A Generalization of Proximity Functions for K-means

Junjie Wu[1], Hui Xiong[2], Jian Chen[1], Wenjun Zhou[2]
[1]Research Center for Contemporary Management, Key Research Institute
of Humanities and Social Sciences at Universities, Tsinghua University, China
E-mail: {wujj, chenj}@sem.tsinghua.edu.cn
[2]Management Science and Information Systems Department, Rutgers University, USA
E-mail: hxiong@andromeda.rutgers.edu, wjzhou@pegasus.rutgers.edu

Abstract

K-means is a widely used partitional clustering method. A large amount of effort has been made on finding better proximity (distance) functions for K-means. However, the common characteristics of proximity functions remain unknown. To this end, in this paper, we show that all proximity functions that fit K-means clustering can be generalized as K-means distance, which can be derived by a differentiable convex function. A general proof of sufficient and necessary conditions for K-means distance functions is also provided. In addition, we reveal that K-means has a general uniformization effect; that is, K-means tends to produce clusters with relatively balanced cluster sizes. This uniformization effect of K-means exists regardless of proximity functions. Finally, we have conducted extensive experiments on various real-world data sets, and the results show the evidence of the uniformization effect. Also, we observed that external clustering validation measures, such as Entropy and Variance of Information (VI), have difficulty in measuring clustering quality if data have skewed distributions on class sizes.

1. Introduction

K-means [18] is a widely-used prototype-based clustering algorithm. A key design issue of K-means clustering is the use of proximity functions. Intuitively, one can easily understand that different choices of proximity functions for K-means can lead to quite different clustering results. In the literature, while a large amount of research work has been proposed on finding better proximity (distance) functions which can lead to a quick convergence, the common characteristics of proximity functions that fit K-means clustering (i.e. quickly converge to a solution) remain unknown.

Along this line, in this paper, we present a concept of "K-means distance", which can be used to guide the choices of proximity functions that can fit K-means clustering. Indeed, we show that all proximity functions that fit K-means clustering can be generalized as K-means distance, which can be derived by a differentiable convex function. A general proof of sufficient and necessary conditions for K-means distance functions is also provided.

While K-means clustering can be used for a wide variety of data types, it cannot be used for all the data types. For instance, Xiong et al. [25] revealed K-means has troubles in dealing with the case that the distributions of "true" cluster sizes of the data are skewed. In particular, they showed that, if Euclidean distance is used as the proximity function, K-means tends to produce clusters with relatively balanced cluster sizes (this is also called the uniformization effect of K-means). However, this paper highlights that this uniformization effect exists regardless of proximity functions used in K-means as long as these proximity functions are K-means distances. Specifically, we show that some well-known proximity functions of K-means, such as the cosine similarity, the coefficient of correlation [7], and the Bregman divergence [6], are K-means distances. When these proximity functions are used for K-means clustering, the uniformization effect cannot be avoided for data with skewed class distributions.

In addition, we have conducted extensive experiments on a number of real-world data sets from different application domains. Our experimental results show that K-means tends to produce the clusters in which the variation of the cluster sizes is smaller. This data variation is measured by the Coefficient of Variation (CV) [7]. The CV, described in more detail later, 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.

Finally, as shown in our experimental results, after K-means clustering, the distributions of the resultant cluster sizes are in a much narrower interval compared to the dis-

1550-4786/07 $25.00 © 2007 IEEE
DOI 10.1109/ICDM.2007.59

IEEE computer society

tributions of the "true" cluster sizes. Indeed, the CV values of resultant cluster sizes are normally distributed and the 95% confidence interval is [0.09, 0.85]. The significance of the normal distribution is tested by the χ^2 statistic and the Shapiro-Wilk W statistic [22]. Also, we observed that some external clustering validation measures, such as Entropy and Variance of Information (VI) [20], have difficulty in measuring clustering quality if data have skewed distributions on class sizes. When dealing with data sets with skewed distributions on the class sizes, both Entropy and VI have the favor on clustering algorithms, such as K-means, which tend to reduce high variation on the cluster sizes.

2. The K-means Distance

In this section, we characterize the distance (proximity) functions that fit K-means. We prove that, under certain assumptions, any distance function which can be used for K-means clustering must take some specific expression derived from a differentiable convex function. We call the family of such functions: ***The K-means distance***. For instance, the Bregman divergence as well as the cosine similarity are two different cases of the K-means distance.

K-means [18] is a prototype-based, simple partitional clustering technique which attempts to find the user-specified K clusters. These clusters are represented by their centroids (a cluster centroid is typically the mean of the points in the cluster). K-means has an ***objective function***: $\min \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)$, where C_i denotes cluster i, and c_i is its centroid. The clustering process of K-means is as follows. First, K initial centroids are selected. Then the ***two-phase iterations*** are launched. That is, in the first phase, every point in the data is assigned to the closest centroid, and each collection of points assigned to a centroid forms a cluster; then in the second phase, the centroid of each cluster is updated based on the points assigned to it. This process is repeated until no point changes clusters.

Note that there are several types of centroids can be used, e.g., the mean, the median, or the medroid. Since the ***mean*** has the unique advantage of high computational efficiency among other types of centroids, and can adapt to most of the existed distance functions, we restrict our analysis on the traditional K-means algorithm which takes the mean of the instances in each cluster as the centroid of the cluster.

Definition 1 *We say that a distance function F fits K-means, if the value of the K-means objective function using F can be continuously (not strictly) decreased by the two-phase iterations.*

Apparently, the distance functions that fit K-means must have the ability to facilitate the convergence of the two-phase iterations. Next, we give a lemma as follows.

Lemma 1 *A distance function $F(x, y)$: $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ fits K-means, if and only if $\forall \mathbb{C} = \{x_1, x_2, \cdots, x_n\} \subset \mathbb{R}^d$,*

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \in \{y | \arg\min_{y \in \mathbb{R}^d} \sum_{i=1}^{n} F(x_i, y)\}$$

Proof: If we can prove that by using the distance function and the mean as the centroid, the value of the objective function of K-means can decrease continuously, then the sufficient condition holds.

Let $D(A^k, U^k)$ denote the value of the objective function after k iterations, where A denotes the phase of assigning instances to the nearest centroids, and U denotes the phase of updating the centroids. Then, in the next iteration, $D(A^{k+1}, U^k) \leq D(A^k, U^k)$, for each instance finds its closest centroid after the reassignment. Furthermore, the updated centroid, i.e., the new mean of each cluster, is the minimizer of the sum of the distances in that cluster, which implies that $D(A^{k+1}, U^{k+1}) \leq D(A^{k+1}, U^k)$. Therefore $D(A^{k+1}, U^{k+1}) \leq D(A^k, U^k)$, and the equality holds if and only if there is no re-assignment in the $k+1$ iteration. Thus the sufficient condition holds.

On the contrary, if $\overline{x}$ is not one of the minimizers of $\sum_{i=1}^{n} F(x_i, y)$, then $D(A^{k+1}, U^{k+1})$ may be larger than $D(A^{k+1}, U^k)$. In other words, the update of the centroid for $\{x_1, \cdots, x_n\}$ may inversely increase the value of the objective function. So the necessary condition holds. □

Lemma 2 *A differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if $\forall x, y \in \mathbb{R}^d$, $\phi(x) - \phi(y) - (x-y)^t \nabla \phi(y) \geq 0$. Further, if the equality holds if and only if $x = y$, then ϕ is strictly convex.*

The proof of Lemma 2 can be found on page 70 in [5]. This lemma often serves as the sufficient and necessary conditions for a function being a (strictly) convex function.

Now the question is: How to characterize all the distance functions that fit K-means? To this end, we have the following theorem on one dimensional data.

Theorem 1 *Assume that $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a non-negative function such that: (1) $F(x, x) = 0$, $\forall x \in \mathbb{R}$, (2) F and F_x are continuous, and (3) F_y is continuously differentiable on x, then F fits K-means if and only if there exists some differentiable convex function $\phi : \mathbb{R} \to \mathbb{R}$ such that*

$$F(x, y) = \phi(x) - \phi(y) - (x - y)\phi'(y).$$

Proof: Note that F_x and F_y here denote the partial derivatives of F on x and y, respectively. We first prove the sufficient condition. For any cluster $\{x_1, \cdots, x_n\}$, let $c^* = \sum_{i=1}^{n} x_i / n$, then $\forall y \in \mathbb{R}$

$$\begin{aligned}\Delta &= \sum_{i=1}^{n} F(x_i, y) - \sum_{i=1}^{n} F(x_i, c^*)\\ &= -n\phi(y) - \sum_{i=1}^{n}(x_i - y)\phi\prime(y) + n\phi(c^*) + \sum_{i=1}^{n}(x_i - c^*)\phi\prime(c^*).\end{aligned}$$

Furthermore, since $\sum_{i=1}^{n}(x_i - c^*) = \mathbf{0}$, and $\sum_{i=1}^{n}(x_i - y) = n(c^* - y)$, we have

$$\Delta = n(\phi(c^*) - \phi(y) - (c^* - y)\phi\prime(y)) = nF(c^*, y) \geq 0.$$

Thus $c^* = \sum_{i=1}^{n} x_i/n$ is one of the minimizers of $\sum_{i=1}^{n} F(x_i, y)$, so the sufficient condition follows from Lemma 1.

Then we prove the necessary condition. For each cluster $\{x_1, \cdots, x_n\}$, since c^* is one of the minimizers of $\sum_{i=1}^{n} F(x_i, y)$, we can have

$$\sum_{i=1}^{n} F_y(x_i, c^*) = 0. \tag{1}$$

Without loss of generality, let $x'_1 = x_1 + \delta$, $x'_2 = x_2 - \delta$, $\delta > 0$, we still have $c^* = (x'_1 + x'_2 + x_3 + \cdots + x_n)/n$ which means

$$F_y(x'_1, c^*) + F_y(x'_2, c^*) + \sum_{i=3}^{n} F_y(x_i, c^*) = 0. \tag{2}$$

Subtracting Equation (2) by Equation (1), we have

$$F_y(x_1+\delta, c^*) - F_y(x_1, c^*) = -(F_y(x_2-\delta, c^*) - F_y(x_2, c^*)). \tag{3}$$

Dividing both sides of Equation (3) by δ and let $\delta \to 0$, we have $F_{yx}(x_1, c^*) = F_{yx}(x_2, c^*)$. Similarly we have

$$F_{yx}(x_1, c^*) = \cdots = F_{yx}(x_n, c^*).$$

Therefore it must be $F_{yx}(x, y) = -H(y)$. So

$$F_y(x, y) = -H(y)x + I(y). \tag{4}$$

We know that for $n = 2$, $F_y(x_1, c^*) + F_y(x_2, c^*) = 0$. If we replace $F_y(x, y)$ by Equation (4), then we have $I(c^*) = H(c^*)c^*$, which implies $I(y) = H(y)y$. Therefore, we have

$$F_y(x, y) = (y - x)H(y).$$

Let $\phi'(y) = \int H(y)dy$, then

$$\begin{aligned}F(x, y) &= \int_x^y F_y(x, y)dy = \int_x^y (y - x)H(y)dy\\ &= \int_x^y (y - x)d\phi'(y) = \phi(x) - \phi(y) - (x - y)\phi'(y).\end{aligned}$$

Since $\forall x, y \in \mathbb{R}, F(x, y) \geq 0$, $\phi(\cdot)$ is a convex function, which follows from Lemma 2. Thus the necessary condition holds. So we complete the proof. □

It is valuable to extend Theorem 1 to the multi-dimensional data case.

Theorem 2 *Assume that $F : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a non-negative function such that: (1) $F(x, x) \geq 0$, $\forall x \in \mathbb{R}^d$, (2) F and F_x are continuous, and (3) F_y is continuously differentiable on x, then F fits K-means if and only if there exists some differentiable convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that*

$$F(x, y) = \phi(x) - \phi(y) - (x - y)^t \nabla\phi(y).$$

Due to the page limitation, we have to omit the proof of Theorem 2, which is similar to the proof of Theorem 1. Based on Theorem 2, we can have the precise definition of the K-means distance as follows.

Definition 2 *We say that a distance function F is a K-means distance, if there exists some differentiable convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that*

$$F(x, y) = \phi(x) - \phi(y) - (x - y)^t \nabla\phi(y).$$

According to Theorem 2, the K-means distance fits the K-means clustering. And under certain acceptable assumptions, the K-means distance is the only distance that fits K-means when the centroid type is the mean. Furthermore, please note that the K-means distance is a family of distance functions with different ϕ. Finally, we must point out that a K-means distance is not necessary to be a metric; that is, a K-means distance may not have symmetry and triangle inequality properties.

Theorem 3 *Given a differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$, let $F(x, y) = \phi(x) - \phi(y) - (x - y)^t \nabla\phi(y)$, then ϕ is strictly convex if and only if $\forall \mathbb{C} = \{x_1, x_2, \cdots, x_n\} \subset \mathbb{R}^d$,*

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \arg\min_{y \in \mathbb{R}^d} \sum_{i=1}^{n} F(x_i, y).$$

Proof: For any cluster $\{x_1, \cdots, x_n\} \subset \mathbb{R}^d$, $\forall y \in \mathbb{R}^d$, $\Delta = \sum_{i=1}^{n} F(x_i, y) - \sum_{i=1}^{n} F(x_i, \overline{x}) = nF(\overline{x}, y)$. Therefore, according to Lemma 2, that ϕ *is strictly convex* is equivalent to that $\Delta \geq 0$ *and the equality holds if and only if* $y = \overline{x}$. And the latter equivalent condition implies that $\overline{x}$ is the unique minimizer of $\sum_{i=1}^{n} F(x_i, y)$. So both the sufficient and necessary conditions hold. □

Remark: Theorem 3 can help us further divide the family of the K-means distances into two types. Type I are derived from strictly convex functions ϕ, which implies that the mean is the unique minimizer of the sum of the distances in each cluster. On the contrary, type II K-means distances are derived from convex but not strictly convex functions ϕ, which means the sum of the distances in each cluster has more than one minimizers. Next, we would like to introduce some widely used K-means distances of different types.

Example 1 *Let $\phi(x) = \|x\|^2$, then we can have a familiar K-means distance: $F(x, y) = \|x - y\|^2$, i.e., the squared Euclidean distance.*

Example 2 *Let $\phi(x) = -H(x)$, where x is a discrete probabilistic distribution, and $H(x)$ is the entropy of x. Then we produce a K-means distance: $F(p, q) = D(p\|q)$, where $D(p\|q)$ is the relative entropy of distributions p and q.*

Remark: We can proved that both $\|x\|^2$ and $-H(x)$ are strictly convex. That means the squared Euclidean distance and the relative entropy are type I K-means distances. Actually, type I K-means distance has another name: Bregman divergence, first introduced by Bregman [6], and recently studied by Banerjee et al. [3]. Therefore, the Bregman divergence is not the K-means distance itself but just one type of it. □

Example 3 *Let $\phi(x) = \|x\|$, then we can induce a K-means distance*

$$F(x, y) = \|x\| - \|y\| - (x - y)^t \nabla_y(\|y\|) = \|x\| - \frac{x^t y}{\|y\|}.$$

Remark: Since $x^t y/\|y\| = \|x\| \cos(\theta) \leq \|x\|$, thus $F(x, y) \geq 0$, which implies ϕ is convex. Furthermore, $\forall \mathbb{C} = \{x_1, \cdots, x_n\} \subset \mathbb{R}^d$ with $\|x_i\| = 1$, $k \sum_{i=1}^n x_i$ is the minimizer of $\sum_{i=1}^n F(x_i, y)$, where k can be any positive real number. Thus according to Theorem 3, this distance is a type II K-means distance. Finally, we can show

$$\min_y \sum_{i=1}^n (1 - \frac{x_i^t y}{\|y\|}) \Leftrightarrow \max_y \sum_{i=1}^n \frac{x_i^t y}{\|y\|} = \max_y \sum_{i=1}^n \cos(x_i, y).$$

Therefore $F(x, y)$ is equivalent to the cosine similarity in K-means clustering. In other words, the cosine similarity can be transformed into a type II K-means distance. A similar measure, i.e., the coefficient of correlation , can also be equivalently transformed into a type II K-means distance, if the instances have been centralized and standardized to be $x' = (x - \mu_x)/\|x - \mu_x\|$. □

In summary, under the generalized framework of the K-means distance, we can unify the type I and type II distances. Specifically, it is interesting to show that the cosine similarity, which has long been regarded as a directional measure compared with the Bregman divergence, can be also equivalently transformed into a K-means distance.

3. The Uniformization Effect

In this section, we describe the uniformization effect of K-means. First, we provide the definition of the uniformization effect of K-means as follows.

Definition 3 *We say that the K-means clustering shows the uniformization effect, if the distribution of the cluster sizes by K-means is more uniform than the distribution of the class sizes.*

3.1 The Uniformization Effect of K-means

Here we illustrate the existence of the uniformization effect of K-means. As we know, the proximity function that fits K-means is the K-means distance, so the objective function of K-means can be rewritten as follows.

$$obj = \sum_{i=1}^{k} \sum_{j=1}^{n_i} kd(x_j^{(i)}, \overline{x}^{(i)}),$$

where $kd(x, y)$ is the K-means distance, n_i and $\overline{x}^{(i)}$ are the size and the centroid (mean) of cluster i respectively. If we substitute $kd(x, y)$ by $\phi(x) - \phi(y) - (x - y)^t \nabla \phi(y)$, we have

$$obj = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \phi(x_j^{(i)}) - \sum_{i=1}^{k} n_i \phi(\overline{x}^{(i)}). \quad (5)$$

So to minimize obj is to maximize

$$obj' = \sum_{i=1}^{k} n_i \phi(\overline{x}^{(i)}). \quad (6)$$

Let $\overline{x} = \sum_{i=1}^k n_i \overline{x}^{(i)}/n$, where $n = \sum_{i=1}^k n_i$. Apparently $\overline{x}$ is the mean of all instances. Here, we consider the case that the cluster number $k = 2$ and illustrate the uniformization effect of K-means.

First, we use the Taylor's expansion on ϕ with the expansion point being $\overline{x}$:

$$\phi(\overline{x}^{(i)}) \simeq \phi(\overline{x}) + \nabla \phi(\overline{x})^t (\overline{x}^{(i)} - \overline{x}) + \frac{1}{2!} (\overline{x}^{(i)} - \overline{x})^t \nabla^2 \phi(\overline{x}) (\overline{x}^{(i)} - \overline{x}). \quad (7)$$

We assume that the higher order infinitesimal items are small enough to be ignored. Also, it's trivial to show that, given $n = n_1 + n_2$, when cluster number $k = 2$

$$\overline{x}^{(1)} - \overline{x} = \frac{n_2}{n} (\overline{x}^{(1)} - \overline{x}^{(2)}), \quad \overline{x}^{(2)} - \overline{x} = \frac{n_1}{n} (\overline{x}^{(2)} - \overline{x}^{(1)}). \quad (8)$$

Therefore, if we substitute $\phi(\overline{x}^{(i)})$ in Equation (6) by Equation (7) and use Equation (8), we can get

$$obj' \simeq n\phi(\overline{x}) + \frac{n_1 n_2}{n}(\overline{x}^{(1)} - \overline{x}^{(2)})^t \nabla^2 \phi(\overline{x})(\overline{x}^{(1)} - \overline{x}^{(2)}). \quad (9)$$

Since ϕ is convex, $\nabla^2 \phi$ is semi-positive definite, i.e.,

$$(\overline{x}^{(1)} - \overline{x}^{(2)})^t \nabla^2 \phi(\overline{x})(\overline{x}^{(1)} - \overline{x}^{(2)}) \geq 0.$$

Therefore, if we isolate the effect of $\overline{x}^{(1)} - \overline{x}^{(2)}$, the maximization of obj' implies the maximization of $n_1 n_2$, which leads to $n_1 = n_2 = n/2$.

In the case of $k > 2$, i.e., the cluster number is greater than two, the situation is much more complicated. So we leave it to the experimental part.

3.2 The Uniformization Effect and the Centroid Distance

Here, we explore the relationship of the uniformization effect and the centroid distances between the pairs of classes. Please note that the distance function used here is also the K-means distance.

Assume D is a data set with two classes $\mathbb{C}_1 = \{x_i\}_{i=1}^m$ and $\mathbb{C}_2 = \{y_j\}_{j=1}^n$. Let $\overline{x} = \sum_{i=1}^m x_i/m$ and $\overline{y} = \sum_{j=1}^n y_j/n$ denote their centroids respectively. Similar to the proof of Theorem 1, we can have

$$\triangle_{1-2} = \sum_{i=1}^{m} kd(x_i, \overline{y}) - \sum_{i=1}^{m} kd(x_i, \overline{x}) = m \times kd(\overline{x}, \overline{y}),$$

$$\triangle_{2-1} = \sum_{j=1}^{n} kd(y_j, \overline{x}) - \sum_{j=1}^{n} kd(y_j, \overline{y}) = n \times kd(\overline{y}, \overline{x}).$$

That means *on average* the distance between the instances of class $\mathbb{C}_1$ ($\mathbb{C}_2$) and the centroid of class $\mathbb{C}_2$ ($\mathbb{C}_1$) is larger than the distance between the same instances and their own centroid, and the gap is right the distance of the two centroids. In other words, the closer the two centroids can get, the closer the instances of one class and the centroid of the other class can be. Under such condition, we can expect statistically that more instances from the larger class can be assigned to the centroid of the smaller class, which may result in a relatively balanced distribution — that is what we called the uniformization effect. By contrast, if the centroid distance is large, the uniformization effect can be ignorable even if the two classes are highly imbalanced.

Therefore, in general, the uniformization effect of K-means is strongly related to the centroid distances between the classes. Specifically, small centroid distances tend to induce a significant uniformization effect.

4. Experimental Evaluation

In this section, we present experimental results to demonstrate the uniformization effect of K-means clustering. We also empirically study the cluster validation issues related to the uniformization effect.

Table 2. Some Notations.

CV_0: the CV value of the class sizes
CV_1: the CV value of the cluster sizes after clustering
DCV: $\text{CV}_1 - \text{CV}_0$

4.1 Experimental Measurements

In this subsection, we first introduce some important measures used in our experiments.

Proximity Measures. We used three kinds of K-means distances as shown in Table 1. In general, cosine distance shows merits on clustering high-dimensional data such as the document data and the gene expression data, and Euclidean distance prefers data sets with normal dimensionality. KL-divergence has information theoretical meanings on clustering words in document data sets [9].

Cluster Validity Measures. We used two external clustering validation measures: Entropy and Variation of Information (VI), which are based on the information theory. Entropy measures the purity of the clusters with respect to the given class labels. It has been widely used in data mining community, and the details can be found in [25, 26, 23]. VI, a new measure introduced by an axiomatic view [20], measures the amount of information that is lost or gained in changing from the class set to the cluster set. In particular, VI is a true metric that satisfies some important axioms uniquely [20]. Details of VI can be found in [19]. In general, the lower the Entropy or the VI value is, the better the clustering performances.

A Measure of Dispersion Degree. Here we introduce the Coefficient of Variation (CV) [7], which measures the dispersion degree of a data set. 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 $\text{CV} = s/\bar{x}$ where $\bar{x} = \sum_{i=1}^n x_i/n$ and $s = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2/(n-1)}$. Note that there are some other statistics, such as standard deviation and skewness [7], which can also be used to characterize the dispersion of a data distribution. However, the standard deviation has no scalability; that is, the dispersion of the original data and stratified sample data is 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

Table 1. Various K-means Distances.

K-means Distance	$\phi(x)$	$Kd(x,y)$
squared Euclidean distance	$\|x\|^2$	$\|x-y\|^2$
KL-divergence (relative entropy)	$\sum_{i=1}^{d} p_i \log_2 p_i$	$\sum_{i=1}^{d} p_i \log_2(p_i/q_i)$
cosine distance†	$\|x\|$	$\|x\| - \frac{x^t y}{\|y\|}$

†: Actually we use equivalent cosine similarity instead of cosine distance in the experiments.

Table 3. Experimental Data Sets.

Data set	Source	#objects	#features	#classes	Min class size	Max class size	CV_0
Document Data Sets							
hitech	TREC	2301	126373	6	116	603	0.495
sports	TREC	8580	126373	7	122	3412	1.022
tr11	TREC	414	6429	9	6	132	0.882
tr12	TREC	313	5804	8	9	93	0.638
tr23	TREC	204	5832	6	6	91	0.935
tr31	TREC	927	10128	7	2	352	0.936
tr41	TREC	878	7454	10	9	243	0.913
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
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
Other Data Sets							
ecoli	UCI	336	7	8	2	143	1.160
optdigits	UCI	5620	64	10	554	572	0.012
pendigits	UCI	10992	16	10	1055	1144	0.042
segment	UCI	2310	19	7	330	330	0.000
a1a	LIBSVM	1605	123	2	395	1210	0.718
fourclass	LIBSVM	862	2	2	307	555	0.407
diabetes_scale	LIBSVM	768	8	2	268	500	0.427
dna.scale	LIBSVM	2000	180	3	464	1051	0.500
german.numer	LIBSVM	1000	24	2	300	700	0.566
ijcnn1	LIBSVM	49990	22	2	4853	45137	1.140
ionosphere_scale	LIBSVM	351	34	2	126	225	0.399
satimage.scale	LIBSVM	4435	36	6	415	1072	0.425

mean values. In general, the larger the CV value is, the greater the variability is in the data.

4.2 The Experimental Setup

Experimental Tools. In our experiments, we used the MATLAB 7.1 [1] and CLUTO 2.1.1 [14] implementations of K-means. The MATLAB version is suitable for dense data sets, and we modified it so as to incorporate more K-means distances, such as KL-divergence. CLUTO is used to handle sparse data sets; that is, all and only the experimental results with cosine similarity as the proximity measure were produced by CLUTO. Note that the parameters of K-means in CLUTO were set to match the ones in MATLAB for the comparison purpose, and the cluster number K was set to match the class number of each data set.

Experimental Data Sets. For our experiments, 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 3. The relevant notations can be found in Table 2.

Document Data Sets. 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 `tr11`, `tr12`, `tr23`, `tr31`, `tr41` and `tr45` were derived from the TREC-5[24], TREC-6 [24], and TREC-7 [24] collections. The classes of these data sets correspond to the documents that were judged relevant to particular queries. The `la2` data set is part of the TREC-5 collection [24] and contains news articles from the Los Angeles Times. The `ohscal` data set was obtained from the OHSUMED collection [11], 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 [15]. The data sets `k1a` and `wap` were from the WebACE project (WAP) [10]; each document corresponds to a web page listed in

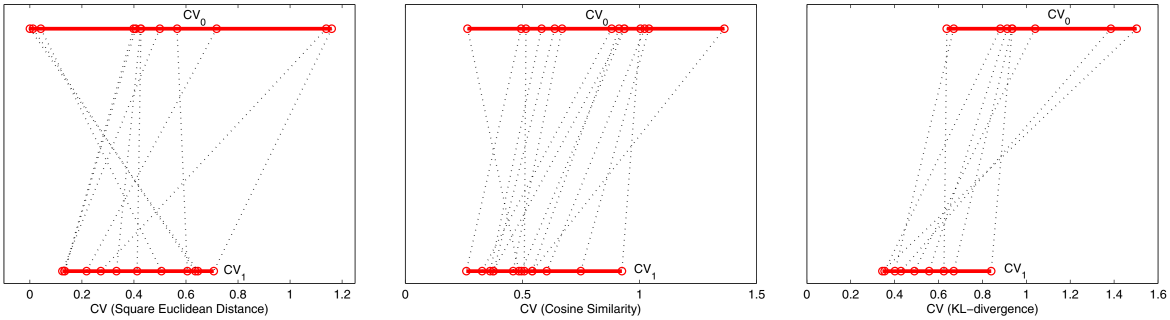


Figure 1. The Uniformization Effect of K-means.

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 [21].

Biomedical Data Sets. `LungCancer` and `Leukemia` data sets are from the Kent Ridge Biomedical Data Set Repository (KRBDSR) [16]. 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 of samples that do not fit in any of the above 6 subtypes, and each subtype is described by 12558 genes.

Other Data Sets. Besides the above high-dimensional data sets, we also used some data sets with normal dimension sizes. Among them, the `ecoli`, `optdigits`, `pendigits` and `segment` data sets are from the well-known UCI repository, which have been widely used in data mining community. The rest eight data sets are from LIBSVM repository [2], which contains data sets for the classification purpose originally. Note that some of these data sets are right from the UCI repository which have been standardized by LIBSVM for the classification task.

4.3 The Uniformization Effect of K-means

In this subsection, we demonstrate the existence of the uniformization effect of K-means. We first applied K-means with different types of distances on different types of data sets, then computed the CV_1 values for the resultant distributions of the cluster sizes. Thus we can compute the corresponding DCV values, which indicate the effect of K-means on the data distributions. Finally, we evaluated the clustering performances by two widely used measures: Entropy and VI.

Specifically, for cosine similarity, each document instance x was standardized to have unit length before applying clustering, i.e., $\|x\| = 1$. For KL-divergence, we used the so-called co-clustering scheme; that is, we viewed the document data set as a "word-document" matrix, and employed K-means with KL-divergence on the words first so as to get 100 word clusters, then reduced the features according to the word clusters, and finally employed K-means with KL-divergence again on the modified data set to get the final document clusters. The reason of co-clustering can be found in [9]. Also note that in practice we standardized each instance x to make $\sum_{i=1}^{d} x_i = 1$, where x_i is the ith attribute value of x.

Table 4 shows the results. As can be seen, no matter what K-means distance we used, for the data sets with large CV_0 values, K-means tends to reduce the variation on the cluster sizes of the clustering results, as indicated by the negative DCV values. This result indicates that, for data sets with highly imbalanced class sizes, the uniformization effect is dominant in the objective function. Also, there are five data sets including `optdigits`, `pendigits`, `segment`, `ohscal` and `german.numer`, which have small CV_0 values. Indeed, for these data sets with relatively balanced class sizes, the uniformization effect of K-means is not significant and can be dominated by other factors, such as the centroid distances, the densities or the shapes of the classes.

Now let's take a closer look on the CV_1 values, i.e., the distribution of the cluster sizes. Figure 1 shows the comparisons of CV_1 and CV_0 values when applying different K-means distances. Please note that a dot line represents a data set in Table 4. As can be seen, after applying K-means, the distributions of the cluster sizes, i.e., the CV_1 values, are in a much narrower interval. Also, we test the hypothesis that all the CV_1 values are normally distributed. The results show that the p values of the Shapiro-Wilk W statistic and χ^2 statistic are 0.887 and 0.478 respectively, which implies not to reject the null hypothesis (given the significance level $\alpha = 0.05$). Therefore we can compute the 95% confidence interval of CV_1: [0.09, 0.85], given the mean and the standard deviation of CV_1 values are 0.470 and 0.192, respectively. In other words, for data sets with CV_0 values greater than 0.85, the uniformization effect of K-means will take place with a very high probability.

Table 4. Experimental Results.

Distance	Dataset	#cluster	CV_0	CV_1	DCV	E	VI
Square	ecoli	8	1.160	0.707	-0.454	0.664	1.746
Euclidean	optdigits	10	***0.012***	***0.636***	***0.624***	1.357	2.461
Distance	pendigits	10	***0.042***	***0.506***	***0.464***	1.058	1.965
	dna	3	0.500	0.125	-0.374	0.955	2.015
	satimage	6	0.425	0.333	-0.092	0.935	1.913
	a1a	2	0.718	0.218	-0.500	0.691	1.559
	ijcnn1	2	1.140	0.273	-0.867	0.460	1.433
	ionosphere_scale	2	0.399	0.133	-0.266	0.821	1.694
	german.numer	2	***0.566***	***0.605***	***0.040***	0.869	1.721
	fourclass	2	0.407	0.135	-0.272	0.877	1.808
	diabetes_scale	2	0.427	0.412	-0.015	0.875	1.754
	Segment	7	***0.000***	***0.646***	***0.646***	1.164	1.983
Min	-	2	0.000	0.125	-0.867	0.460	1.433
Max	-	10	1.160	0.707	0.646	1.357	2.461

Distance	Dataset	#cluster	CV_0	CV_1	DCV	E	VI
Cosine	hitech	6	0.495	0.261	-0.234	1.631	3.394
Similarity	sports	7	1.022	0.604	-0.418	1.112	2.662
	la2	6	0.516	0.508	-0.008	1.563	3.147
	ohscal	10	***0.266***	***0.486***	***0.219***	2.156	4.223
	tr11	9	0.882	0.328	-0.554	0.883	2.171
	tr12	8	0.638	0.377	-0.261	0.965	2.082
	tr23	6	0.935	0.375	-0.559	1.475	3.362
	tr31	7	0.936	0.496	-0.440	1.228	2.886
	tr41	10	0.913	0.542	-0.371	1.159	2.666
	tr45	10	0.669	0.460	-0.209	1.310	2.775
	k1a	20	1.004	0.925	-0.079	1.547	3.266
	wap	20	1.040	0.749	-0.291	1.536	3.339
	LungCancer	5	1.363	0.542	-0.820	0.608	1.877
	Leukemia	7	0.582	0.363	-0.219	1.522	3.178
Min	-	5	0.266	0.261	-0.820	0.608	1.877
Max	-	20	1.363	0.925	0.219	2.156	4.223

Distance	Dataset	#cluster	CV_0	CV_1	DCV	E	VI
KL-divergence	tr11	9	0.882	0.355	-0.527	0.418	1.220
	tr12	8	0.638	0.625	-0.013	0.219	0.453
	tr23	6	0.935	0.670	-0.265	0.639	1.525
	tr31	7	0.936	0.840	-0.096	0.290	0.766
	tr41	10	0.913	0.557	-0.355	0.335	0.988
	tr45	10	0.669	0.402	-0.268	0.431	1.050
	wap	20	1.040	0.490	-0.550	0.692	1.827
	re0	13	1.502	0.345	-1.157	1.128	3.253
	re1	25	1.385	0.429	-0.956	1.056	2.829
Min	-	6	0.638	0.345	-1.157	0.219	0.453
Max	-	25	1.502	0.840	-0.013	1.128	3.253

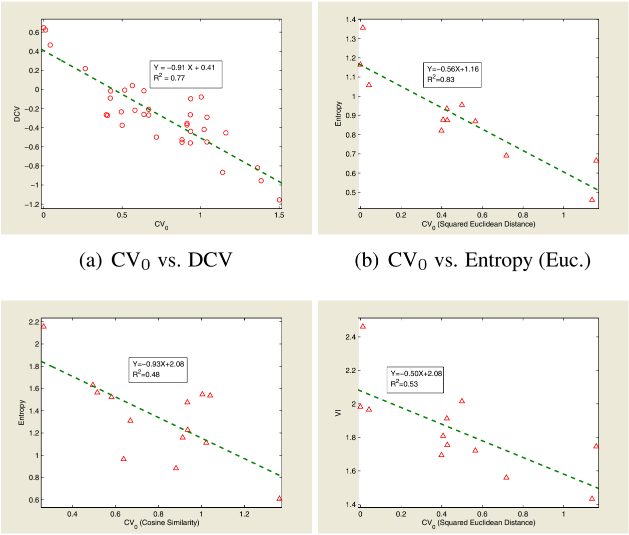
Since CV_1 values have such a narrow interval, we can expect that as the skewness of the distribution of class sizes increases, the results by K-means clustering tend to be farther away from the true ones. Figure 2(a) indeed shows this trend; that is, the absolute DCV values increase as the CV_0 values increase. Therefore, in general, applying K-means clustering on highly imbalanced data sets is not very effective, which can be indicated by the DCV measure.

4.4 The Uniformization Effect and the Cluster Validity Problem

Here, we illustrate that the uniformization effect of K-means can make negative impact on the cluster validity. More specifically, some well-known clustering validation measures may not have the ability to identify the uniformization effect of K-means, so as to deliver unreliable scores on the clustering results. In this subsection, we will focus on two measures: Entropy and Variation of Information (VI). The former has been widely used in data mining community, and the latter was well established on the information theory and a set of axioms.

Entropy. We first perform the analysis using the Entropy measure. Figure 2(b) and 2(c) show the Entropy values along data sets with increasing variation on the class sizes. The common ground of these figures is that the Entropy values tend to be systematically lower (better) on results of highly imbalanced data sets. This seriously contradicts the findings above: K-means tends to produce poorer partitions on highly imbalanced data sets due to the uniformization effect. Therefore, we can suspect that, Entropy may not be suitable for evaluating the results produced by K-means with squared Euclidean distance or cosine similarity.

Furthermore, we explore the reason why Entropy is unreliable on K-means. The key point is, Entropy only assesses the purity of each cluster, but does not promise to find every class of the data set. In detail, for a highly imbalanced data set, due to its uniformization effect, K-means tends to break down the large class and assign the pieces to different clusters. Since the number of instances of the large class



(a) CV_0 vs. DCV

(b) CV_0 vs. Entropy (Euc.)

(c) CV_0 vs. Entropy (cos.)

(d) CV_0 vs. VI (Euc.)

Figure 2. Relationships between CV_0 and Validation Measures.

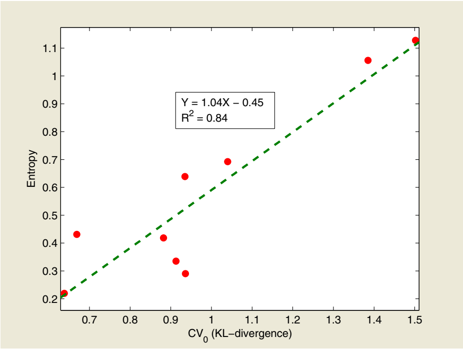


Figure 3. CV_0 vs. Entropy (KL-divergence).

can be “huge” compared with other classes, the pieces of it can still be dominant in their corresponding clusters, so as to make the clusters be rather “pure”, which results in a low Entropy value. Actually, this seeming “purity” is at the cost of missing many small classes in the data! A detailed illustration can also be found in [25].

Variation of Information (VI). VI is a more sophisticated clustering validation measure which implicitly takes the integrality of all classes in data into consideration. As a result, VI is also more complicated than Entropy — to directly understand its relationship with K-means and the uniformization effect is rather difficult. Therefore, we here only evaluate them empirically.

Figure 2(d) shows the VI values along data sets with increasing variation on class sizes. A similar trend as Entropy can be found; that is, the VI values tend to be systematically lower (better) on results of highly imbalanced data sets. Therefore, we can suspect that VI also has troubles on evaluating the K-means clustering, especially for highly imbalanced data sets.

Please note that Figure 2(d) is based on the Euclidean distance, however this observation can be extended to the situation when applying cosine similarity. We omit the results based on cosine similarity due to the page limitation.

Measures for K-means with KL-divergence. Here, we study the effectiveness of the two measures on evaluating results produced by K-means with KL-divergence. An interesting observation is that, compared with their previous performances on squared Euclidean distance or cosine similarity, both Entropy and VI measures show rather different behaviors on KL-divergence.

As indicated by Figure 3, the Entropy values increase as the CV_0 values go up. This is opposite to the trends in Figure 2(b) and 2(c). This implies that, given KL-divergence as the K-means distance, Entropy can identify poor results by K-means on highly imbalanced data sets. As to VI, we can simply compute the coefficient of correlation between VI values and Entropy values. The 0.99 result implies that VI acts in a way very similar to Entropy; that is, the VI values increase as the distributions of the resultant cluster sizes being farther away from the true ones. In other words, VI is also effective on evaluating results produce by K-means with KL-divergence.

5. Related Work

Here, we highlight some research results which are mostly related to the main theme of this paper.

First, people have studied the impact of high dimensionality on the performance of K-means, and found that the traditional Euclidean notion of proximity is not effective for K-means clustering on high-dimensional data sets. To meet this challenge, one research direction is to make use of dimensionality reduction techniques, such as multidimensional scaling (MDS) [4], principal components analysis (PCA) [13], and singular value decomposition (SVD) [8]. Another direction for this problem is to redefine the notions of proximity, e.g., by the Shared Nearest Neighbors (SNN) similarity [12]. Some other similarity measures, such as the cosine similarity, have also been used as proximity functions for clustering high-dimensional document data sets [26].

Second, there are a number of choices for the proximity function that can be used in K-means. For instance, the Euclidean distance has been widely used, and the cosine similarity, KL-divergence as well as Itakura-Saito distance have also shown their advantages for document clustering [26], words clustering [9] and power spectra analysis [17], respectively. Recently, in his inspiring work, Banerjee [3] proposed a general framework for K-means clustering that uses proximity functions based on Bregman divergences [6]. This work shares some common grounds with our work. However, in this paper, we propose a general concept of K-means distance for proximity functions that fit K-means. Some well-known proximity functions of K-means, such as the cosine similarity, the coefficient of correlation, and the Bregman divergence, are instances of K-means distance.

Finally, in their previous work [25], Xiong et al. preliminarily studied the uniformization effect of K-means and the cluster validation issues. However, their focus is merely on the squared Euclidean distance and the Entropy measure. Therefore, this paper is a natural extension of [25]. Indeed, we show that the uniformization effect exists no matter which proximity function is used in K-means as long as this proximity function fits K-means clustering. Also, we show the impact of skewed cluster distributions on the performance of some other external cluster validation measure, such as Variation of Information (VI), which has been well established in the machine learning community [19, 20].

6. Conclusions

In this paper, we studied the generalization issues of proximity functions for K-means. Specifically, we showed that a proximity function that fits K-means can be derived from a differentiable convex function. We call such proximity functions as K-means distance. Also, we theoretically proved that some widely used proximity functions, such as the Bregman divergence and the cosine similarity, are the instances of K-means distance. In addition, we revealed that K-means has a general uniformization effect; that is, K-means tends to produce clusters with relatively uniform sizes regardless proximity functions used. Finally, experimental results on real-world data sets show the uniformization effect of K-means. We also observed that both Entropy and VI have difficulty in measuring clustering quality if the class sizes of data have skewed distributions.

7. Acknowledgements

This research was partially supported by the National Science Foundation of China (NSFC) (Nos. 70621061, 70518002). Also, this research was supported in part by a Faculty Research Grant from Rutgers Business School-Newark and New Brunswick.

References

[1] K-means clustering in statistics toolbox of matlab 7.1.

[2] Libsvm. In *www.csie.ntu.edu.tw/ cjlin/libsvm/*.

[3] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clusteringwith bregman divergences. *JMLR*, 6:1705–1749, 2005.

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

[5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[6] L. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[7] M. DeGroot and M. Schervish. *Probability and Statistics*. Addison Wesley; 3 edition, 2001.

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

[9] I. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *JMLR*, 3(4):1265–1278, 2003.

[10] E.-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore. Webace: A web agent for document categorization and exploration. In *Proc. of Agents'98*, 1998.

[11] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *SIGIR-94*, 1994.

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

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

[14] G. Karypis. In *http://www-users.cs.umn.edu/ karypis/cluto/*.

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

[16] J. Li and H. Liu. In *http://sdmc.i2r.a-star.edu.sg/rp/*.

[17] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

[18] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Proc. of the 5th BSMSP, Volume I, Statistics*. University of California Press, September 1967.

[19] M. Meila. Comparing clusterings by the variation of information. In *Proc. of the 16th COLT*, 2003.

[20] M. Meila. Comparing clusterings- an axiomatic view. In *Proc. of the 22th ICML*, 2005.

[21] M. F. Porter. An algorithm for suffix stripping. In *Program, 14(3)*, 1980.

[22] J. Stevens. *Applied Multivariate Statistics for the Social Sciences*. Lawrence Erlbaum Associates, 2001.

[23] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[24] TREC. Text retrieval conference. In *http://trec.nist.gov*.

[25] H. Xiong, J. Wu, and J. Chen. K-means clustering versus validation measures: A data distribution perspective. In *Proc. of the 12th ACM SIGKDD*, August 2006.

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