# Exploring the Performance Limit of Cluster Ensemble Techniques

Xiaoyi Jiang and Daniel Abdala*

Department of Mathematics and Computer Science, University of Münster
Einsteinstrasse 62, D-48149 Münster, Germany
{xjiang,abdalad}@math.uni-muenster.de

**Abstract.** Cluster ensemble techniques are a means for boosting the clustering performance. However, many cluster ensemble methods are faced with high computational complexity. Indeed, the median partition methods are $\mathcal{NP}$-complete. While a variety of approximative approaches for suboptimal solutions have been proposed in the literature, the performance evaluation is typically done by means of ground truth. In contrast this work explores the question how well the cluster ensemble methods perform in an absolute sense *without ground truth*, i.e. how they compare to the (unknown) optimal solution. We present a study of applying and extending a lower bound as an attempt to answer the question. In particular, we demonstrate the tightness of the lower bound, which indicates that there exists no more room for further improvement (for the particular data set at hand). The lower bound can thus be considered as a means of exploring the performance limit of cluster ensemble techniques.

## 1 Introduction

Clustering, or finding partitions[1], of data is a fundamental task in multivariate data analysis. It receives increasingly importance due to the ever increasing amount of data. A large variety of clustering algorithms [20] have been proposed in the past. A recent development is constrained clustering [4], which accommodates additional information or domain knowledge. Cluster ensemble techniques provide another means for boosting the clustering performance.

Motivated by the success of multiple classifier systems, the idea of combining different clustering results emerged. Given a data set, a cluster ensemble technique consists of two principal steps: ensemble generation and consensus computation. In the first step, an ensemble (with sufficient diversity) is computed. For this purpose different clustering algorithms or the same algorithm with varying parameter settings can be applied. Other options include the use of different subsets of features and projection of the data into different subspaces. The main

* Daniel D. Abdala thanks the CNPq, Brazil-Brasilia for granting him a Ph.D. scholarship under the process number 290101-2006-9.

[1] Recently, efforts have been undertaken to go beyond the traditional understanding of clustering as partitions, i.e. [16]. This is, however, not the focus of this work.

E.R. Hancock et al. (Eds.): SSPR & SPR 2010, LNCS 6218, pp. 405–414, 2010.

challenge of cluster ensemble techniques lies in an appropriate way of computing a final clustering, which disagrees least overall with the input ensemble.

There exist two main approaches for consensus computation: co-occurrence based and median partition methods. The fundamental assumption of co-occurrence based methods is that patterns belonging to a "natural" cluster are very likely to be co-located in the same cluster in different data partitions. Therefore, a matrix with such co-location information can serve as plausibility values that two patterns should be clustered together. Typically, a subsequent step based on this matrix is designed to compute a final clustering; see for instance the evidence accumulation method [6]. Median partition methods are based on an optimization formulation of consensus computation; see for instance [17]. Since this optimization problem is typically $\mathcal{NP}$-complete [3], various suboptimal solutions have been proposed.

The focus of this work is performance assessment of cluster ensemble techniques *without using any ground truth information*. In the literature the experimental validation is typically done by means of ground truth. In contrast we explore the question how well the cluster ensemble methods perform in an absolute sense without ground truth, i.e. how they compare to the (unknown) optimal solution. This paper presents a study of applying and extending the lower bound presented in [11] as an attempt to answer the question.

## 2 Problem Statement

Given the data set $X = \{x_1, x_2, \ldots, x_n\}$ of $n$ patterns $x_i$, a cluster ensemble is a set $P = \{P_1, P_2, \ldots, P_N\}$, where $P_i$ is a clustering of $X$. We denote the set of all possible clusterings of $X$ by $\mathcal{P}_X$ ($P \subset \mathcal{P}_X$). The goal of cluster ensemble techniques is to find a consensus clustering $P^* \in \mathcal{P}_X$, which optimally represents the ensemble $P$.

In median partition methods this optimality is formulated as:

$$P^* = \arg\min_{P \in \mathcal{P}_X} \sum_{i=1}^{N} d(P, P_i)$$

where $d()$ is a distance (dissimilarity) function between two clusterings. Note that this definition is a special instance of the so-called generalized median problem, which has been intensively investigated in structural pattern recognition, see [12,10] for the case of strings and graphs.

The median partition problem has been proven to be $\mathcal{NP}$-complete [3]. An exhaustive search in $\mathcal{P}_X$ is computationally intractable. In practice suboptimal approaches [14,17] are thus developed to solve the optimization problem.

Given a suboptimal solution $\tilde{P} \in \mathcal{P}_X$, however, the question of its accuracy arises. In [11] a lower bound is proposed to answer this question (for the general case of generalized median problems). For an approximate solution $\tilde{P}$ the following relationship holds:

$$\mathrm{SOD}(\tilde{P}) \;=\; \sum_{i=1}^{N} d(\tilde{P}, P_i) \;\geq\; \sum_{i=1}^{N} d(P^*, P_i) \;=\; \mathrm{SOD}(P^*)$$

where SOD stands for sum of distances. The quality of $\tilde{P}$ can be absolutely measured by the difference $\mathrm{SOD}(\tilde{P}) - \mathrm{SOD}(P^*)$. Since $P^*$ and $\mathrm{SOD}(P^*)$ are unknown in general, we resort to a lower bound $\Gamma$ with

$$0 \;\leq\; \Gamma \;\leq\; SOD(P^*) \;\leq\; SOD(\tilde{P})$$

and measure the quality of $\tilde{P}$ by $\mathrm{SOD}(\tilde{P}) - \Gamma$ instead. Obviously, the trivial lower bound $\Gamma = 0$ is useless. We require $\Gamma$ to be as close to $\mathrm{SOD}(P^*)$ as possible.

In [11] a lower bound based on linear programming is proposed for metric spaces. Assuming a metric distance function $d()$, the lower bound for the median partition problem is specified by the solution $\Gamma$ of the following linear program:

$$\begin{array}{l} \text{minimize } x_1 + x_2 + \cdots + x_N \text{ subject to} \\ \forall i, j \in \{1, 2, \ldots, N\},\ i \neq j, \ \begin{cases} x_i + x_j \;\geq\; d(P_i, P_j) \\ x_i + d(P_i, P_j) \;\geq\; x_j \\ x_j + d(P_i, P_j) \;\geq\; x_i \end{cases} \\ \forall i \in \{1, 2, \ldots, N\},\ x_i \geq 0 \end{array}$$

Given a suboptimal solution $\tilde{P}$ and the computed lower bound, the deviation $\Delta = \mathrm{SOD}(\tilde{P}) - \Gamma$ can thus give a hint of the absolute accuracy of $\tilde{P}$. In particular, if $\Delta \approx 0$, then it can be safely claimed that there is hardly room for further improvement (for the particular data set at hand).

In this paper we present a study of the lower bound $\Gamma$ using two cluster ensemble methods and eleven data sets. Among others it will be demonstrated that this lower bound can (almost) be reached by the computed solution. This tightness indicates the limited room for further improvement. Therefore, the lower bound $\Gamma$ represents a means of exploring the performance limit of cluster ensemble techniques.

The remainder of this paper is organized as follows. Section 3 describes the experimental settings of our study. The experimental results are presented in Section 4. Later in Section 5 the study is extended to deal with weighted cluster ensemble techniques. Finally, some further discussions conclude this paper.

## 3 Experimental Settings

In this section we give the details of designing our study: Metric distance functions, cluster ensemble methods, and data sets used in the experiments and the test protocol.

### 3.1 Metric Distance Functions

Many distance functions have been suggested to measure the dissimilarity of two partitions of the same data set; see [13] for a detailed discussion. For our study the following three were selected, which are provably metric.

**Variance of information:** This metric is an information-theoretic one. Given two partitions $P$ and $Q$ of $X$, it is defined by

$$d_{vi}(P, Q) = H(P) + H(Q) - 2I(P, Q)$$

where $H(P)$ and $H(Q)$ are the entropy of $P$ and $Q$, respectively, and $I(P, Q)$ represents the mutual information of $P$ and $Q$; see [13] for a proof of the metric property.

**van Dongen metric:** Fundamental to this distance function [18] is a (non-optimal) matching of the two sets of clusters.

$$d_{vd}(P, Q) = 1 - \frac{1}{2n} \cdot \Big( \sum_{C_p \in P} \max_{C_q \in Q} |C_p \cap C_q| + \sum_{C_q \in Q} \max_{C_p \in P} |C_q \cap C_p| \Big)$$

**Mirkin metric:** Let $a$ equal to the number of pairs of patterns co-clustered in $P$ but not in $Q$ and $b$ equal to the number of pairs of patterns co-clustered in $Q$ but not in $P$. Then, the Mirkin metric belongs to the class of distance functions based on counting pairs and is simply defined by $d_m(P, Q) = a + b$. A proof of the metric property can be found in [7].

### 3.2 Cluster Ensemble Methods

We used two cluster ensemble methods in our experiments. The first one is the evidence accumulation method [6]. It computes the co-occurrence matrix, which is interpreted as a new similarity measure between the patterns. The consensus partition is then obtained by using a hierarchical clustering algorithm. We report the results based on the average-linkage variant (`EAC-AL`) only, since it mostly outperforms the single-linkage variant. The second cluster ensemble method (`RW`) is based on the co-occurrence matrix as well. But it adapts a random walker segmentation algorithm to produce a final clustering [1].

### 3.3 Data Sets

For our experiments we used two data sources. Nine UCI data sets [2] as summarized in Table 1. Special remarks need to be made about the Mammographic Mass (Mammo) and the Optical Recognition of Handwritten Digits (Optic) data sets. For Mammo, all patterns with missing values were removed, reducing this way the number of patterns from 961 to 830. For the Optic data set we extracted a subset of the first 100 patterns of each digit, producing a subset of 1000 patterns.

Two artificial data sets from [17] were included into our experiments. The first data set (2D2K) contains 500 2D points from two Gaussian clusters and the second data set (8D5K) contains 1000 points from five multivariate Gaussian distributions (200 points each) in 8D space.

**Table 1.** Summary of test data sets

| Data set | $n$ | # attributes | # clusters |
|---|---|---|---|
| Iris | 150 | 4 | 3 |
| Wine | 178 | 13 | 3 |
| Breast | 683 | 9 | 2 |
| Optic | 1000 | 64 | 10 |
| Soy | 47 | 35 | 2 |
| Glass | 218 | 9 | 7 |
| Haberman | 306 | 3 | 2 |
| Mammo | 830 | 5 | 2 |
| Yeast | 1484 | 8 | 10 |
| 2D2K | 500 | 2 | 2 |
| 8D5K | 1000 | 8 | 5 |

### 3.4 Test Protocol

Given an ensemble $P$, we compute a final clustering $\tilde{P}$ using either `EAC-AL` or `RW`. The following measures are used to characterize the performance: $\mathrm{SOD}(\tilde{P})$, the lower bound $\Gamma$ (for the ensemble), and the deviation

$$\Delta' = (\mathrm{SOD}(\tilde{P}) - \Gamma)/\mathrm{SOD}(\tilde{P})$$

(in percentage). For each data set, this procedure is repeated ten times (i.e. ten different ensembles) and the average measures are reported.

## 4 Experimental Results and Discussions

For the two cluster ensemble methods `EAC-AL` and `RW` the performance measures are shown in Table 2. The deviation $\Delta'$ can be interpreted as the potential of further improvement. For three data sets (Haberman, Mammo, and 2D2K) $\mathrm{SOD}(\tilde{P})$ almost reaches the lower bound $\Gamma$ for all three distance functions, indicating practically no room for improvement. To some extent the same applies to the data set Soy and 8D5K in conjunction with `EAC-AL`. In these cases the lower bound turns out to be extremely tight. On the other hand, if the deviation is large, we must be careful in making any claims. The large deviation may be caused by two reasons: The lower bound is not tight enough in that particular case or the computed solution $\tilde{P}$ is still far away from the (unknown) optimal solution $P^*$.

The second case is certainly more delicate. But we may interpret as of some, although uncertain, potential of further improvement. Given such an ensemble, we could generate more ensembles and compute additional candidates for consensus clustering. The measure SOD can then be used for selecting a final solution. This strategy has been suggested in [17] (although in a different context): "Our objective function has the added advantage that it allows one to add a stage that selects the best consensus function without any supervisory information,

**Table 2.** Deviation $\Delta'$

Evidence accumulation method `EAC-AL`

| | $d_{vi}$ | | | $d_{vd}$ | | | $d_m$ | | |
|---|---|---|---|---|---|---|---|---|---|
| dataset | SOD($\tilde{P}$) | $\Gamma$ | $\Delta'$(%) | SOD($\tilde{P}$) | $\Gamma$ | $\Delta'$(%) | SOD($\tilde{P}$) | $\Gamma$ | $\Delta'$(%) |
| Iris | 8.22 | 7.24 | 12.0 | 2.26 | 2.16 | 4.3 | 27621 | 25113 | 9.1 |
| Wine | 2.01 | 1.86 | 7.7 | 0.35 | 0.33 | 5.1 | 7232 | 6777 | 6.3 |
| Breast | 1.16 | 1.08 | 7.3 | 0.16 | 0.15 | 3.8 | 71244 | 68392 | 4.0 |
| Optic | 7.50 | 6.37 | 15.0 | 2.06 | 1.85 | 10.0 | 378439 | 315016 | 16.8 |
| Soy | 3.90 | 3.79 | 2.9 | 1.65 | 1.62 | 1.9 | 1616 | 1591 | 1.6 |
| Glass | 5.20 | 4.66 | 10.4 | 1.37 | 1.24 | 9.4 | 39909 | 33939 | 15.8 |
| Haberman | 7.60 | 7.58 | 0.3 | 2.84 | 2.84 | 0.0 | 233417 | 232994 | 0.2 |
| Mammo | 1.77 | 1.77 | 0.0 | 0.38 | 0.38 | 0.0 | 248649 | 248649 | 0.0 |
| Yeast | 13.94 | 11.40 | 18.3 | 3.85 | 3.34 | 13.4 | 3512666 | 3010184 | 14.3 |
| 2D2K | 4.86 | 4.69 | 3.0 | 1.18 | 1.15 | 3.0 | 1037580 | 978050 | 5.7 |
| 8D5K | 4.97 | 4.91 | 1.8 | 1.69 | 1.66 | 2.0 | 585462 | 579262 | 1.1 |

Random walker based method `RW`

| | $d_{vi}$ | | | $d_{vd}$ | | | $d_m$ | | |
|---|---|---|---|---|---|---|---|---|---|
| dataset | SOD($\tilde{P}$) | $\Gamma$ | $\Delta'$(%) | SOD($\tilde{P}$) | $\Gamma$ | $\Delta'$(%) | SOD($\tilde{P}$) | $\Gamma$ | $\Delta'$(%) |
| Iris | 8.40 | 7.24 | 13.8 | 2.28 | 2.16 | 5.2 | 28067 | 25113 | 10.5 |
| Wine | 2.09 | 1.86 | 10.0 | 0.35 | 0.33 | 4.5 | 7242 | 6777 | 5.8 |
| Breast | 1.49 | 1.08 | 27.7 | 0.20 | 0.15 | 23.9 | 90032 | 68392 | 24.0 |
| Optic | 11.38 | 6.37 | 44.0 | 3.90 | 1.85 | 50.9 | 749459 | 315016 | 57.7 |
| Soy | 6.19 | 3.79 | 36.9 | 4.08 | 1.62 | 52.0 | 3433 | 1591 | 49.3 |
| Glass | 7.96 | 4.66 | 41.1 | 2.53 | 1.24 | 45.9 | 69186 | 33940 | 49.3 |
| Haberman | 7.70 | 7.58 | 1.5 | 2.86 | 2.84 | 0.7 | 234484 | 232995 | 0.6 |
| Mammo | 1.77 | 1.77 | 0.0 | 0.38 | 0.38 | 0.0 | 248650 | 248650 | 0.0 |
| Yeast | 18.60 | 11.40 | 38.2 | 10.51 | 3.34 | 67.5 | 6606869 | 3010185 | 53.4 |
| 2D2K | 4.69 | 4.69 | 0.0 | 1.15 | 1.15 | 0.0 | 978050 | 978050 | 0.0 |
| 8D5K | 5.24 | 4.91 | 5.9 | 2.43 | 1.66 | 15.0 | 721412 | 579262 | 11.3 |

by simply selecting the one with the highest ANMI" (ANMI is the particular SOD used in that work). In doing so, a tight lower bound may give us a hint to continue or terminate the procedure without any knowledge of ground truth.

There is also the issue of inconsistency among different distance functions. Sometimes it happens that the deviation values for two distance functions vary, partly substantially. This observation is not really surprising. Different distance functions may not share the same view of dissimilarity, thus the quality of a consensus clustering. It is up to the user to decide which distance function is more suitable for a particular data clustering task.

Finally, we want to point out that the two cluster ensembles methods used in our study do not belong the class of median partition techniques. But even in this case the lower bound still provides useful information about the optimality of the computed consensus clustering.

**Table 3.** Comparison of lower bounds $\Gamma$ and $\Gamma_m$

| dataset | $\Gamma$ | $\Gamma_m$ | $(\Gamma_m - \Gamma)/\Gamma(\%)$ |
|---|---|---|---|
| Iris | 25113 | 26377 | 5.0 |
| Wine | 6777 | 6820 | 0.6 |
| Breast | 68392 | 71196 | 4.1 |
| Optic | 315016 | 335678 | 6.6 |
| Soy | 1591 | 1599 | 0.5 |
| Glass | 33940 | 34513 | 1.7 |
| Haberman | 232995 | 233273 | 0.1 |
| Mammo | 248650 | 248650 | 0.0 |
| Yeast | 3010185 | 3224160 | 7.1 |
| 2D2K | 978050 | 1168728 | 8.4 |
| 8D5K | 579262 | 584848 | 1.0 |

**Special case $d_m$:** The cluster ensemble problem with Merkin distance $d_m$ has been intensively investigated [7,8]. This is mainly due to the simplicity of $d_m$, which allows to obtain deep insight into this particular consensus clustering problem. In particular, several suboptimal algorithms have been proposed with known approximation factor. In addition, a lower bound specific to $d_m$ only can be defined:

$$\Gamma_m \;=\; \sum_{i<j} \min\Big(\sum_{k=1}^{N} X_{ij}^{(k)},\; N - \sum_{k=1}^{N} X_{ij}^{(k)}\Big)$$

where $X_{ij}^{(k)}$ is the Bernoulli random variable as 1 if $x_i$ and $x_j$ are co-clustered in partition $P_k$ and 0 otherwise. $\Gamma_m$ takes the specific properties of $d_m$ into account, whereas $\Gamma$ is based on the general properties of a metric only. $\Gamma_m$ is thus better informed and expected to be tighter than $\Gamma$. In Table 3 we compare the closedness of the two lower bounds. It is remarkable that without any knowledge of $d_m$ and using the metric properties alone, the general lower bound $\Gamma$ almost reaches $\Gamma_m$.

## 5 Extension to Weighted Cluster Ensemble Techniques

Cluster ensembles techniques can be extended by assigning a weight $w_i$ to each involved partition $P_i$, which represents the estimated relative merit of the partitions. In [19], for instance, four weights are considered: inter-cluster distance, intra-cluster distance, mean size of clusters, and difference between the cluster sizes. Then, the weighted median partition problem can be stated as:

$$P^* \;=\; \arg\min_{P \in \mathcal{P}_X} \sum_{i=1}^{N} w_i \cdot d(P, P_i)$$

Here we assume that smaller weights mean favorable partitions. The extension of the linear program lower bound $\Gamma$ to deal with the weighted cluster ensemble problem is straightforward, resulting in a lower bound $\Gamma_w$.

$$\text{minimize } w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_N \cdot x_N \text{ subject to}$$

$$\forall i, j \in \{1, 2, \ldots, N\},\ i \neq j,\ \begin{cases} x_i + x_j \ \geq \ d(P_i, P_j) \\ x_i + d(P_i, P_j) \ \geq \ x_j \\ x_j + d(P_i, P_j) \ \geq \ x_i \end{cases}$$

$$\forall i \in \{1, 2, \ldots, N\},\ x_i \geq 0$$

Many cluster ensembles methods can be easily extended to integrate such weights. In co-occurrence based techniques such as `EAC-AL` and `RW` this can be done when computing the co-occurrence matrix. In our case we have used the inter-cluster distance as weights only.

For these weighted algorithms the performance measures are shown in Table 4. Compared to the unweighted results in Table 2 the things have not changed

**Table 4.** Deviation $\Delta'$ (weighted versions)

Weighted evidence accumulation method `EAC-AL`

| | $d_m$ | | | $d_{vd}$ | | | $d_{vi}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| dataset | SOD($\tilde{P}$) | $\Gamma_w$ | $\Delta'$(%) | SOD($\tilde{P}$) | $\Gamma_w$ | $\Delta'$(%) | SOD($\tilde{P}$) | $\Gamma_w$ | $\Delta'$(%) |
| Iris | 0.78 | 0.68 | 12.1 | 0.21 | 0.12 | 4.5 | 2599 | 2356 | 9.2 |
| Wine | 0.20 | 0.19 | 7.5 | 0.04 | 0.03 | 5.0 | 723 | 678 | 6.2 |
| Breast | 0.12 | 0.11 | 7.3 | 0.02 | 0.02 | 3.8 | 7119 | 6834 | 4.0 |
| Optic | 0.75 | 0.64 | 14.7 | 0.21 | 0.19 | 9.7 | 36742 | 31492 | 13.9 |
| Soy | 0.39 | 0.38 | 2.2 | 0.16 | 0.16 | 1.4 | 160 | 158 | 1.2 |
| Glass | 0.52 | 0.47 | 10.5 | 0.14 | 0.12 | 9.6 | 3996 | 3423 | 12.5 |
| Haberman | 0.77 | 0.76 | 1.5 | 0.29 | 0.29 | 0.8 | 23754 | 23303 | 1.9 |
| Mammo | 0.17 | 0.17 | 0.0 | 0.04 | 0.04 | 0.0 | 23794 | 23794 | 0.0 |
| Yeast | 1.40 | 1.14 | 18.4 | 0.38 | 0.33 | 13.2 | 353189 | 299571 | 15.0 |
| 2D2K | 0.52 | 0.52 | 0.0 | 0.13 | 0.13 | 0.0 | 107322 | 107834 | 0.0 |
| 8D5K | 0.49 | 0.48 | 1.3 | 0.16 | 0.16 | 1.6 | 56825 | 56218 | 1.0 |

Weighted random walker based method `RW`

| | $d_m$ | | | $d_{vd}$ | | | $d_{vi}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| dataset | SOD($\tilde{P}$) | $\Gamma_w$ | $\Delta'$(%) | SOD($\tilde{P}$) | $\Gamma_w$ | $\Delta'$(%) | SOD($\tilde{P}$) | $\Gamma_w$ | $\Delta'$(%) |
| Iris | 0.81 | 0.68 | 16.0 | 0.22 | 0.20 | 9.1 | 2753 | 2356 | 14.4 |
| Wine | 0.64 | 0.19 | 70.8 | 0.11 | 0.03 | 70.8 | 2303 | 677 | 70.6 |
| Breast | 0.22 | 0.11 | 50.6 | 0.03 | 0.02 | 50.5 | 13819 | 6834 | 50.5 |
| Optic | 1.12 | 0.64 | 43.0 | 0.36 | 0.19 | 46.0 | 55409 | 31492 | 42.2 |
| Soy | 0.52 | 0.38 | 25.5 | 0.39 | 0.16 | 47.6 | 307 | 157 | 42.3 |
| Glass | 0.85 | 0.47 | 44.7 | 0.30 | 0.13 | 51.9 | 6436 | 3422 | 42.5 |
| Haberman | 0.80 | 0.76 | 4.3 | 0.29 | 0.29 | 1.4 | 24101 | 23303 | 3.3 |
| Mammo | 0.17 | 0.17 | 0.0 | 0.04 | 0.04 | 0.0 | 23794 | 23794 | 0.0 |
| Yeast | 1.85 | 1.14 | 38.8 | 1.02 | 0.33 | 66.7 | 511552 | 299571 | 40.8 |
| 2D2K | 0.52 | 0.52 | 0.9 | 0.13 | 0.13 | 0.9 | 108495 | 107833 | 0.5 |
| 8D5K | 0.52 | 0.48 | 5.8 | 0.24 | 0.16 | 15.0 | 70603 | 56218 | 11.3 |

much. For the three data sets Haberman, Mammo, and 2D2K, SOD($\tilde{P}$) again almost reaches the lower bound $\Gamma_w$ for all three distance functions, indicating practically no room for further improvement. In conjunction with `EAC-AL` the same can be said for the data set 8D5K. In these cases the lower bound turns out to be extremely tight. On the other hand, if the deviation is larger, we must be careful in making any claims. Also here we can take the deviation as a hint for continuing optimization.

## 6 Discussions and Conclusion

In this paper we have presented a study of the lower bound $\Gamma$ using eleven data sets. It could be shown:

- In some cases this lower bound can (almost) be reached by the computed solution. This tightness implies that there exists no more room for further improvement for this particular data set (with respect to the used distance function). Larger deviation may indicate some, although uncertain, potential of improvement and thus serves as a hint for continuing optimization.
- The same observation can be made also for weighted version of cluster ensemble methods.
- The tightness of $\Gamma$ can be even demonstrated in case of Merkin distance $d_m$ by comparing with another lower bound, which is derived from the special nature of $d_m$.

Based on these facts we consider the lower bound $\Gamma$ (and $\Gamma_m$ in case of $d_m$) a means of exploring the performance limit of cluster ensemble techniques.

The lower bound defined in [11] presumes a metric distance function $d()$. The triangle inequality of a metric excludes cases in which $d(P, R)$ and $d(R, Q)$ are both small, but $d(P, Q)$ is very large. In practice, however, there may exist distance functions which do not satisfy the triangle inequality. The work [5] extends the concept of metrics to a relaxed triangle inequality. Instead of the strict triangle inequality, the relation:

$$d(P, R) + d(R, Q) \;\geq\; \frac{d(P, Q)}{1 + \varepsilon}$$

is required, where $\varepsilon$ is a small nonnegative constant. This is also called quasi-metric in mathematics [9]. As long as $\varepsilon$ is not very large, the relaxed triangle inequality still retains the human intuition of similarity. Note that the strict triangle inequality is a special case with $\varepsilon = 0$. The lower bound $\Gamma$ can be easily extended to quasi-metric distance functions by changing the inequalities in the linear program accordingly. This extended lower bound can be expected to be useful in working with cluster ensemble methods based on quasi-metrics.

## References


1. Abdala, D.D., Wattuya, P., Jiang, X.: Ensemble clustering via random walker consensus strategy. In: Proc. of ICPR, Istanbul (2010)
2. Asuncion, A., Newman, D.J.: UCI machine learning repository (2010)
3. Barthelemy, J.P., Leclerc, B.: The median procedure for partition. In: Partitioning Data Sets. AMS DIMACS Series in Discrete Mathematics, pp. 3–34 (1995)
4. Basu, S., Davidson, I., Wagstaff, K.L. (eds.): Constrained Clustering: Advances in Algorithms, Theory, and Applications. CRC Press, Boca Raton (2009)
5. Fagin, R., Stockmeyer, L.: Relaxing the triangle inequality in pattern matching. Int. Journal on Computer Vision 28(3), 219–231 (1998)
6. Fred, A., Jain, A.K.: Combining multiple clusterings using evidence accumulation. IEEE Trans. on PAMI 27(6), 835–850 (2005)
7. Gionis, A., Mannila, H., Tsapara, P.: Clustering Aggregation. ACM Trans. on Knowledge Discovery from Data 1(1) (2007)
8. Goder, A., Filkov, V.: Consensus clustering algorithms: Comparison and refinement. In: Proc. of Workshop on Algorithm Engineering and Experiments, pp. 109–117 (2008)
9. Heinonen, J.: Lectures on Analysis on Metric Spaces. Springer, New York (2001)
10. Jiang, X., Münger, A., Bunke, H.: On median graphs: Properties, algorithms, and applications. IEEE Trans. on PAMI 23(10), 1144–1151 (2001)
11. Jiang, X., Bunke, H.: Optimal lower bound for generalized median problems in metric space. In: Caelli, T., Amin, A., Duin, R.P.W., Kamel, M., de Ridder, D. (eds.) SPR 2002 and SSPR 2002. LNCS, vol. 2396, pp. 143–151. Springer, Heidelberg (2002)
12. Lopresti, D., Zhou, J.: Using consensus sequence voting to correct OCR errors. Computer Vision and Image Understanding 67(1), 39–47 (1997)
13. Meila, M.: Comparing clusterings - an information based distance. Journal of Multivariate Analysis 98(5), 873–895 (2007)
14. Luo, H., Jing, F., Xie, X.: Combining multiple clusterings using information theory based genetic algorithm. In: Proc. of Int. Conf. on Computational Intelligence and Security, pp. 84–89 (2006)
15. Mirkin, B.G.: Mathematical Classification and Clustering. Kluwer Academic Press, Dordrecht (1996)
16. Pelillo, M.: What is a Cluster? Perspectives from Game Theory. In: NIPS Workshop on "Clustering: Science of Art?" (2009)
17. Strehl, A., Ghosh, J.: Cluster ensembles – a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3, 583–617 (2002)
18. van Dongen, S.: Performance criteria for graph clustering and Markov cluster experiments. Technical Report INSR0012, Centrum voorWiskunde en Informatica (2000)
19. Vega-Pons, S., Correa-Morris, J., Ruiz-Shulcloper, J.: Weighted cluster ensemble using a kernel consensus function. In: Ruiz-Shulcloper, J., Kropatsch, W. (eds.) CIARP 2008. LNCS, vol. 5197, pp. 195–202. Springer, Heidelberg (2008)
20. Xu, R., Wunsch II, D.: Survey of clustering algorithms. IEEE Trans. on Neural Networks 16(3), 645–678 (2005)