A New Mallows Distance Based Metric For Comparing Clusterings

Ding Zhou DZHOU@CSE.PSU.EDU
Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16801, USA

Jia Li JIALI@STAT.PSU.EDU
Department of Statistics, The Pennsylvania State University, University Park, PA 16801, USA

Hongyuan Zha ZHA@CSE.PSU.EDU
Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16801, USA

Abstract

Despite of the large number of algorithms developed for clustering, the study on comparing clustering results is limited. In this paper, we propose a measure for comparing clustering results to tackle two issues insufficiently addressed or even overlooked by existing methods: (a) taking into account the distance between cluster representatives when assessing the similarity of clustering results; (b) constructing a unified framework for defining a distance based on either hard or soft clustering and ensuring the triangle inequality under the definition. Our measure is derived from a complete and globally optimal matching between clusters in two clustering results. It is shown that the distance is an instance of the Mallows distance—a metric between probability distributions in statistics. As a result, the defined distance inherits desirable properties from the Mallows distance. Experiments show that our clustering distance measure successfully handles cases difficult for other measures.

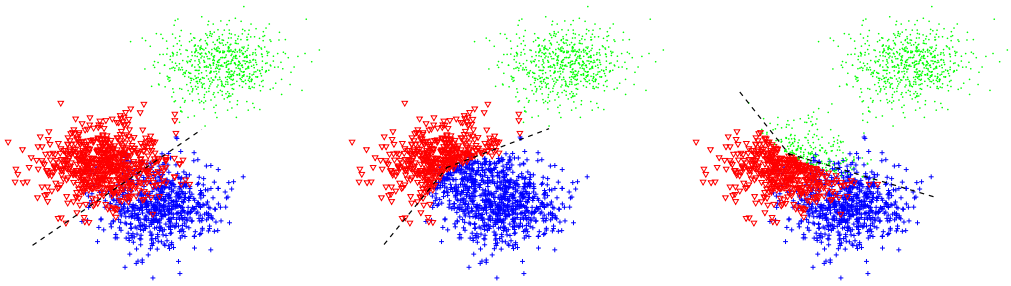
1. Introduction

As a primary knowledge discovery technique used in various fields, clustering has been extensively studied for decades. With the existence of many clustering algorithms, it is often desirable to assess the extent of "agreement" between two clustering results (Meila, 2002). For brevity, hereafter, we refer to a clustering result as clustering and a set of objects grouped together as a cluster.

There have been some, if not many, efforts on comparing clusterings. Existing methods form three categories (Meila, 2002): (1) pair counting (Ben-Hur, 2002; Fowlkes, 1983; Hubert, 1985; Rand, 1971), (2) set matching (Meila, 2002; Dongen, 2000) and (3) variation of information (VI) (Meila, 2002). The *pair counting* method evaluates the similarity between two clustering algorithms by examining how likely they are to group a pair of objects together, or, separate them in different clusters. All pair counting methods are restricted to handling hard clustering. Other drawbacks of pair counting methods are also discussed in (Fowlkes, 1983). The *set matching* method seeks for a match between clusters, that is, the sets of objects grouped together in two clusterings respectively. Existing set matching approaches perform matching in a stepwise manner without a global optimization objective. In the case when two clusterings possess different numbers of clusters, some clusters may even be ignored and play no role in the comparison. The *VI* measure computes the amount of information that is lost or gained in changing from one clustering to the other. It also addresses the problem of soft clustering. However, VI no longer maintains the triangle inequality when handling soft clustering.

All the aforementioned methods compare clusterings based only on the memberships of objects to clusters. An important aspect neglected in the comparison is the variation of similarity between pairs of cluster representatives. In particular, for vector data, a cluster representative can be the mean of that cluster. Consider the following example. Given three clusters A,B and C, suppose A and B are relatively similar in the sense of their representatives while C is quite different from both, as shown in Figure 1(a). Suppose in another two clustering results, some data in A are mislabeled as cluster B and C respectively, as shown in Figure 1(b) and 1(c). Existing methods yield the same distance between the original clustering and each of the other two clusterings. We will propose a measure to take into account the similarity between cluster representatives. Un-

Appearing in *Proceedings of the 22^{nd} International Conference on Machine Learning*, Bonn, Germany, 2005.



(a) Dataset (b) Clustering (I) (c) Clustering (II)

Figure 1. Clusters with different inter-distances

der this new measure, the distance between the clusterings formed by assigning objects in A to B will be smaller than that by assigning the objects to C instead.

In this paper, we propose new measures for the distance between clusterings under two scenarios: (1) comparing clusterings based on the memberships of objects to clusters alone; (2) based on memberships as well as the similarity between cluster representatives. For brevity, we refer to the first case as categorical clustering comparison and the second as comparison with similarity differentiation. Our work advances existing clustering comparison techniques in several aspects: (1) The similarity between cluster representatives is taken into consideration when comparing clusterings; (2) For categorical clustering comparison, all the clusters are guaranteed to affect the measure as a result of using a globally optimal matching approach. (3) The clustering distance developed for the categorical comparison treats hard and soft clustering in a seamlessly unified manner and the distance satisfies intrinsic properties of a metric, e.g., the triangle inequality. For most existing clustering distance measurements, however, the triangle inequality is violated.

The rest of the paper is organized as follows. Section 2 and Section 3 provide background, notations and introduction to the Mallows distance. Section 4 describes our measures in two scenarios. Experiments are presented in Section 5. We conclude in Section 6

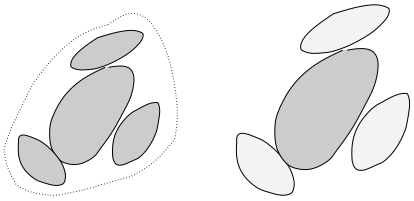
2. Preliminaries and Motivations

Clustering is the process of dividing a data set D into clusters so that objects within each cluster are highly similar to each other and those in different clusters differ as much as possible.

For *hard clustering*, each object is associated with only one cluster. For *soft clustering*, an object is associated with every cluster to a certain extent indicated by a weight. The weight assigned to cluster k for object i is denoted by $p_{i,k}$. Normally, $0 \leq p_{i,k} \leq 1$ and $\sum_{k=1}^{K} p_{i,k} = 1$. In the light of statistical modeling based clustering, $p_{i,k}$ is the posterior probability for object i belonging to cluster k. The clustering result is hence summarized by a membership probability matrix: $P_{N\times K} = (p_{i,j}), 1 \leq i \leq N, 1 \leq j \leq K$.

Hard clustering can be considered as a special case of $P_{N\times K}$ with each row containing exactly one element equal to 1 and the rest 0. If $p_{i,k} = 1$, object i is assigned exclusively to cluster k.

Let the two clustering results be denoted by Cls_1 and Cls_2, whose membership matrices are $P_{N\times K}$ and $Q_{N\times J}$. The distance between Cls_1 and Cls_2 in the categorical comparison case is a function of the two matrices: $D(Cls_1, Cls_2) = \Gamma(P_{N\times K}, Q_{N\times J})$. On the other hand, for the comparison that differentiates clusters based on their labels as well as representatives, the distance between Cls_1 and Cls_2 can be denoted by: $D(Cls_1, Cls_2) = \Delta(P_{N\times K}, Q_{N\times J}, R_P, R_Q)$, where R_P and R_Q are matrices containing cluster representative vectors generated by Cls_1 and Cls_2.



(a) Low threshold (b) High threshold

Figure 2. Clustering with high, low thresholds

Our motivation for this paper is based on the the following observations: First, existing comparing methods lack similarity consideration for clusters within each clustering. Second, existing set matching methods all greedily search for a "best match" for each cluster and then add up the contributions of the matches found. By doing so, these methods ignore the "unmatched" part of each cluster. Such a case is illustrated in (Meila, 2002).

Figure 2(a) and Figure 2(b) illustrate an example when one algorithm performs on the same dataset with different parameter settings. Consider a density-based clustering algorithm (e.g. DBSCAN (Ester, 1996)): it uses a small threshold and generates one big cluster as in Figure 2(a); but in Figure 2(b), under a higher density threshold, the algorithm only identifies the inner part of this cluster, which is of high density. The outskirt data portion is divided into several small clusters distributed evenly. Intuitively, the two clustering results should be somehow "close" considering that these two results are for the same dataset obtained by the same clustering algorithm, only with different parameters. Now consider how an existing set matching method (Robardet, 2000) works for this example: it searches for the "best match" between the only cluster in Figure 2(a) and a cluster in Figure 2(b). It finds the biggest cluster in Figure 2(b) as the "best match" and stops. However, all the other small clusters in Figure 2(b) are not even considered. A possibly better way is to match the big cluster in Figure 2(a) with all the clusters in Figure 2(b), distributing weights on the match with both big and small clusters. *A key issue is how to determine the weights assigned to the matching between each pair of clusters in a certain optimal sense.*

Another problem concerning most existing clustering comparing methods is that they were primarily designed to handle comparison of hard clusterings. When comparing clusters only based on hard labels, we may encounter situations where the comparison conflicts intuition. Here we give one example to show the importance of comparing soft clustering:

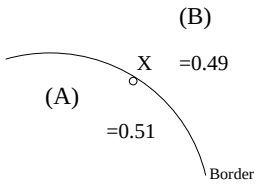


Figure 3. Data object near cluster border

As shown in Figure 3, consider a data object X on the border of two clusters (A) and (B). By soft clustering Cls_1, we get the probability of 0.51 that X belongs to (A) and 0.49 that X belongs to (B). Suppose there is another clustering Cls_2 that generate the opposite probability: $p(X \in (A)) = 0.49$ and $p(X \in (B)) = 0.51$.

Using hard clustering comparison, we merely compare the labels after firstly label X with (A) in Cls_1 while (B) for X in Cls_2. In a case when there are many objects on the border, Cls_1 and Cls_2 tend to be quite different under hard comparison though in fact they are not. As a result the original meanings of the clustering results may be misinterpret if only hard clustering comparison is available. *Hence, it is compelling for us to design a measure that is capable of handling soft clustering*.

3. Monge-Kantorovich mass transfer and Mallows distance

The distance measure we discuss in this paper has its roots in optimal mass transport problems and Mallows distance for measuring the difference between two multivariable probability distributions (Rachev, 1984; Mallows, 1972). In this section, we provide some background materials on those topics.

The original mass transport problem proposed by Monge in 1700s asks how to move a pile of dirt to a fill with the least amount of work. In the 1940s, Kantorovich gave a relaxed formulation of the problem and proposed a dual variational principle for solving the problem. Consider two probability distributions P and Q on R^n. Define

$$M = \{\text{probability distribution } \mu(x,y) \text{ on } R^n \times R^n\} \mid \int_y d\mu(x,y) = P(x), \int_x d\mu(x,y) = Q(y)\}$$

Let $C(x,y)$ indicates the work to move a unit amount of mass from x to y. Then, we seek to minimize the cost functional,

$$J(\mu) = \int C(x,y) d\mu(x,y)$$

among all $\mu \in M$.

Unaware of the work in mass transport, in 1972 (Mallows, 1972) Mallows proposed to measure the difference between two probability distributions P and Q on R^n as

$$\text{Mallow}_p(P,Q) = \min_{\mu}(E_\mu \|x-y\|_p^p)^{1/p}$$

subject to:

$$\int_y d\mu(x,y) = P(x), \int_x d\mu(x,y) = Q(y)$$

where the $\|\cdot\|_p$ denotes the L_p norm, and $1 \leq p < +\infty$. Clearly, Mallows distance is a special case of Kantorovich's mass transport problem with $C(x,y) = \|x-y\|_p^p$.

For two discrete distributions $P = \{(x_1,p_1),...,(x_n,p_n)\}$ and $Q = \{(y_1,q_1),...,(y_m,q_m)\}$ with $\sum p_i = 1$ and $\sum q_i = 1$, minimizing the cost functional reduces to

$$\min_{\mu} \sum_{i=1}^{n} \sum_{j=1}^{m} \mu(i,j) C(x_i, y_j),$$

subject to

$$\mu(i,j) \geq 0 \text{ and } \sum_{j=1}^{m} \mu(i,j) = p_i;$$

$$\sum_{i=1}^{n} \mu(i,j) = q_j \text{ and } 1 \leq i \leq n; 1 \leq j \leq m.$$

The dual of the above linear programming problem is to find $u = [u_1, \ldots, u_n]$ and $v = [v_1, \ldots, v_m]$ to

$$\max \sum_{i=1}^{n} p_i u_i + \sum_{i=1}^{m} q_i v_i$$

subject to $u_i + v_j \leq C(x_i, y_j), i = 1, \ldots, n, j = 1, \ldots, m$. By solving the dual problem, we may achieve better computational efficiency.

We mention that another recent topic in which the mass transport problem and Mallows distance play a role is the measuring of texture and color similarities for image retrieval (Rubner, 1998), under the name of $Earth\ Mover's\ Distance(EMD)$. It is pointed out in (Levina, 2001) that the EMD is in fact a special case of the mass transport problem when both sides have equal amount of mass.

4. Clustering Comparison

By matching clusters in two clusterings in a globally optimal manner with all the clusters involved in the comparison, we develop clustering distance measures for the two cases: categorical comparison and comparison with similarity differentiation. For the categorical clustering distance, we address soft clustering directly. Hard clustering is a special case of soft clustering and requires no special treatment. It will be shown that the matching based categorical distance is equivalent to a Mallows distance and hence inherits its metric properties. The distance developed for comparison with similarity differentiation will be referred to as the cluster similarity sensitive (CSS) distance.

4.1. Categorical clustering distance

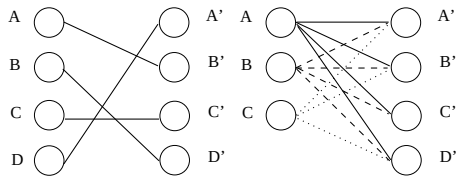
Given a dataset $D = \{x_1, x_2, ...x_N\}$, suppose two clustering results Cls_1 and Cls_2 are obtained. Cls_1 contains K clusters and Cls_2 J. For soft clustering, let the probability matrix generated by Cls_1 be $P = (p_{i,j})$, where $p_{i,j}$ denotes the probability that object x_i belongs to cluster C_j. Let the corresponding matrix generated by Cls_2 be $Q = (q_{i,j})$. Denote the K clusters in Cls_1 by C_1, ..., C_K and the J clusters in Cls_2 by C'_1, ..., C'_J.

A cluster, for instance, C_j is characterized by the N-dimensional vector $(p_{1,j}, p_{2,j}, ..., p_{N,j})^t$, denoted by ζ_j. That is, C_j is determined by the probability of each object x_i belonging to it. Similarly, denote the vector characterizing cluster C'_j by $\gamma_j = (q_{1,j}, q_{2,j}, ..., q_{N,j})^t$. To reflect the significance of each cluster in Cls_1 and Cls_2 for the purpose of comparing the two, we assign a weight to each cluster. Let the weights assigned to C_j be α_j, $\sum_{j=1}^{K} \alpha_j = 1$, and those to C'_j be β_j, $\sum_{j=1}^{J} \beta_j = 1$. Example values for α_j are $1/K$ if all the clusters are weighted equally, or the percentage of data assigned to cluster C_j with respect to the whole data set. To summarize, a clustering can be represented by a discrete distribution on the space R^N. In particular, Cls_1 corresponds to the distribution: $\mathcal{P} = \{(\zeta_1, \alpha_1), ...(\zeta_K, \alpha_K)\}$; and Cls_2 to: $\mathcal{Q} = \{(\gamma_1, \beta_1), ...(\gamma_J, \beta_J)\}$. To measure the distance between Cls_1 and Cls_2, we adopt the Mallows distance for the two distributions $\mathcal{P}$ and $\mathcal{Q}$. Without difficulty of extension, we assume that L_1 norm is used in the Mallows distance. The distance between Cls_1 and Cls_2 is thus given by

$$D(Cls_1, Cls_2) = \min_{w_{k,j}} \sum_{k=1}^{K} \sum_{j=1}^{J} w_{k,j} \sum_{i=1}^{N} |p_{i,k} - q_{i,j}| \quad (1)$$

subject to $w_{k,j} \geq 0$, $\sum_{k=1}^{K} w_{k,j} = \beta_j$, $\sum_{j=1}^{J} w_{k,j} = \alpha_k$, for all k, j.

The Mallows distance $D(Cls_1, Cls_2)$ can be interpreted as an optimal cluster matching scheme. The distance between two clusters C_k and C'_j is measured by the L_1 distance of their characterizing vectors: $\| \zeta_k - \gamma_j \|_1 = \sum_{i=1}^{N} |p_{i,k} - q_{i,j}|$. Since it is not known beforehand which pairs of clusters in Cls_1 and Cls_2 should be compared against each other, every cluster C_k in Cls_1 is "soft" matched to each C'_j, $j = 1, ..., J$, in Cls_2. The extent of matching between C_k and C'_j is indicated by weight $w_{k,j}$. By ensuring that $\sum_{k=1}^{K} w_{k,j} = \beta_j$ for all j, and $\sum_{j=1}^{J} w_{k,j} = \alpha_k$ for all k, every cluster in both Cls_1 and Cls_2 will play a role in determining the overall distance between the clusterings. In addition, the optimization over the matching weights is global and yields a unique solution for the distance. Linear programming is used to compute (1).



(a) $K = J$ (b) $K \neq J$

Figure 4. Mapping between two clusterings

In the special case when $J = K$ and $\alpha_j = 1/K$, $\beta_j = 1/K$, $j = 1, ..., K$, by the intrinsic property of linear programming, the optimal $w_{k,j}$ satisfy the following: for all j in $w_{k,j}$, $w_{k,j} = 1/K$ for one and only one k and $w_{k',j} = 0$ for all $k' \neq k$. Similarly for all k in $w_{k,j}$, $w_{k,j} = 1/K$ for one and only one j and $w_{k,j'} = 0$ for all $j' \neq j$. From the

matching perspective, this implies that the optimal matching is given by permuting the clusters in one clustering and then matching the clusters in the two clusterings one by one, as shown in 4(a). This is consistent with the intuition that when the pairing between clusters is unknown, we should seek for a pairing that minimizes the resulting average pair-wise distance between clusters.

Since the clustering distance defined is a Mallows distance, it inherits all the properties of the Mallows distance. It is shown in (Robardet, 2000) that the Mallows distance is a metric. In particular, it is nonnegative and equals zero if and only if the two distributions are identical, i.e. $\mathcal{P} = \mathcal{Q}$; it is symmetric; and it satisfies the triangle inequality.

4.2. Cluster similarity sensitive distance

In this section, we develop a clustering distance measure that takes into account the distances between cluster representatives. We motivated the usefulness of such a distance by an example in section 2. Again we follow the approach of global cluster matching.

Consider hard clustering with $J = K$ first. Suppose we seek a permutation between clusters in the two clusterings and let C_j be matched to $C'_{\rho(j)}$. An object x_i will contribute to the difference between Cls_1 and Cls_2 if it is assigned to unmatched clusters in the two clusterings. The amount of contribution depends on the similarity between the representatives, also referred to as centroids, of the two clusters. Let the distance between the centroids of C_k and C'_j be $L(k, j)$. The object-wise difference resulting from x_i is:

$$d_i = \sum_{k=1}^{K}\sum_{j=1}^{K} p_{i,k}q_{i,j}L(k,j)(1 - I(\rho(j) = k)) \ . \qquad (2)$$

As usual, $I(\cdot)$ is the indicator function that equals 1 when the argument is true and zero otherwise. For hard clustering, only one term in the above summation is possibly nonzero. Suppose $x_i \in C_k$ and $x_i \in C'_j$. Then the only possibly nonzero term is $p_{i,k}q_{i,j}L(k,j)(1 - I(\rho(j) = k))$, because $p_{i,k'} \times q_{i,j'} = 0$ if $k' \neq k$ or $j' \neq j$. If C'_j is matched to C_k, then this term will also be zero and x_i causes no difference for Cls_1 and Cls_2. We can write (2) equivalently as $d_i = \sum_{k=1}^{K}\sum_{j=1}^{K} p_{i,k}q_{i,j}L(k,j)(1 - Kw_{k,j})$, where $w_{k,j} = 0$ or $1/K$ and $\sum_k w_{k,j} = 1/K$, $\sum_j w_{k,j} = 1/K$. When $w_{k,j} = 1/K$, C_k is matched to C'_j.

Although we derived d_i by considering hard clustering with $J = K$, it is straightforward to extend (2) to soft clustering and the case $J \neq K$, where soft matching between clusters is needed. As in the previous section, let the weight assigned to the match between C_k and C'_j be $w_{k,j}$. Then, in general, the object-wise difference is

$$d_i = \sum_{k=1}^{K}\sum_{j=1}^{J} p_{i,k}q_{i,j}L(k,j)(1 - \frac{2}{\alpha_k + \beta_j}w_{k,j}) \ , \qquad (3)$$

where $w_{k,j} \geq 0$, $\sum_{k=1}^{K} w_{k,j} = \beta_j$ for all j and $\sum_{j=1}^{J} w_{k,j} = \alpha_k$ for all k. The overall distance between Cls_1 and Cls_2 is then defined as:

$$D(Cls_1, Cls_2) = \min_{w_{k,j}} \sum_{i=1}^{N} d_i$$

$$= \min_{w_{k,j}} \sum_{k=1}^{K}\sum_{j=1}^{J}(1 - \frac{2}{\alpha_k + \beta_j}w_{k,j}) \sum_{i=1}^{N} p_{i,k}q_{i,j}L(k,j) \ . (5)$$

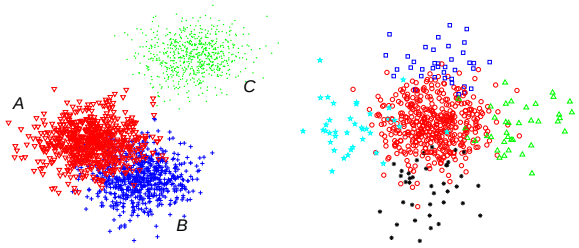
In our experiment, for both types of clustering distance, we used equal weights for all the clusters, i.e., $\alpha_k = 1/K$ for all k and $\beta_j = 1/J$ for all j. Extension to the general case causes no essential difference in the computation of the distances.

Now let us see how the defined measure works. Consider again the example concerning three clusters A, B and C illustrated in Figure 1. $Cls_{correct}$ provides the correct labels. Cls_1 mislabels objects in A to B, and Cls_2 mislabels the same objects to C instead. When comparing Cls_1 and $Cls_{correct}$, according to Equation (2), the clustering difference reflected by object i is $L(A, B')$ since only the term involving clusters A and B' is nonzero. Similarly, when comparing Cls_2 and $Cls_{correct}$, the clustering difference form object i is $L(A, C')$. As $L(A, B') < L(A, C')$, the difference resulting from object i when comparing with Cls_2 is larger than with Cls_1, which coincides with the intuition.

5. Experiments

In this section, we present empirical evaluation of several existing clustering comparing methods as well as the new measures we proposed. Four real world datasets are employed. They are the Sequoia benchmark 2000 and the $sonar$, $liver$ and $diabetes$ datasets from UCI Machine Learning Repository. Also we generated two synthetic datasets based on different purposes of comparisons: (1) dataset DS_1: we design this dataset according to our motivation for the case in Figure 1(a). Three clusters named A,B and C are produced with different inter-cluster distances, as illustrated in Figure 5(a). Three clusters are of the same size, each containing 400 2-d data objects. The distance between centroids of A and C is two times the distance between A and B; (2) dataset DS_2: 4 small clusters center around a big cluster in a satellite formation. The

dataset contains 2000 2-d points in all, 1000 of which are in the center cluster and each small cluster has around 250 points.



(a) DS_1 (b) DS_2

Figure 5. Experimental datasets

We experiment with four popular clustering algorithms: K-Means, DBSCAN (Ester, 1996), EM (Dempster, 1977) and CEM. The first two are hard clustering and the rest soft. We design the comparison of these algorithms by implementing the state-of-the-art comparing measure VI in (Meila, 2002) as well as the classical method $Rand$ proposed in (Rand, 1971). We first evaluate our categorical clustering (CC) measure for soft clustering. Then we present experimental results of our second measure, the clustering similarity sensitive (CSS) measure. In Section 5.3, we examine the computational complexity of our metric in comparison with existing methods.

5.1. Measure for soft clustering

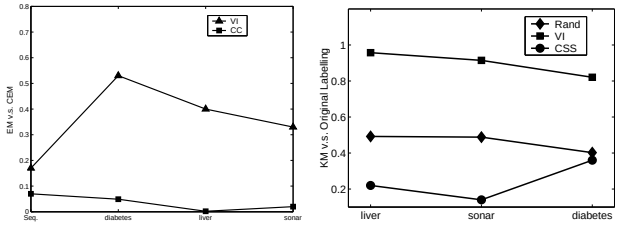
In this section, we firstly compare soft clustering comparison methods based on results produced by EM and CEM on the synthetic and real world datasets we introduced. We present the comparisons between our first CC measure and the VI method in (Meila, 2002) because VI is the only existing measure for soft clustering as far as we know.

As can be derived from Eq. 1, our CC measure is lower bounded by zero and upper bounded by the number of data points. The VI distance is lower bounded by zero and upper bounded by $log(K) + log(J)$, where K and J are the number of clusters in each clustering. In Table 1, we avoid confusion on scales by normalizing the comparison results.

	VI	CC		VI	CC
DS_1	0.19	0.0068	$diabetes$	0.53	0.049
DS_2	0.42	0.39	$sonar$	0.33	0.02
$Seq.$	0.17	0.07	$liver$	0.40	0.002

Table 1: Clustering comparing by VI and CC

First let us consider the synthetic datasets DS_2 and real dataset $Sonar$. The corresponding clustering results are illustrated in Figure 6 (results on $Sonar$ after PCA). Obviously, EM and CEM perform more differently on DS_2 than on $Sonar$. As shown in Table 1, our CC measure well matches such intuition on both datasets. By contrast, VI does not clarify the differences between EM and CEM for the two datasets by giving the close measures, 0.42 and 0.33. VI commits the large distance between Figure 6(c) and Figure 6(d) because there are many points on the boundary of two classes that are misclassified. CC measure, however, smoothes such differences in Eq. 1.



(a) Soft clustering (b) Hard clustering

Figure 7. CC/ CSS v.s. VI, $Rand$ on real world datasets

Comparisons between EM and CEM are also presented in Table 1 on three other real world datasets and one other synthetic dataset. Some real world datasets are all high dimensional datasets and we perform dimension reduction (PCA) yielding better separated clusters. By doing this we expect that EM and CEM will generate very similar clustering results for all of them. In Figure 7(a), we plot both normalized comparison results by VI and our CC measure. We can see CC coincides with our expectations while VI produces instable comparison results.

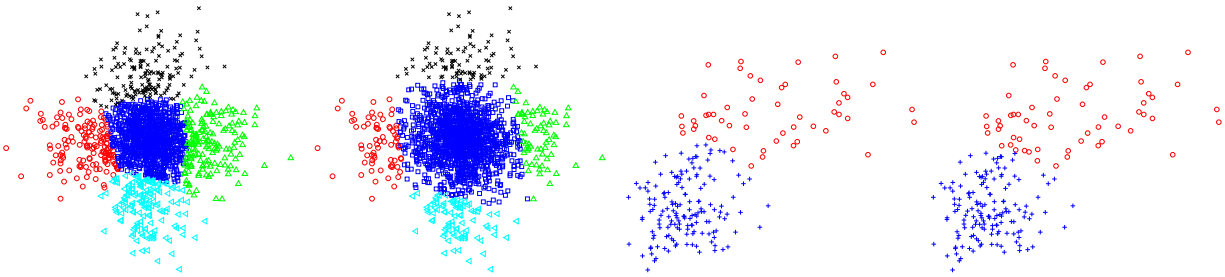
In addition to these convincing examples, from theoretical perspective, our CC measure is more appealing as it is positive, symmetric and transitive, which is inherited from the properties of $Mallows$ distance.

5.2. Measure with cluster similarity differentiation

In this section, we produce several cases of cluster labelling and implement two hard clustering algorithms: DBSCAN (Ester, 1996) and K-Means. The same two synthetic datasets and four real world datasets are used.

For DS_1, we label $1/3$ of the points in cluster A to cluster B, producing labelling L_{ab} and $1/3$ of the points in cluster A to C, producing L_{ac}. Denote the original labels by Ori.

In the second column of Table 2, different labelling on DS_1 are compared with the original labels using VI, $Rand$ and our second CSS measure. From the normalized results of VI and $Rand$, we can see that both VI and $Rand$ produce the same distance (similarity) for L_{ac}/Ori and L_{ac}/Ori cases, which intuitively should be different. Such a phe-



(a) EM DS_2 (b) CEM DS_2 (c) EM $Sonar$ (d) CEM $Sonar$

Figure 6. EM v.s. CEM on DS_1 and DS_2

nomenon, as discussed earlier in this paper, results from the lack of cluster similarity differentiation in both measures. By contrast, our CSS measurement gives larger distance to L_{ac} because cluster C is farther than cluster B, which is a more plausible comparison[1].

	L_{ab}/Ori.	L_{ac}/Ori.	KM/Ori.	DB/Ori
VI	0.202	0.202	0.186	0.263
$Rand$	0.881	0.881	0.931	0.761
$CSS_{/10^2}$	0.287	0.658	0.0841	0.564

Table 2: Hard clustering comparing on DS_1

We also run our CSS measure to compare clusterings generated by DBSCAN and K-Means. As illustrated in the rightmost column of Table 2, three clustering comparison methods all conclude the K-Means performs better than DBSCAN on DS_1.

	Uni-Label/Ori.	DB/Ori.	KM/Ori.
VI	0.58	0.59	0.48
$Rand$	0.57	0.53	0.39
$CSS_{/10^2}$	0.06	0.058	1.19

Table 3: Hard clustering comparing on DS_2

For DS_2, we compare the original labelling with three labellings produced manually and by clustering algorithms.

Firstly, we compare Ori. with that all points are labelled as in the same cluster. As explained earlier such two labelling should intuitively be similar given they may be produced by different settings of clustering parameters. Accordingly, CSS yields relatively low distance for this case while both VI and $Rand$ differentiate them unfairly. Secondly, we test with real clustering algorithm DBSCAN by setting low threshold, as illustrated in Figure 2. Similarly, CSS outperforms VI and $Rand$. Finally, we compare Ori. with labels produced by K-Means. In the K-Means results (KM), many points from the central are labelled to the outskirt, shrinking the size of primary cluster in the center. This sheerly conflicts with the tendency of data distribution, where outskirt points are generated from the center. With this comparison, we expect a large distance between KM and Ori. However, VI returns relatively small distance. CSS and $Rand$ detect the inappropriateness of KM on DS_2 by giving the large distance or small similarity.

For real world datasets, which are mostly high-dimensional, we plot the normalized VI, $Rand$ results and scaled CSS wrt. each dataset in Figure 7(b). We can see that CSS differs from VI and $Rand$ in most real world cases by considering cluster similarities.

5.3. Computational complexity

In this section, we will examine the computational complexity of the two measures we propose.

5.3.1. Analytic complexity

Computing Mallows distance is a special case of the discrete mass transport problem. The complexity for the transport problem is the same as for the minimum cost flow problem. As for the minimum cost flow problem, the best complexity is due to (Orlin, 1988): Given two clusterings on N data points, each producing J and K clusters respectively, the complexity for setting up the optimization problem is linear in N: $O(N)$. The solution to a transportation problem corresponds to a bipartite graph and the worst case complexity is $O((J*K)^2 log(J+K) + J*K(J+K)(log(J+K))^2)$. On average, the algorithm runs much faster, and there are also many faster heuristic methods for finding an approximate solution for large problems.

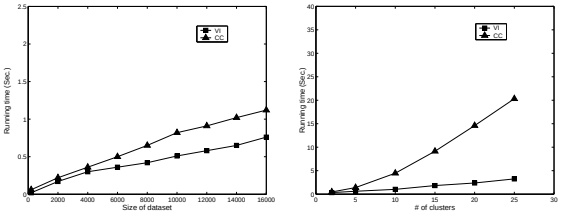
5.3.2. Empirical evaluation

The computational complexity of both CC and CSS are evaluated empirically in terms of both the scale of datasets

[1]Note that the inter-cluster distances for various clusterings are unpredictable. Therefore it becomes an open problem to normalize our CSS measure. Nevertheless, in order to better compare the quantities under CSS measure, we divide all CSS results by 10^2.

and the number of clusters. We compare both metrics with VI and $Rand$.

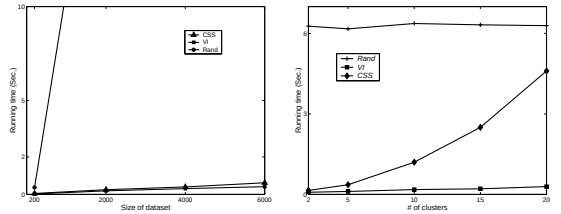
For soft clustering comparison, we first generate a set of two soft labelling matrices for datasets sized from 200 to 16000. In Figure 8(a), we can see both VI and CC scale very well to the number of data points. The computational complexity to number of clusters are experimented in Figure 8(b). The setting of datasets is sized 5000 varying from from 2 to 25 clusters. Comparatively, CC is slightly more sensitive to the increase of cluster number.



(a) Runtime wrt. data size (b) Runtime wrt. cluster #

Figure 8. Runtime for soft clustering comparison

For hard clustering comparison, we use two sets of clustering results, one of various numbers of data points from 200 to 6000, the other containing 2 to 20 clusters. Obviously, it is illustrated in Figure 9(a) that the computation of CSS largely advances $Rand$ in terms of dataset size. The runtime of $Rand$ is relatively stable to the increase of cluster number. However, CSS still outperforms $Rand$ with large number of clusters in Figure 9(b).



(a) Runtime wrt. data size (b) Runtime wrt. cluster #

Figure 9. Runtime for hard clustering comparison

6. Conclusions

This paper addresses the need for clustering comparison with similarity differentiation. By introducing the ideas of data distribution assessment from statistics into the clustering comparison, we proposed two metrics to compare clustering results, which are positive, symmetric and transitive. Our metrics consider cluster similarity differences and is also able to handle soft clustering comparison. The implementation of our comparison measures relies on linear programming.

References

A. Ben-Hur, A. Elisseeff and I. Guyon, *A stability based method for discovering structure in clustered data*, In Pacific Symposium on Biocomputing, pages 6-17, 2002.

A. Dempster, N. Laird, and D. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, Series B, 39(1):138, 1977.

S. Dongen, *Performance criteria for graph clustering and Markov cluster experiments*, Technical Report INS-R0012, Centrum voor Wiskundeen Informatica, 2000.

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 KDD, 1996.

E.B.Fowlkes and C.L.Mallows, *A method for comparing two hierarchical clusterings*, Journal of the American Statistical Association, 78(383):553-569,1983.

L. Hubert and P. Arabie, *Comparing partitions*, Journal of Classification, 2:193-218, 1985.

E. Levina and P. Bickel, *The Earth Mover's distance is the Mallows distance: some insights from statistics*, In proceedings of IEEE International Conference on Computer Vision, Vol. 2, pages 251-256. Vancouver, BC, Canada, 2001.

C.L. Mallows, *A note on asymptotic joint normality*, Annals of Mathematical Statistics, 43(2): 508-515, 1972.

M. Meila, *Comparing Clusterings*, Technical Report, Statistics, University of Washington, 2002.

J.B. Orlin, *A faster strongly polynomial minimim cost flow algorithm*, Proc. 20th ACM Symposium on the Theory of Computing, 1988.

S.T. Rachev, *The Monge-Kantorovich mass transference problem and its stochastic applications*, Theory of Probability and its Applications, 29:647-676, 1984.

W.M. Rand, *Objective criteria for the evaluation of clustering methods*, Journal of the American Statistical Association, 66:846-850, 1971.

C. Robardet and F. Feschet, *A new methodology to compare clustering algorithms*, In proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning, HK, China, 2000.

Y. Rubner, C. Tomasi and L.J. Guibas, *A metric for distribution with applications to image databases*, In proceedings of IEEE International Conference on Computer Vision, pages 59-66. Bombay, India, 1998.