D_{k-1}^{(i)}$. It is trivial to know that $D_{k-1}^{(i)} < D_k$, and they have the relationship as follows.

$$
D_k = D_{k-1}^{(p)} + d(C_p, C_p) + 2 \sum_{1 \le j \le k, j \ne p} d(C_p, C_j) \tag{7}
$$

Note that Equation (7) holds for any $p = 1, 2, \cdots, k$, so actually we have k equations. We sum up these k equations and get

$$
k D_k = \sum_{p=1}^{k} D_{k-1}^{(p)} + \sum_{p=1}^{k} d(C_p, C_p) + 4 \sum_{1 \le i < j \le k} d(C_i, C_j) \tag{8}
$$

According to the assumption for the case that the cluster number is $k - 1$, we have

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

So the first part of the right hand side of Equation (8) is

$$\sum_{p=1}^{k} D_{k-1}^{(p)} = (k-2)\left(\sum_{l=1}^{k}[\frac{n}{n_l}d(C_l,C_l)] + 2\sum_{1\le i<j\le k}[n_i n_j \| m_i - m_j \|^2]\right) + \sum_{l=1}^{k} d(C_l,C_l) \quad (9)$$

So we can further transform Equation (8) into

$$kD_k = (k-2)\left(\sum_{l=1}^{k}[\frac{n}{n_l}d(C_l,C_l)] + 2\sum_{1\le i<j\le k}[n_i n_j \| m_i - m_j \|^2]\right) + 2\left[\sum_{l=1}^{k} d(C_l,C_l) + 2\sum_{1\le i<j\le k} d(C_i,C_j)\right] \quad (10)$$

According to Equation (2), we know that the second part of the right hand side of Equation (10) is exactly $2D_k$. So we can finally get

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

In conclusion, for the case that the cluster number is k, Proposition 1 also holds.

Proposition 2: Let

$$F_D^{(k)} = D_k - 2nF_k \quad (11)$$

we have

$$F_D^{(k)} = 2\sum_{1\le i<j\le k}[n_i n_j \| m_i - m_j \|^2] \quad (12)$$

Proof: If we substitute F_k in Equation (4) and D_k in Equation (5) into Equation (11), we can know that Proposition 2 is true.

Discussions. By Equation (11), 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 D_k and n are constants for a given data set. For Equation (12), 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., $n_1 = n_2 = \cdots = n_k = n/k$. Again, to simplify the discussion, we have isolated the effect of two components: $\| m_i - m_j \|^2$ and $n_i n_j$ in the above analysis. However, for real-world data sets, these two components can have the impact on each other.

III. The Relationship between K-means Clustering and Validation Measures

In this section, we illustrate the relationship between K-means clustering and validation measures. Generally speaking, there are two types of clustering validation techniques [1], [2], [8], [20], [21], [27], [29], [14], [17], which are based on external criteria and internal criteria respectively. The focus of this paper is on the evaluation of external clustering validation measures including Entropy, Purity, and F-measure, which are three commonly used external clustering validation measures for K-means clustering [37], [45]. As external criteria, these measures use external information — class labels in this case.

Entropy measures the purity of the clusters with respect to the given class labels. Thus, if all clusters consist of objects with only a single class label, the entropy is 0. However, as the class labels of objects in a cluster become more varied, the entropy increases.

To compute the entropy of a set of clusters, we first calculate the class distribution of the objects in each cluster, i.e., for each cluster j we compute p_{ij}, the probability that a member of cluster j belongs to class i. 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, as shown in the equation

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

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

Purity. In a similar fashion, we can compute the purity of a set of clusters. First, we calculate the purity in each cluster. For each cluster j, we have the purity $P_j = \frac{1}{n_j} \max_i(n_j^i)$, where n_j^i is the number of objects in cluster j with class label i. In other words, P_j is the fraction of the overall cluster size that the largest class of objects assigned to that cluster represents. The overall purity of the clustering solution is obtained as a weighted sum of the individual cluster purities and is given as $Purity = \sum_{j=1}^{m} \frac{n_j}{n} P_j$, where n_j is the size of cluster j, m is the number of clusters, and n is the total number of data points. In general, we believe that the larger the values of purity, the better the clustering solution is.

F-measure combines the precision and recall concepts from information retrieval [36]. We treat each cluster as if it were the result of a query and each class as if it were the desired set of documents for a query. We then calculate the recall and precision of that cluster for each given class as

$$Recall(i,j) = \frac{n_{ij}}{n_i} \qquad and \qquad Precision(i,j) = \frac{n_{ij}}{n_j}$$

where n_{ij} is the number of objects of class i that are in cluster j, n_j is the number of objects in cluster j, and n_i is the number of objects in class i. The F-measure of cluster j and class i is then given by the following equation

$$F(i,j) = \frac{2Recall(i,j)Precision(i,j)}{Precision(i,j) + Recall(i,j)}$$

For an entire hierarchical clustering, the F-measure of any class is the maximum value it attains at any node (cluster) in the tree, and an overall value for the F-measure is computed by taking the weighted

average F-measures for each class, as given by the equation

$$F = \sum_i \frac{n_i}{n} max\{F(i,j)\}$$

where the max is taken over all clusters at all levels, and n is the number of documents. The F-measure values are in the interval [0,1] and larger F-measure value indicates higher clustering quality.

A. The Dispersion in A Data Distribution

Before we describe the relationship between clustering validation measures and K-means clustering, we first introduce the Coefficient of Variation (CV) [9], which is a measure of the data dispersion. The CV is defined as the ratio of the standard deviation to the mean. Given a set of data objects $X = \{x_1, x_2, \ldots, x_n\}$, we have CV $= \frac{s}{\bar{x}}$, where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$.

Note that there are some other statistics, such as standard deviation and skewness [9], which can also be used to characterize the dispersion of a data distribution. However, the standard deviation has no scalability; that is, the dispersion degrees of the original data and the proportionally stratified sample data are not equal if the standard deviation is used. Indeed, this does not agree with our intuition. Meanwhile, skewness cannot catch the dispersion in the situation that the data is symmetric but has high variance. In contrast, the CV is a dimensionless number that allows comparison of the variation of populations that have significantly different mean values. In general, the larger the CV value is, the greater the variability is in the data.

As shown in the previous subsection, K-means tends to produce clusters with relatively uniform sizes. Therefore, in this paper, we establish a necessary but not sufficient criterion for selecting the right cluster validation measures for K-means as follows.

Necessary Criterion 1: If an external cluster validation measure cannot capture the uniform effect by K-means clustering on data sets with large variation in "true" cluster sizes, this measure is not suitable for validating the results of K-means clustering.

This necessary criterion indicates that, if the CV values of cluster sizes have a significant change after the clustering process, we know that the clustering performance is poor. However, it does not necessarily indicate a good clustering performance if the CV values of cluster sizes only have a minor change after the clustering process.

TABLE I
A SAMPLE DOCUMENT DATA SET.

A Sample Document Data Set
Sports, Sports
Entertainment, Entertainment
Foreign, Foreign, Foreign, Foreign, Foreign
Metro, Metro, Metro, Metro, Metro, Metro, Metro, Metro, Metro, Metro
Politics
CV=1.1187

TABLE II
TWO CLUSTERING RESULTS.

Document Clustering		
Clustering I	1: Sports Sports Sports Sports Sports Sports Sports Sports	CV=0.4213 Purity=0.929 Entropy=0.247 F-measure=0.64
	2: Sports Sports Sports Sports Sports Sports Sports Sports	
	3: Sports Sports Sports Sports Sports Sports Sports Sports	
	4: Metro Metro Metro Metro Metro Metro Metro Metro Metro Metro	
	5: Entertainment Entertainment Foreign Foreign Foreign Foreign Foreign Politics	
Clustering II	1: Sports Foreign	CV=1.2011 Purity=0.952 Entropy=0.259 F-measure = 0.947
	2: Entertainment Entertainment	
	3: Foreign Foreign Foreign	
	4: Metro Metro Metro Metro Metro Metro Metro Metro Metro Metro Foreign	
	5: Politics	

B. The Limitations of the Entropy Measure for Clustering Validation

In our practice, we have observed that entropy 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 created the sample data set as shown in Table I. This data set consists of 42 documents with five class labels. In other words, there are five "true" clusters in this sample data set. The CV value of the cluster sizes of these five "true" clusters is 1.1187.

For this sample document data set, we assume that we have two clustering results by different clustering algorithms as shown in Table II. 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, which is 0.4213. In contrast, for the second clustering result, the CV value of the cluster sizes is 1.2011. This indicates that the five clusters have widely different cluster sizes for the second clustering scheme. Certainly, according to the entropy measure, clustering result I is better than clustering result II (this result is due to the fact that the entropy measure more heavily penalizes a large impure cluster). However, if we look at five "true" clusters carefully, we find that the second clustering result is much closer to the "true" cluster distribution and the first clustering result is actually away from the "true" cluster distribution. This is also reflected by the CV values. The CV value (1.2011) of five cluster sizes in the second clustering result is closer to the CV value (1.1187) of five "true" cluster sizes.

Finally, in Table II, we can also observe that the purity of the second clustering result is better than that of the first clustering result. Indeed, this contradicts to the result from the entropy measure. In summary, this example illustrates that the entropy measure has the favorite on the algorithms, such as K-means, which produce clusters with relatively uniform sizes. This effect is more significant in the situation that the data has highly dispersed "true" cluster sizes. In other words, if the entropy measure is used for validating K-means clustering, the validation result can be misleading.

C. F-measure for K-means Clustering

F-measure was originally designed for validating the results of hierarchical clustering algorithms [25]. Since then it has been widely used in the clustering literature, most cases for hierarchical clustering [45], [37], yet some for partitional clustering [33] along with the entropy measure. However, further investigation is needed to clarify whether the F-measure is suitable for validating flat clusters. To this end, we provide some analysis of F-measure for K-means clustering.

For the sample data set in Table I, the F-measure values for the results of the clustering schemes I and II are 0.640 and 0.947, respectively. According to F-measure, the clustering result II is much better than the cluster result I. This result contradicts to the validation result from the entropy measure, but is consistent with the result from the CV measure. The reason is that K-means has the tendency to divide

a large and pure cluster into several smaller clusters and the entropy measure does not penalize such a division but F-measure does. For instance, for the sample data set in Table I, K-means may divide the large cluster–"Sports" into three smaller clusters as shown in the cluster result I in Table II. According to the entropy measure, the resulting three smaller clusters are perfect; that is, the entropy value is zero. However, in terms of F-measure, the recall of the cluster "Sports" is small. This has negative impact on the F-measure value of the heavily weighted class, and eventually decreases the overall F-measure value for the entire data set. In other words, F-measure can detect and penalize the uniform effect produced by K-means on data sets with highly dispersed cluster sizes. Therefore, from a data distribution perspective, F-measure is more suitable for K-means clustering than the entropy measure.

IV. Experimental Results

In this section, we present experimental results to show the impact of data distributions on the performance of K-means clustering. Specifically, we first present (1) a brief introduction to the experimental setup, then demonstrate: (2) the effect of the true cluster sizes on the performance of K-means clustering; (3) the effect of K-means clustering on the distributions of the clustering results; (4) the validation performance of the entropy measure on the results of K-means clustering; (5) the problem with K-means clustering and the entropy measure; and (6) the validation performance of the F-measure on the results of K-means clustering.

TABLE III
Some Notations.

CV_0: the CV value on the cluster sizes of the "true" clusters
CV_1: the CV value on the cluster sizes of the clustering results
DCV: the difference of CV values before and after K-means clustering

A. The Experimental Setup

The Experimental Tool. In our experiments, we used the CLUTO implementation of K-means [22]. Also, since the Euclidean notion of proximity is not very effective for K-means clustering on real-world high-dimensional data sets, such as gene expression data sets and document data sets, for all the experiments

TABLE IV
SOME CHARACTERISTICS OF EXPERIMENTAL DATA SETS.

Data set	Source	# of objects	# of features	# of classes	Min class size	Max class size	CV_0
Document Data Sets							
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
Biomedical Data Sets							
LungCancer	KRBDSR	203	12600	5	6	139	1.363
Leukemia	KRBDSR	325	12558	7	15	79	0.584
UCI Data Sets							
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

in this paper, the cosine similarity is used in the objective function for K-means. Finally, please note that some notations used in our experiments are shown in Table III.

The Experimental Data Sets. 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 IV. In the table, CV_0 shows the CV values of "true" cluster sizes and "# of classes" indicates the number of "true" clusters.

Document Data Sets. The `fbis` data set was from the Foreign Broadcast Information Service data of the TREC-5 collection [40]. 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 set contains documents about computers, electronics, health, medical, research, and technology; and the `sports` data set contains documents about baseball, basket-ball, bicycling, boxing, football, golfing, and hockey. Data sets `tr23` and `tr45` were derived from the TREC-5[40], TREC-6 [40], and TREC-7 [40] collections. The `la2` data set is part of the TREC-5 collection [40] and contains news articles from the Los Angeles Times. The `ohscal` data set was obtained from the OHSUMED collection [16], 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 [26]. The data sets `k1a` and `k1b` contain

exactly the same set of documents but they differ in how the documents were assigned to different classes. In particular, `k1a` contains a finer-grain categorization than that contained in `k1b`. The data set `wap` was from the WebACE project (WAP) [15]; each document corresponds to a web page listed in the subject hierarchy of Yahoo!. For all document clustering data sets, we used a stop-list to remove common words, and the words were stemmed using Porter's suffix-stripping algorithm [34].

Biological Data Sets. `LungCancer` [4] and `Leukemia` [42] data sets were from Kent Ridge Biomedical Data Set Repository (KRBDSR) which is an online repository of high-dimensional features [28]. 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 6 subtypes of pediatric acute lymphoblastic leukemia samples and 1 group samples that do not fit in any of the above 6 subtypes, and each is described by 12558 genes.

UCI Data Sets. In addition to the above high-dimensional data sets, we also used some UCI data sets with small dimension sizes [32]. 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 has been 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 $0-9$, and the `letter` data set includes the letter information of $A-Z$.

Note that for each data set in Table IV, to void the randomness, all experiments were conducted 10 times and the averaged values are presented in the paper.

B. The Effect of the "True" Cluster Sizes on K-means

Here, we illustrate the effect of the "true" cluster sizes on the results of K-means clustering. In our experiment, we first used K-means to cluster the input data sets, and then computed the CV values for the "true" cluster distribution of the original data and the cluster distribution of the clustering results. The number of clusters k was set as the "true" cluster number for the purpose of comparison.

TABLE V
EXPERIMENTAL RESULTS ON REAL-WORLD DATA SETS.

		Standard Deviation of Sizes		Coefficient of Variation of Sizes				
Data set	Average of Sizes	STD_0	STD_1	CV_0	CV_1	DCV=CV_0-CV_1	Entropy	F-measure
fbis	145	139	80	0.96	0.55	0.41	0.345	0.565
hitech	384	190	140	0.50	0.37	0.13	0.630	0.546
k1a	117	117	57	1.00	0.49	0.51	0.342	0.559
k1b	390	513	254	1.32	0.65	0.66	0.153	0.705
la2	513	264	193	0.52	0.38	0.14	0.401	0.689
ohscal	1116	297	489	0.27	0.44	-0.17	0.558	0.562
re0	116	174	45	1.50	0.39	1.11	0.374	0.388
re1	66	92	22	1.39	0.32	1.06	0.302	0.454
sports	1226	1253	516	1.02	0.42	0.60	0.190	0.742
tr23	34	32	14	0.93	0.42	0.51	0.418	0.545
tr45	69	46	30	0.67	0.44	0.23	0.329	0.663
wap	78	81	39	1.04	0.49	0.55	0.313	0.541
LungCancer	41	55	26	1.36	0.63	0.73	0.332	0.621
Leukemia	46	27	17	0.58	0.37	0.21	0.511	0.565
ecoli	42	49	21	1.16	0.50	0.66	0.326	0.581
page-blocks	1095	2138	1029	1.95	0.94	1.01	0.146	0.685
letter	769	23	440	0.03	0.57	-0.54	0.683	0.282
pendigits	1099	46	628	0.04	0.57	-0.53	0.394	0.668
Min	34	23	14	0.03	0.33	-0.54	0.146	0.282
Max	1226	2138	1029	1.95	0.94	1.11	0.683	0.742
Parameters used in CLUTO: -clmethod=rb -sim=cos -crfun=i2 -niter=30								

Table V shows the experimental results on various real-world data sets. As can be seen, for the data sets with large CV_0, K-means tends to reduce the variation on the cluster sizes of the clustering results as indicated by CV_1. This result indicates that, for data sets with high variation on the cluster sizes of "true" clusters, the "uniform effect" is dominant in the objective function. In other words, K-means tends to reduce the variation on the cluster sizes in the clustering results. Indeed, if we look at Equation (12) in Section II, this result shows that the factor $\| m_i - m_j \|^2$ is dominated by the factor $n_i n_j$ for this case.

Also, for data sets with low CV_0 values, K-means increases the variation on the cluster sizes of the clustering results slightly as indicated by the corresponding CV_1 values. This result indicates that, for data sets with very low variation on the cluster sizes of "true" clusters, the "uniform effect" is not significant. Indeed, for Equation (12) in Section II, this result indicates that the factor $n_i n_j$ is dominated by the factor $\| m_i - m_j \|^2$.

C. The Effect of K-means Clustering on the Distribution of the Clustering Results

In the previous subsection, we showed that K-means tends to reduce the variation on the cluster sizes if the CV_0 is high and increase the variation on the cluster sizes if the CV_0 is very low. In this experiment, we want to get a better understanding about the effect of K-means clustering on the distribution of the

clustering results.

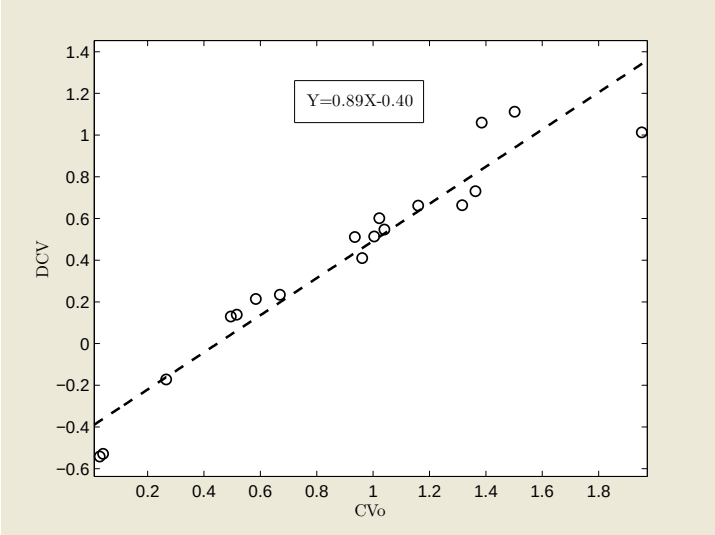


Fig. 2. An Illustration of the Change of CV Values after K-means Clustering.

Figure 2 shows the relationship between DCV and CV_0. In the figure, there is a linear regression fitting line ($y = 0.89x - 0.40$) for all the points (CV_0, DCV). As can be seen, as the increase of CV_0 values, DCV values increase accordingly. For the linear regression fitting line, if $x = 0.45$, then $y = 0$. This indicates that if CV_0 is higher than 0.45, K-means clustering tends to reduce the CV_1 values. Otherwise, if CV_0 is less than 0.45, K-means clustering tends to increase the CV_1 values. In other words, 0.45 is the statistical, empirical threshold of CV_0 value which determines the dispersion degree of clustering results is higher or lower than the original one.

Indeed, Figure 3 shows the relationship between CV_0 and CV_1 for all the experimental data sets listed in Table IV and 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 smaller range from 0.33 to 0.94. We empirically have the value interval of CV_1: [0.3, 1].

D. The Effect of the Entropy Measure on the Results of K-means Clustering

In this subsection, we present the effect of the entropy measure on the K-means clustering results. Figure 4 shows the plot of entropy values for all the experimental data sets in Table IV. A general trend can be observed is that while the differences in CV values before and after clustering increase as the

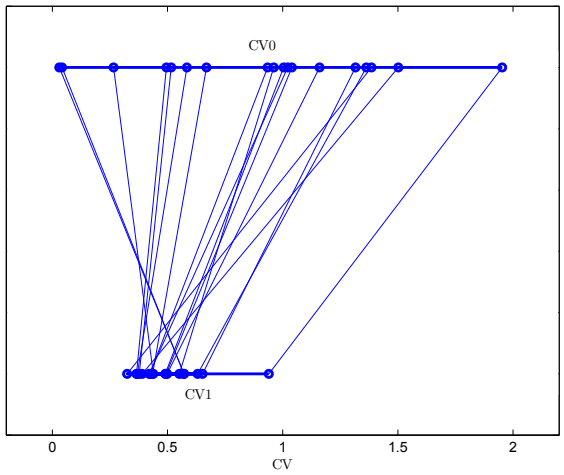


Fig. 3. The Relationships between CV Values before and after K-means Clustering.

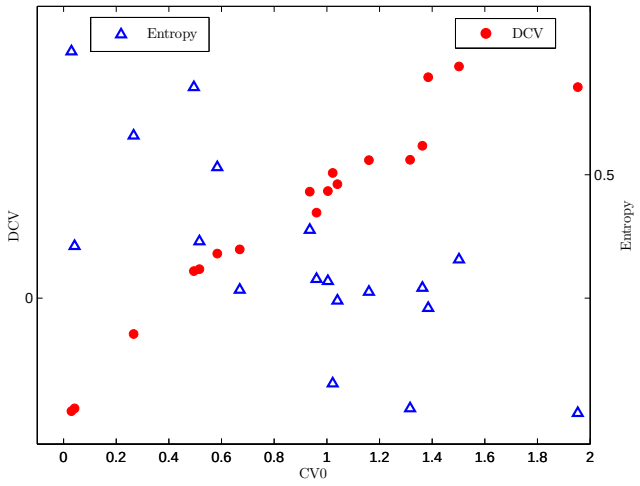


Fig. 4. An Illustration of the "Biased Effect" of the Entropy Measure.

increase of CV_0 values, the entropy values tend to decrease. In other words, there is a disagreement between DCV and the entropy measure on evaluating the clustering quality. Entropy indicates better quality, but DCV shows that the distributions of clustering results are away from the distributions of "true" clusters. This indicates worse clustering quality. The above observation agrees with our analysis in Section III that entropy has a biased effect on K-means.

To strengthen the above observation, we also generated two groups of synthetic data sets from two real-world data sets: `pendigits` and `letter`. These synthetic data sets have wide dispersion degree on their "true" cluster sizes. The first group of synthetic data sets was derived from the `pendigits` data

TABLE VI

EXPERIMENTAL RESULTS ON SAMPLE DATA SETS FROM THE "PENDIGITS" DATA SET.

Data set	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
CV_0	0.00	0.11	0.21	0.30	0.41	0.59	0.82	1.05	1.30	1.50	1.69	1.93	2.15	2.42
DCV	-0.46	-0.27	-0.12	-0.12	-0.09	0.06	0.24	0.48	0.63	0.69	0.78	0.97	1.16	1.47
Entropy	0.373	0.380	0.385	0.391	0.393	0.387	0.375	0.361	0.332	0.309	0.287	0.272	0.239	0.192

TABLE VII

EXPERIMENTAL RESULTS ON SAMPLE DATA SETS FROM THE "LETTER" DATA SET.

Data set	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
CV_0	0.09	0.32	0.54	0.75	1.01	1.30	1.60	1.81	2.02	2.18	2.35
DCV	-0.43	-0.22	0.03	0.30	0.59	0.83	1.10	1.28	1.47	1.60	1.77
Entropy	0.667	0.661	0.647	0.619	0.590	0.551	0.515	0.489	0.463	0.446	0.425

set as shown in Table IV. We applied the following sampling strategy: 1) We first sampled the original data set to get a sample with 10 classes, and the number of objects for each class is $\{1000, 100, 100, 100, 100, 100, 100, 100, 100, 100\}$, respectively. Then based on this sample, 2) we did random sampling on the class with 1000 objects and merged the sampling objects with all the other objects in the rest 9 classes to form an experimental data set. We gradually reduced the sample size of the first cluster to 100, thus obtained various data sets with decreasing dispersion degree. On the other hand, to get data sets whose "true" class distributions with higher dispersion degrees, 3) we did random stratified sampling to the 9 classes with 100 objects each, and merged the sampling objects with the rest 1000 objects to form an experimental data set. We gradually reduced the sample size for each 9 classes to 30, and thus got a series of data sets with increasing dispersion degree. A similar sampling strategy was also applied to the `letter` data set for generating the second group of synthetic data sets. Note that for each dispersion degree we did sampling 10 times, i.e., there are 10 data sets for each dispersion degree, and output the average values as the clustering results.

Table VI and Table VII show the entropy values of the results of K-means clustering on these two groups of synthetic data sets respectively. Also, Figure 5 and Figure 6 show the corresponding plots of the entropy values for these two groups of synthetic data sets respectively. A similar trend has been observed; that is, the entropy values decrease as the increase of CV_0 values.

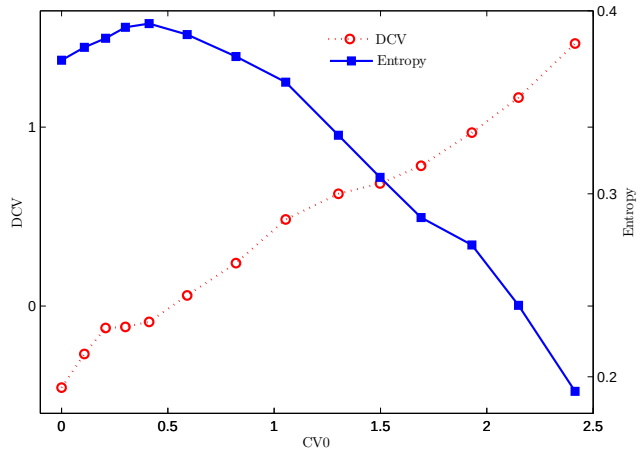


Fig. 5. An Illustration of the "Biased Effect" of Entropy Using Sample Data Sets from the `Pendigits` data set.

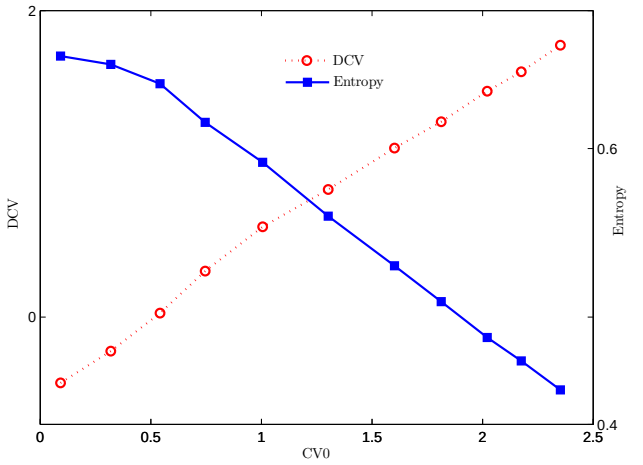


Fig. 6. An Illustration of the "Biased Effect" of Entropy Using Sample Data Sets from the `Letter` data set.

E. The Problem with K-means Clustering and the Entropy Measure

In our experiments, we found one major problem with K-means for the data sets which have high variation on the cluster sizes of "true" clusters. To illustrate this, we selected five data sets with high CV_0 values including the `re0`, `re1`, `wap`, `ecoli`, and `k1a` data sets. We did K-means clustering on these five data sets using the number of "true" clusters as the k for K-means. In the clustering results, we labelled each cluster by the label of the majority objects in the cluster. We found that many "true" clusters were disappeared in the clustering results. Figure 7 shows the percentage of the disappeared "true" clusters in the K-means clustering results for these five data sets. As can be seen, every data set has a significant

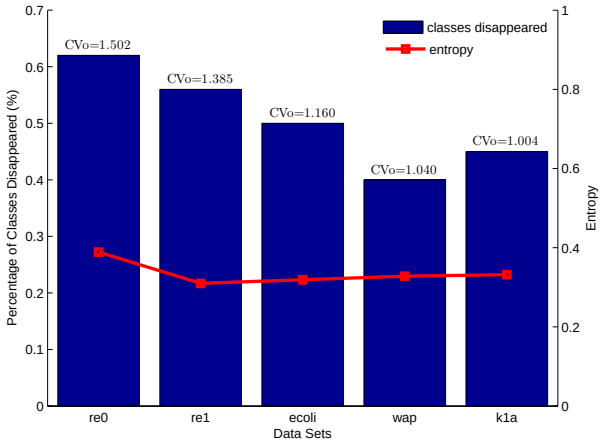


Fig. 7. The Percentage of the Disappeared "True" Clusters in Data Sets with Varied Cluster Sizes.

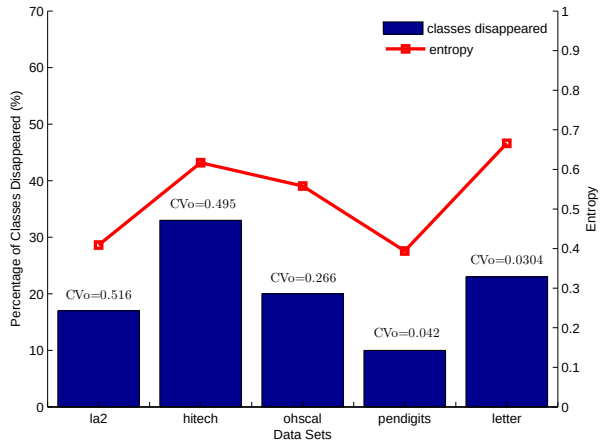


Fig. 8. The Percentage of the Disappeared "True" Clusters for Data Sets with Low Dispersion.

number of "true" clusters disappeared. For instance, for the re0 data set ($CV_0 = 1.502$), more than 60% true clusters have disappeared after K-means clustering.

In addition, in Table V and Figure 7, we can observe that very low entropy values were achieved for these data sets with high CV_0 values. In other words, if the entropy measure is used as the clustering validation measure, the K-means clustering results on these five data sets should be excellent. However, as demonstrated above, the clustering results on these five data sets are actually far away from the 'true" cluster distributions. In summary, this result indicates that (1) K-means may not perform well for data sets with high variation on the cluster sizes of "true" clusters; (2) the entropy measure is not an algorithm-independent clustering validation measure and favorites the K-means clustering.

Finally, for the purpose of comparison, we conducted a similar experiment on five data sets with low CV_0 values. Figure 8 shows the percentage of the disappeared "true" clusters. An interesting observation is that, compared to the results on data sets with high CV_0 values, the percentages of the disappeared "true" clusters became much smaller and the entropy values increased. In other words, the entropy measure on the data sets with relatively uniform "true" cluster sizes is more reliable.

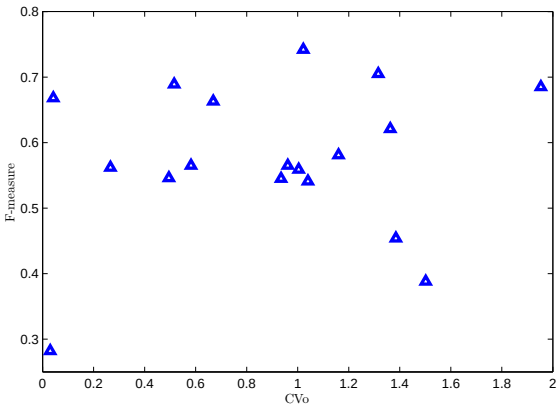
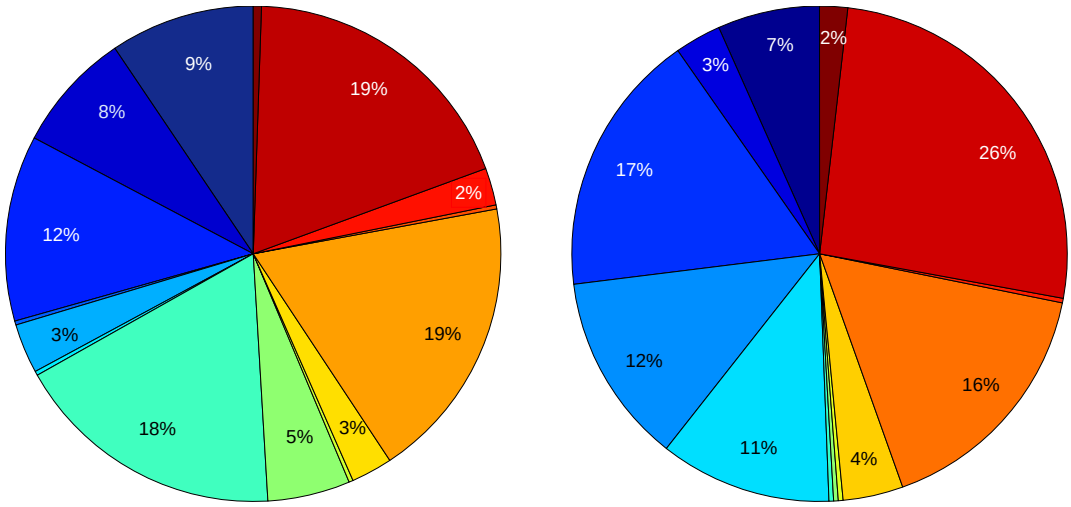


Fig. 9. The F-measure Values for the Clustering Results on All the Experimental Data Sets.



(a) class "grai" (b) class "crud"

Fig. 10. The Partition Behaviors of K-means on two "true" clusters of `re1`.

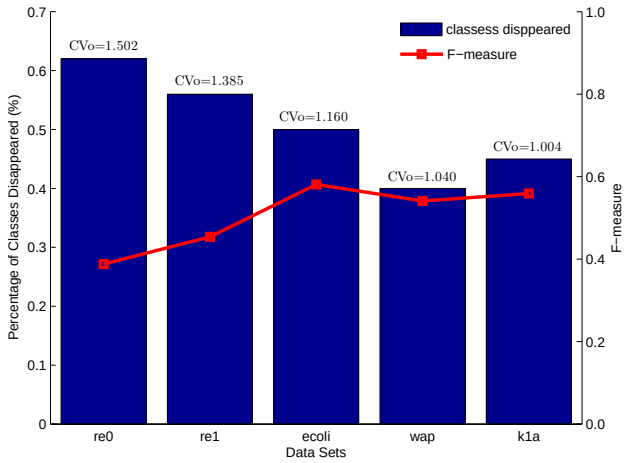


Fig. 11. The Percentage of the Disappeared "True" Clusters for Data Sets with Varied Cluster Sizes.

F. The Validation Performances of the F-measure on the Results of K-means Clustering

Figure 9 shows the F-measure values for the clustering results by K-means on all the experimental data sets. In the figure, we can see that the F-measure values do not show a strong correlation with the CV_0 values. Another observation is that the entropy value for the `re1` data set is 0.310 which ranks 4th (the smaller the better) among the data sets listed in Table IV. However, the corresponding F-measure value for this data set is 0.454 with the rank of 14th (the larger the better). To have a better understanding on this inconsistent validation result, we have carefully looked at the results of k-means clustering on the `re1` data set. Indeed, we noticed that some large "true" clusters were divided into several smaller pieces by K-means. As an example, Figure 10 shows that two largest "true" clusters "grai" and "crud" of `re1` have been divided into 15 and 13 pieces, respectively. 7 out of 15 pieces of "grai" are the dominant classes in their corresponding clusters. Also, 6 out of 13 pieces of "crud" are the dominant classes in their corresponding clusters. This is why the percentage of the disappeared "true" clusters of `re1` is 0.56 (there are 25 "true" clusters in the `re1` data set). As we discussed above, the entropy measure cannot capture this scenario and provides misleading information on the clustering performance. In contrast, the F-measure penalizes the small recalls of "grai" and "crud", and thus shows a small value which indicates a poor clustering performance.

Finally, Figure 11 shows the F-measure values as well as the percentage of the disappeared "true" clusters in the clustering results of K-means on five data sets including `re0`, `re1`, `wap`, `ecoli`, and `k1a`. As can be seen, if there is a significant number of clusters disappeared, the F-measure value is low. In other words, the F-measure can provide a somewhat consistent indication about the loss of "true" clusters caused by K-means clustering.

V. 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 paper.

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 sets, such as gene expression data sets and document data sets. To meet this challenge, one research direction is to make use of dimensionality reduction techniques, such as Multidimensional Scaling (MDS) [5], Principal Components Analysis (PCA) [19], and Singular Value Decomposition (SVD) [10]. Also, several feature transformation techniques have been proposed for high-dimensional document data sets, 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 has been provided by Tang et al. [39]. 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 [18]. Finally, some other similarity measures, e.g., the cosine measure, have also shown appealing effects on clustering document data sets [45].

Second, it has been recognized that K-means has difficulty in detecting the "natural" clusters with non-spherical shapes [38], [17]. To address this issue, one research direction is to modify the K-means clustering algorithm. For instance, Guha et al. [13] 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 on data sets with various shapes than the K-means clustering method [38].

Third, outliers and noise in the data can also degrade the performance of clustering algorithms [23], [43], [46], especially for prototype-based algorithms such as K-means. To deal with this problem, one research direction is to incorporate some outlier removal techniques before conducting K-means clustering. For instance, a simple method [23] of detecting outliers is based on the distance measure. Breunig et al. [7] proposed a density based method using the Local Outlier Factor (LOF) for the purpose of identifying outliers in data sets with varying densities. There are also some other clustering based methods to detect outliers as small and remote clusters [35], or objects that are farthest from their corresponding cluster centroids [25]. Another research direction is to handle outliers during the clustering process. There has been serval techniques designed for such purpose. For example, DBSCAN automatically classifies low-density points as noise points and removes them from the clustering process [12]. Also, SNN density-based clustering [11] and CURE [13] explicitly deal with noise and outliers during the clustering process.

Fourth, many clustering algorithms that work well for small or medium-size data sets are unable to handle large data sets. Along this line, a discussion of scaling K-means clustering to large data sets was provided by Bradley et al. [6]. A broader discussion of specific clustering techniques can be found in [31]. For instance, some representative techniques include CURE [13] and BIRCH [44], etc.

Finally, some researchers have identified some other factors, such as the types of attributes, the types of data sets, and scales of attributes, which may have impact on the performance of K-means clustering. However, in this paper, we focused on understanding the impact of the distributions of "true" cluster sizes on the performance of K-means clustering. Also, we investigate the relationship between K-means and some cluster validation measures [3], [2], [20], [27], [29], [14], [17].

VI. Conclusions and Future Work

In this paper, we presented an organized study of K-means and cluster validation measures from a data distribution perspective. Specifically, our major focus is to characterize the relationships between data

distributions and K-means clustering as well as the entropy measure and F-measure. Along this line, we first theoretically illustrated the relationship between the objective function of K-means and cluster sizes. We also conducted various experiments on a number of real-world data sets. Our experimental results show that K-means tends to reduce variation in cluster sizes if the variation of the "true" cluster sizes is high and increase variation in cluster sizes if the variation of the "true" cluster sizes is very low.

Also, we observed that, no matter what the CV values of "true" cluster sizes are, the CV values of resultant cluster sizes are typically located in a much narrow range of about 0.3 to 1.0. In addition, we found that many "true" clusters were disappeared in the clustering results if K-means was applied for data sets with large variation in "true" cluster sizes; that is, K-means produces the clustering results which are far away from the "true" cluster distributions. However, when the entropy measure was used for cluster validation, it could not capture this uniform effect and provided misleading information about the clustering performance. This motivates the need for using the CV measure as a necessary criterion to validate the clustering results.

There are several potential directions for future research. First, we would like to investigate what measures best reflect the performance of K-means clustering. Second, we plan to improve K-means clustering for better handling data sets with large variation in "true" cluster sizes.

VII. Acknowledgments

This research was partially supported by the National Science Foundation of China (NSFC) (No. 70621061, 70321010) and the Rutgers Seed Funding for Collaborative Computing Research. Also, this research was supported in part by a Faculty Research Grant from Rutgers Business School- Newark and New Brunswick. Finally, we are grateful to the KDD and SMCB anonymous referees for their constructive comments on the paper.

References

[1] Daniel Barbará and Ping Chen. Using self-similarity to cluster large data sets. *Data Mining and Knowledge Discovery*, 7(2):123–152, 2003.

[2] Daniel Barbará, Yi Li, and Julia Couto. Coolcat: an entropy-based algorithm for categorical clustering. In *Proceedings of the 11th ACM International Conference on Information and Knowledge Management*, pages 582–589, 2002.

[3] J.C. Bezdek and N.R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics — Part B*, 28(3):427–436, 1998.

[4] Arindam Bhattacharjee and et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13790–13795, November 2001.

[5] I. Borg and P. Groenen. *Modern Multidimensional Scaling – Theory and Applications*. Springer Verlag, Febuary 1997.

[6] P.S. Bradley, U.M. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In *Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 9–15, August 1998.

[7] M.M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander. Lof: Identifing density based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 427–438, 2000.

[8] Yixin Chen, Ya Zhang, and Xiang Ji. Size regularized cut for data clustering. In *Neural Information Processing Systems (NIPS)*, 2005.

[9] M. DeGroot and M. Schervish. *Probability and Statistics (3 edition)*. Addison Wesley, 2001.

[10] J.W. Demmel. *Applied Numerical Linear Algebra*. Society for Industrial & Applied Mathematics, September, 1997.

[11] L. Ertoz, M. Steinbach, and V. Kumar. A new shared nearest neighbor clustering algorithm and its applications. In *Workshop on Clustering High Dimensional Data and its Applications, the 2nd SIAM International Conference on Data Mining*, 2002.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. 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*, pages 226–231, August 1996.

[13] S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pages 73–84, June 1998.

[14] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: Part i. *SIGMOD Record*, 31(2):40–45, 2002.

[15] Eui-Hong Han, Daniel Boley, Maria Gini, Robert Gross, Kyle Hastings, George Karypis, Vipin Kumar, B. Mobasher, and Jerry Moore. Webace: A web agent for document categorization and exploration. In *Proceedings of the 2nd International Conference on Autonomous Agents*, 1998.

[16] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. 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*, pages 192–201, July 1994.

[17] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1998.

[18] R.A. Jarvis and E.A. Patrick. Clusering using a similarity measure based on shared nearest neighbors. *IEEE Transactions on Computers*, C-22(11):1025–1034, 1973.

[19] I.T. Jolliffe. *Principal Component Analysis (2nd edition)*. Springer Verlag, October 2002.

[20] Istvan Jonyer, Diane J. Cook, and Lawrence B. Holder. Graph-based hierarchical conceptual clustering. *Journal of Machine Learning*

Research, 2:19–43, 2001.

[21] Istvan Jonyer, Lawrence B. Holder, and Diane J. Cook. Graph-based hierarchical conceptual clustering. *Journal of Machine Learning Research*, 2:19–43, 2001.

[22] George Karypis. Cluto – software for clustering high-dimensional datasets, version 2.1.1. In *http://glaros.dtc.umn.edu/gkhome/views/cluto*, 2006.

[23] E.M. Knorr, R.T. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. *VLDB Journal*, 8:237–253, 2000.

[24] K. Krishna and M. Narasimha Murty. Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics — Part B*, 29(3):433–439, 1999.

[25] B. Larsen and C. Aone. 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*, pages 16–22, August 1999.

[26] D. Lewis. Reuters-21578 text categorization text collection 1.0. In *http://www.research.att.com/ lewis*, 2004.

[27] Jing Li, Dacheng Tao, Weiming Hu, and Xuelong Li. Kernel principle component analysis in pixels clustering. In *Web Intelligence*, pages 786–789, 2005.

[28] Jinyan Li and Huiqing Liu. Kent ridge biomedical data set repository. In *http://sdmc.i2r.a-star.edu.sg/rp/*.

[29] Wenyuan Li, Wee Keong Ng, Ying Liu, and Kok-Leong Ong. Enhancing the effectiveness of clustering with spectra analysis. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):887–902, 2007.

[30] J. MacQueen. Some methods for classification and analysis of multivariate observations. In Lucien M. Le Cam and Jerzy Neyman, editors, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume I, Statistics*. University of California Press, 1967.

[31] F. Murtagh. *Clustering Massive Data Sets, Handbook of Massive Data Sets*. Kluwer, 2000.

[32] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. Uci repository of machine learning databases, 1998.

[33] A. Ozgur and E. Alpaydm. Unsupervised machine learning techniques for text document clustering. In *www.cmpe.boun.edu.tr/pilab/MLmaterial/ozgur04unsupervised.pdf*, April 2004.

[34] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.

[35] L. Portnoy, E. Eskin, and S.J. Stolfo. Intrusion detection with unlabeled data using clustering. In *Proceedings of the 2001 ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001)*, 2001.

[36] C. J. Van Rijsbergen. *Information Retrieval (2nd Edition)*. Butterworths, London, 1979.

[37] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In *Workshop on Text Mining, the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2000.

[38] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[39] Bin Tang, Michael Shepherd, Malcolm I. Heywood, and Xiao Luo. Comparing dimension reduction techniques for document clustering. In *Canadian Conference on AI*, pages 292–296, 2005.

[40] TREC. Text retrieval conference. In *http://trec.nist.gov*, 1996.

[41] Hui Xiong, Junjie Wu, and Jian Chen. K-means clustering versus validation measures: a data distribution perspective. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 779–784, 2006.

[42] Eng-Juh Yeoh and et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, March 2002.

[43] J.-S. Zhang and Y.-W. Leung. Robust clustering by pruning outliers. *IEEE Transactions on Systems, Man, and Cybernetics — Part B*, 33(6):983–998, 2003.

[44] T. Zhang, R. Ramakrishnan, and M.Livny. Birch: An efficient data clustering mehtod for very large databases. In *Proceedings of 1996 ACM SIGMOD International Conference on Management of Data*, pages 103–114, June 1996.

[45] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. *Machine Learning*, 55(3):311–331, June 2004.

[46] Aoying Zhou, Feng Cao, Ying Yan, Chaofeng Sha, and Xiaofeng He. Distributed data stream clustering: A fast em-based approach. In *Proceedings of the 23rd International Conference on Data Engineering*, pages 736–745, 2007.

Hui Xiong is currently an Assistant Professor in the Management Science and Information Systems department at Rutgers, the State University of New Jersey. He received the B.E. degree in Automation from the University of Science and Technology of China, China, the M.S. degree in Computer Science from the National University of Singapore, Singapore, and the Ph.D. degree in Computer Science from the University of Minnesota, USA. His general area of research is data and knowledge engineering, with a focus on developing effective and efficient data analysis techniques for emerging data intensive applications. He has published over 50 technical papers in peer-reviewed journals and conference proceedings. He is a co-editor of Clustering and Information Retrieval (Kluwer Academic Publishers, 2003) and a co-Editor-in-Chief of Encyclopedia of GIS (Springer, 2008). He is an associate editor of the Knowledge and Information Systems journal and has served regularly in the organization committees and the program committees of a number of international conferences and workshops. He was the recipient of the 2007 Junior Faculty Teaching Excellence Award and the 2008 Junior Faculty Research Award at the Rutgers Business School. He is a senior member of the IEEE, and a member of the ACM, the ACM SIGKDD, and Sigma Xi.

Junjie Wu received his Ph.D. in Management Science and Engineering from Tsinghua University, China. He also holds a B.E. degree in Civil Engineering from the same university. He is currently an Assistant Professor in Information Systems Department, School of Economics and Management, Beihang University, China. His general area of research is data mining and statistical modeling, with a special interest on solving the problems raised from the real-world business applications. He has published 3 papers in KDD and 1 paper in ICDM. He is the co-chair of ”Data Mining in Business”, a special track in AMIGE 2008. He has also been a reviewer for the leading academic journals and many international conferences in his area. He is the recipient of the Outstanding Yound Research Award at School of Economics and Management, Tsinghua University. He is a member of AIS.

Jian Chen (M’95-SM’96-F’08) received the B.Sc. degree in Electrical Engineering from Tsinghua University, Beijing, China, in 1983, and the M.Sc. and the Ph.D. degree both in Systems Engineering from the same University in 1986 and 1989, respectively. He is Professor and Chairman of Management Science Department, Director of Research Center for Contemporary Management, Tsinghua University. His main research interests include supply chain management, E-commerce, decision support systems, modeling and control of complex systems. Dr. Chen has published over 100 papers in refereed journals and has been a principal investigator for over 30 grants or research contracts with National Science Foundation of China, governmental organizations and companies. He has been invited to present several plenary lectures. He is the recipient of Ministry of Education Changjiang Scholars, Fudan Management Excellence Award(3rd), Science and Technology Progress Awards of Beijing Municipal Government; the Outstanding Contribution Award of IEEE Systems, Man and Cybernetics Society; Science and Technology Progress Award of the State Educational Commission; Science & Technology Award for Chinese Youth. He has also been elected to IEEE Fellow. He serves as Chairman of the Service Systems and Organizations Technical Committee of IEEE Systems, Man and Cybernetics Society, Vice President of Systems Engineering Society of China, Vice President of China Society for Optimization and Overall Planning, a member of the Standing Committee of China Information Industry Association. He is the editor of “the Journal of Systems Science and Systems Engineering”, an area editor of “Electronic Commerce Research and Applications”, an associate editor of ”IEEE Transactions on Systems, Man and Cybernetics: Part A”, “IEEE Transactions on Systems, Man and Cybernetics: Part C”, and “Asia Pacific Journal of Operational Research” and serves on the Editorial Board of “International Journal of Electronic Business”, “International Journal of Information Technology and Decision Making” and “Systems Research and Behavioral Science”.