ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 3, pp. 262–273.
© The Author(s), 2024 published by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2024.
Russian Text © The Author(s), 2024, published in Avtomatika i Telemekhanika, 2024, No. 3, pp. 6–22.

TOPICAL ISSUE

Self-Adjusted Consensus Clustering with Agglomerate Algorithms

B. G. Mirkin[*,**,a] **and A. A. Parinov**[*,b]

[*] *National Research University Higher School of Economics, Moscow, Russia*
Birkbeck University of London, United Kingdom
e-mail: [a] *bmirkin@hse.ru,* [b] *aparinov@hse.ru*

Received July 8, 2023
Revised October 21, 2023
Accepted January 20, 2024

Abstract—This paper reports of theoretical and computational results related to an original concept of consensus clustering involving what we call the projective distance between partitions. This distance is defined as the squared difference between a partition incidence matrix and its image over the orthogonal projection in the linear space spanning the other partition incidence matrix. It appears, provided that the ensemble clustering is of a sufficient size, agglomerate clustering with the semi-average within-cluster similarity criterion effectively solves the problem of consensus partition and, moreover, of the number of clusters in it.

Keywords: consensus clustering, agglomerate clustering, consensus matrix, semi-average criterion, shifting data

DOI: 10.31857/S0005117924030022

1. INTRODUCTION

Methods for consensus partitioning are relevant to the network research, as pointed out by Lancichinetti and Fortunato [8] and Liu et al. [9]. The goal of this paper is to present and justify a self-adjusting method for finding a consensus partition over a partition ensemble. The method is, basically, conventional locally optimal agglomeration of clusters over the consensus matrix supplemented with several traits. The traits include: (i) shifting the similarity values to a zero sum; (ii) the semi-average within-cluster criterion; (iii) zeroing the diagonal similarities at each agglomeration step. The semi-average clustering criterion appears to implement the concept of consensus partitioning using what we call the projective distance between partitions. This distance is defined as the squared difference between a partition incidence matrix and its image over the orthogonal projection in the linear space spanning the other partition's incidence matrix [11, 12]. The experimental part is based on a novel synthetic generator of partition ensembles. The generator involves a "mutation" probability which controls both the diversity of generated partitions and their relation to a ground truth partition. Methods for maximizing within-cluster summary criterion, including most popular modularity clustering method [14] and algorithm Louvain [2], are used as a benchmark.

2. CONSENSUS MATRIX AND ITS SHIFTING

Given a number of partitions $R_1, R_2, \ldots, R_M$ on a set of N objects I, a consensus partition is conventionally defined over what is referred to as consensus matrix, that is an $N \times N$ matrix $A = (a_{ij})$ whose (i, j)th entry a_{ij} is defined as the number of those partitions R_m $(m = 1, 2, \ldots, M)$ in which both i and j belong in the same part $(i, j \in I)$. Any partition R on I can be one-by-one

Table 1. Data of five partitions over six objects represented by columns of cluster labels

#	R1	R2	R3	R4	R5
1	1	1	2	3	3
2	1	1	1	3	2
3	1	2	1	2	2
4	2	2	2	2	3
5	2	2	2	1	1
6	2	3	2	1	1

Table 2. Consensus matrix for partitions in Table 1

#	1	2	3	4	5	6
1	5	3	1	2	1	1
2	3	5	3	0	0	0
3	1	3	5	2	1	0
4	2	0	2	5	3	2
5	1	0	1	3	5	4
6	1	0	0	2	4	5

Table 3. Consensus matrix (on the left), random interactions matrix, in the middle, and modularity shifted consensus matrix (on the right), for the partitions in Table 1

#	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
1	5	3	1	2	1	1	2.22	1.88	2.05	2.39	2.39	2.05	2.78	1.12	–1.05	–0.39	–1.39	–1.05
2	3	5	3	0	0	0	1.88	1.59	1.74	2.03	2.03	1.74	1.12	3.41	1.26	–2.03	–2.03	–1.74
3	1	3	5	2	1	0	2.05	1.74	1.89	2.21	2.21	1.89	–1.05	1.26	3.11	–0.21	–1.21	–1.89
4	2	0	2	5	3	2	2.39	2.03	2.21	2.58	2.58	2.20	–0.39	–2.03	–0.21	2.42	0.42	–0.21
5	1	0	1	3	5	4	2.39	2.03	2.21	2.58	2.58	2.21	–1.39	–2.03	–1.21	0.42	2.42	1.79
6	1	0	0	2	4	5	2.05	1.74	1.89	2.21	2.21	1.89	–1.05	–1.74	–1.89	–0.21	1.79	3.11

Table 4. Scale shifted consensus matrix for the partitions in Table 1

	1	2	3	4	5	6
1	1.96	–0.04	–2.04	–1.04	–2.04	–2.04
2	–0.04	1.96	–0.04	–3.04	–3.04	–3.04
3	–2.04	–0.04	1.96	–1.04	–2.04	–3.04
4	–1.04	–3.04	–1.04	1.96	–0.04	–1.04
5	–2.04	–3.04	–2.04	–0.04	1.96	0.96
6	–2.04	–3.04	–3.04	–1.04	0.96	1.96

represented by its binary $N \times N$ adjacency matrix $r = (r_{ij})$ such that $r_{ij} = 1$ if both i and j belong to the same part in R, and $r_{ij} = 0$, otherwise. It is well-known that $A = \Sigma_m r_m$ where r_m is the adjacency matrix for R_m $(m = 1, 2, \ldots, M)$.

The consensus values express the extent of similarity between objects in I according to the partitions R_m $(m = 1, 2, \ldots, M)$. The most similar are those objects i, j, that belong to the same part in all the partitions. Consensus value $a_{ij} = M$ for them. In contrast, most dissimilar are objects which are never in the same part; $a_{ij} = 0$ for them. All elements of matrix A are not negative. For the goals of partitioning, it should be beneficial to shift them so that the average value of A becomes zero. There are two shifting transformations of the similarities described in the literature, modularity shift [ ] and scale shift [ ], defined as follows:

• **Modularity shift.** This is based on the concept of random interactions between objects. Let the summary similarities $a_{i.} = \Sigma_j a_{ij}$ and $a_{.j} = \Sigma_i a_{ij}$ express the respective "charge" of row i and column j. Then the random interaction between items i and j is proportional to the product of the charges, $a_{i.}a_{.j}$, and can be expressed as the ratio $a_{i.}a_{.j}/a_{..}$ where is the total "charge" $a_{..} = \Sigma_j a_{.j} = \Sigma_i a_{i.}$, to be rated in the same units as the charges. Modularity shift cleans the similarities from the random interactions:

$$a_{ij} \Leftarrow a_{ij} - a_{i.}a_{.j}/a_{..}. \tag{1}$$

It is not difficult to prove that the total sum of the modularity shifted similarities a_{ij} is equal to zero, so that the average similarity after the modularity shifting is 0 as well.

• **Scale shift.** This transformation is a simple shift of the origin of the scale of measurement of the similarities a_{ij} into the point of the average similarity value, which is $\overline{a} = \dfrac{\sum_{i,j} a_{ij}}{N^2} = \dfrac{\sum_i a_{i.}}{N^2} = \dfrac{\sum_j a_{.j}}{N^2}$

$$a_{ij} \Leftarrow a_{ij} - \overline{a}. \tag{2}$$

Of course, the total sum of the scale shifted similarities a_{ij} is equal to zero too, and the average similarity after the scale shifting is 0 as well.

Example. Consider a set of 6 objects $I = \{1, 2, 3, 4, 5, 6\}$ and five partitions of this set presented in Table 1.

Table 2 presents consensus matrix over the data in Table 1.

This matrix together with the matrix of random interactions and that of the modular transformation is presented in Table 3.

To obtain scale shifted matrix, one needs to subtract the average similarity, 3.04, at this dataset, from all the consensus matrix entries (see in Table 4).

One can notice that the modularity shifting leaves much more entries positive than the scale shifting. In the right subtable in Table 3, one can see two rows with three positive entries each, the 2d and the 5th rows. In contrast, the number of positive entries in the scale shifted matrix of Table 4 is extremely limited: only one outside the main diagonal! This is important because only positive entries can force entities to merge in a cluster, as we will see further on.

3. PARTITIONING CRITERIA

Out of several clustering criteria to occur in the literature, we consider two based on within-cluster similarities. That is, the consensus matrix entries, a_{ij}, are considered as similarities between i and j. One criterion is just the sum of within cluster similarities. That is, for any partition $R = \{R_1, R_2, \dots, R_K\}$ of I in K parts (clusters) the within-cluster R_k $(k = 1, 2, \dots, K)$ similarity is scored with the sum of similarities a_{ij} over all pairs (i, j) of objects from R_k, so that the total within-cluster similarity is expressed as

$$f(R) = \sum_{k=1}^{K} \sum_{i,j \in R_k} a_{ij}, \tag{3}$$

which should be made as high as possible. It is obvious that at non-negative a_{ij}, this criterion leads to a trivial solution: the maximum of within-cluster sum is reached at the universal cluster embracing all the N objects. This is why preliminarily transforming the similarities with either modularity shifting or scale shifting are beneficial for the goal of partitioning. In fact, the popular modularity clustering criterion [14, 3] is but the summary within-cluster similarity criterion (3) after the modularity shift [11].

Another criterion under consideration is the semi-average within-cluster similarity [11]:

$$g(R) = \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i,j \in R_k} a_{ij}, \tag{4}$$

which is very similar to criterion (3), except that the within-cluster sums here are regularized by dividing them over the cluster size. This criterion emerges in the context of approximate clustering including the additive clustering [11]. This criterion does not necessarily lead to the universal cluster at non-negative similarities; however, it is beneficial to apply it after preliminarily shifting the similarities, as we will see later in the text. We employ this criterion as the most appropriate choice for consensus clustering, as we will see later.

4. AGGLOMERATE ALGORITHMS

There are several types of algorithms customarily applied for obtaining suboptimal solutions: agglomeration of smaller clusters into larger ones, division of larger clusters into smaller ones, obtaining clusters one by one, exchange between clusters by objects. Of them, we concentrate here only on agglomeration, because we are going to demonstrate that a version of agglomerate algorithm for the criterion (4) is effective indeed for the goal of consensus clustering.

The classical agglomerate algorithm with respect the summary criterion in (3) can be formulated as follows:

Algorithm AgSu:

1. Initialization. Take singleton clusters as the initial partition to consist of N singletons $\{1\}, \{2\}, \ldots, \{N\}$. Define the similarity matrix between them, $B = (b_{ij})$, to be equal to $A = (a_{ij})$, meaning that similarity b_{ij} between singletons $\{i\}$ and $\{j\}$ is equal to a_{ij}.

2. General step. Given a partition $S = \{S_1, S_2, \ldots, S_m\}$ of I and $m \times m$ matrix $B = (b_{st})$ of summary similarity values between clusters $(s, t = 1, 2, \ldots, m)$, find the maximum $b_{s^*t^*} = \max_{s \neq t} b_{st}$. If $b_{s^*t^*} > 0$, merge clusters S_{s^*} and S_{t^*} into a union, $S_{s^*} \Leftarrow S_{s^*} \cup S_{t^*}$ with a follow-up recomputation of the summary similarity by adding row t^* to row s^*: $b_{s^*t} \Leftarrow b_{s^*t} + b_{t^*t}$ for all $t = 1, 2, \ldots, m$, after which the same applies to columns: $b_{ss^*} \Leftarrow b_{ss^*} + b_{st^*}$ for all $s = 1, 2, \ldots, m$. Then row and column t^* is removed from matrix B, and m is decreased by 1. If $b_{s^*t^*} < 0$, stop. If $m < 3$, stop.

This algorithm is locally optimal: at each merger, taking the maximum of b_{st} maximizes the value of criterion (3): the difference between its values after and before the merger is equal to $2b_{st}$.

A similar algorithm can be formulated for the semi-average criterion (4). The only thing that needs to be formulated before is the level of criterion change at merging two clusters. Let us consider a current partition $S = \{S_1, S_2, \ldots, S_m\}$ and the result of merging clusters S_s and S_t in it, $S(s,t) = \{S_1, \ldots, S_{s-1}, S_s \cup S_t, \ldots, S_m\}$. The difference between $g(S(s,t))$ and $g(S)$ can be expressed as

$$\Delta g(s,t) = g(S(s,t)) - g(S) = \frac{2b_{st} - N_t b_{ss}/N_s - N_s b_{tt}/N_t}{N_s + N_t}, \tag{5}$$

where b_{st}, b_{ss}, b_{tt} are elements of the matrix B of summary similarities between/within clusters. This formula can be proved with elementary transformations of elements of the difference $g(S(s,t)) - g(S)$.

- **Algorithm AgSa:**

1. Initialization. Take singleton clusters as the initial partition to consist of N singletons $\{1\}, \{2\}, \ldots, \{N\}$. Define the similarity matrix between them, $B = (b_{ij})$, to be equal to $A = (a_{ij})$, meaning that similarity b_{ij} between singletons $\{i\}$ and $\{j\}$ is equal to a_{ij}.

2. General step. Given a partition $S = \{S_1, S_2, \ldots, S_m\}$ of I and $m \times m$ matrix $B = (b_{st})$ of summary similarity values between clusters $(s, t = 1, 2, \ldots, m)$, find the maximum $\Delta g(s^*, t^*) = \max_{s \neq t} \Delta g(s,t)$, see (5). If $\Delta g(s^*, t^*) > 0$, merge clusters S_{s^*} and S_{t^*} into a union, $S_{s^*} \Leftarrow S_{s^*} \cup S_{t^*}$ with a follow-up recomputation of the summary similarity by adding row t^* to row s^*: $b_{s^*t} \Leftarrow b_{s^*t} + b_{t^*t}$ for all $t = 1, 2, \ldots, m$, after which the same applies to columns: $b_{ss^*} \Leftarrow b_{ss^*} + b_{st^*}$ for all $s = 1, 2, \ldots, m$. Then row and column t^* are removed from matrix B, and m is decreased by 1. If $\Delta g(s^*, t^*) < 0$, stop. If $m < 3$, stop as well.

The formulations of AgSu and AgSa algorithms have that advantage that at similarities after the modular shifting or scale shifting, the number of clusters is determined in both AgSu and AgSa automatically: the computation stops at that m at which all the non-diagonal elements of B are negative in AgSu, or $\Delta g(s,t) < 0$ for $s \neq t$. At this, no cluster merger can further increase the value of criterion (3) or (4), respectively.

As is well known, the agglomeration algorithms are time-consuming because at each agglomeration step they look for a minimum among all the elements in B, which is of the order of N^2, especially at the beginning. Efforts at decreasing the computation went in the direction of finding and using properties of clustering criteria to allow using results of the previous iterations [13]. Then paper [2] came up with a revolutionary idea of dropping off the need in comprehensively searching through all the pairs of similarities. The authors used the summary modularity criterion (3) to formulate and substantiate experimentally a general idea they dubbed as Louvain algorithm: let us take one element, s, in pair (s, t) to be random, so that the maximum of b_{st} is determined only by enumeration of t, which is of the order of N, not N^2. Of course, this idea can be used with any criterion, not just the modularity based. Thus, we utilize two versions of the Louvain algorithm. We formulate here a general version which can be used with any criterion $c(R)$ to be maximized over all partitions $R = \{R_1, R_2, \ldots, R_K\}$. Assume that one can easily compute the difference $\Delta c(s, t) = c(S(s, t)) - c(S)$.

General algorithm Louvain GAL.

1. Specify the N-part singleton partition $R = \{\{1\}, \{2\}, \ldots, \{N\}\}$ as the starting partition.

2. At any given partition $R = \{R_1, R_2, \ldots, R_K\}$, run a loop over arbitrary ordering of clusters $1, 2, \ldots, K$.

Within this loop,

2.1. At any given s, find t^* maximizing the difference $\Delta c(s, t)$ over all $t = 1, 2, \ldots K$.

2.2. Merge clusters R_s and R_{t^*} together and change R for partition $R(s, t^*)$; decrease K by 1.

2.3. Check stop condition: either all $\Delta g(s, t) < 0$ for $s \neq t$ or $K < 3$. If true, stop. If not, return to the loop.

2.4. If the loop is finished, go to 2 with the current R.

There is one additional operation, which is to be performed before running any of the agglomerate algorithms:

ZD: Zeroing diagonal elements of the similarity matrix.

This is done by making every diagonal element of the current cluster-to-cluster similarity matrix B equal to zero.

We tested whether it is beneficial to execute ZD before every agglomerate step, not only in the beginning. It appears, that is beneficial only at AgSa agglomeration algorithm, out of all the techniques considered above. Thus, further on, we consider AgSa involving ZD at every agglomeration, whereas all the other options involve ZD only in the beginning.

5. EXPERIMENTAL COMPUTATIONS

Surprisingly, the business of experiments with consensus clustering algorithms is based on rather shaky foundations. Typically, a dataset is taken, and a clustering algorithm applies many times at various parameter settings to generate a clustering ensemble, that is a set of partitions to be used for finding a consensus clustering. Therefore, the issue of consensus clustering here is combined with specifics of the clustering algorithms and datasets. This generates issues related to the "quality" of clustering ensembles, their diversity, their representativeness, etc. [4, 15, 1]. In our view, the issue of consensus clustering should be relieved from issues related to producing clustering ensembles and, more so, issues related to quality of clustering algorithms producing those ensembles. In other words, the experimental setting for a consensus clustering algorithm should straightforwardly generate clustering ensembles so that their diversity and representativeness are easily controllable.

We propose here such a setting. First, our synthetic data generator produces a "ground truth" partition of an entity set. To do so, we specify three parameters: the size of the entity set N, the

Table 5. Parameters of the experiments: N number of objects, K number of clusters, M ensemble size, m minimum cluster size, p mutation probability

N	K	M	m	p	Data shift	Algorithm
1000 3000	4 9 15	10 40	2	0.8	Modularity Scale	Agglomeration Louvain

number of parts in the partition, K, and their minimum size, m. This latter parameter is useful when applying clustering algorithms whose parameters are based on probabilities. To decently estimate these parameters, one may need at least m elements. Then we put mK elements in different parts (adding to them m-element portions); the other $N - mK$ elements are assigned with either of K cluster labels randomly. This can be done with a $randi(K, T)$ procedure, which assigns each of $T = N - Km$ elements with either of K different integer labels.

After the ground truth partition S is generated, we generate an ensemble of partitions $R_1, R_2, \ldots, R_M$ to represent it. To this end, we specify a probability, p, $0 < p < 1$, of "mutation". To generate R_1, we reassign $100p\%$ of the entities to any randomly chosen cluster. Other partitions are generated similarly. By increasing p, one increases the diversity of the ensemble. The fact that such an ensemble is representative for the ground truth partition follows from the construction.

Some may find our mutation mechanism overly simplistic. For example, all the partitions generated with this have the same number of clusters as the ground truth. Indeed, one can propose more complex mutation schemas to involve, for example, random mergers and splits of the ground truth clusters. Let us, however, point to good properties of our data generator. First, we can build rather diverse partitions indeed by increasing the mutation probability p. Second, by decreasing the number of partitions in our ensembles, we can create really difficult situations for consensus clustering algorithms, say, by making the number of them less than the number of parts in the ground truth partition, $M < K$.

In the follow-up computations we used two values for the number of objects N, $N = 1000$ and $N = 3000$, three values for the number of clusters K, $K = 4$, $K = 9$, and $K = 15$, and two values for the size of partition ensemble, M, $M = 40$ and $M = 10$. These are summarized in Table 5, in which algorithms under consideration are listed also.

The quality of the results is scored by two characteristics: the number of clusters obtained and ARI (Adjusted Rand index), an index of similarity between the ground truth partition and that obtained by an algorithm [6].

ARI is based on the number of pairs of objects that are consistent in partitions under comparison, that is, either belong to a same cluster, or to different clusters, in both partitions:

$$ARI\,(A, B) = \frac{\binom{N}{2} * \sum_{s=1}^{K_A}\sum_{t=1}^{K_B}\binom{n_{st}}{2} - \sum_{s=1}^{K_A}\binom{a_s}{2}\sum_{t=1}^{K_B}\binom{b_t}{2}}{\frac{1}{2}\binom{N}{2}\left[\sum_{s=1}^{K_A}\binom{a_s}{2} + \sum_{t=1}^{K_B}\binom{b_t}{2}\right] - \sum_{s=1}^{K_A}\binom{a_s}{2}\sum_{t=1}^{K_B}\binom{b_t}{2}}. \tag{6}$$

In (6), A and B are two partitions of the entity set with K_A and K_B parts, respectively; a_s and b_t are cardinalities of parts in A and B, respectively; n_{st} – frequencies in the joint AB distribution; $\binom{n}{2}$ is a binomial term equal to $n(n-1)/2$.

The closer the value of ARI to unity, the more similar are the partitions; $ARI = 1.0$ shows that $A = B$. If one of the partitions consists of just one part, the set I itself, then $ARI = 0$. ARI can be negative as well, which may happen rather rarely, as, say, at specially defined "dual" pairs of partitions [7].

Table 6. Consensus clustering results at $N = 1000$, $M = 40$

		Summary criterion				Semi-average criterion			
		Louvain		Agglomeration		Louvain		Agglomeration	
Gen		Mod	SShift	Mod	SShift	Mod	SShift	Mod	SShift
4	ARI	0.84/0.09	0.88/0.03	0.89/0.03	0.90/0.01	0.98/0.01	0.98/0.01	0.99/0.0	0.99/0.0
	#	3.8/0.44	4/0	4/0	4/0	8.6/1.1	9/1.4	4/0	4/0
9	ARI	0.44/0.03	0.43/0.04	0.45/0.01	0.50/0.03	0.99/0.0	0.99/0.00	1.0/0.0	1.0/0.0
	#	4.2/0.45	4.2/0.45	4/0	4.8/0.45	11.6/1.5	11.8/1.3	9/0	9/0
15	ARI	0.29/0.01	0.28/0.01	0.33/0.02	0.34/0.01	0.99/0.0	0.99/0.0	1/0	1/0
	#	4/0	4/0	4.8/0.45	5/0	17.4/0.5	17.6/0.89	15/0	15/0

Table 7. Consensus clustering results at $N = 3000$, $M = 40$

		Summary criterion				Semi-average criterion			
		Louvain		Agglomeration		Louvain		Agglomeration	
Gen		Mod	SShift	Mod	SShift	Mod	SShift	Mod	SShift
4	ARI	0.88/0.01	0.87/0.01	0.88/0.01	0.88/0.01	0.98/0.00	0.98/0.00	0.99/0.0	0.99/0.0
	#	4/0	4/0	4/0	4/0	12.2/1.1	13/1	4/0	4/0
9	ARI	0.40/0.02	0.40/0.03	0.43/0.05	0.41/0.02	0.99/0.0	0.99/0.00	1.0/0.0	1.0/0.0
	#	4.2/0.45	3.8/0.45	4.2/0.45	4/0	14.6/0.55	13.8/0.45	9/0	9/0
15	ARI	0.26/0.01	0.27/0.01	0.29/0.02	0.28/0.02	1.0/0.0	0.99/0.0	1/0	1/0
	#	4/0	4/0	4.4/0.55	4.4/0.55	19.4/1.1	19.8/1.3	15/0	15/0

Table 8. Consensus clustering results at $N = 1000$, $M = 10$

		Summary criterion				Semi-average criterion			
		Louvain		Agglomeration		Louvain		Agglomeration	
Gen		Mod	SShift	Mod	SShift	Mod	SShift	Mod	SShift
4	ARI	0.44/0.02	0.43/0.02	0.41/0.03	0.40/0.02	0.70/0.02	0.70/0.02	0.67/0.03	0.66/0.03
	#	3/0	4/0	4/0	3.2/0.45	13.8/0.84	14/1	4/0	4/0
9	ARI	0.18/0.02	0.16/0.01	0.21/0.05	0.21/0.02	0.73/0.01	0.79/0.01	0.73/0.01	0.74/0.01
	#	3/0	4/1	3/0	3.8/0.84	20.2/1.1	20.0/1.2	9/0	9/0
15	ARI	0.12/0.01	0.11/0.01	0.14/0.01	0.13/0.01	0.81/0.01	0.81/0.01	0.76/0.03	0.76/0.03
	#	3/0	4/0.71	3.4/0.55	4.2/0.45	26.2/1.9	26.2/1.3	14.2/0.84	14.4/0.55

Table 9. Consensus clustering results at $N = 3000$, $M = 10$

		Summary criterion				Semi-average criterion			
		Louvain		Agglomeration		Louvain		Agglomeration	
Gen		Mod	SShift	Mod	SShift	Mod	SShift	Mod	SShift
4	ARI	0.39/0.03	0.40/0.01	0.41/0.01	0.40/0.02	0.71/0.01	0.71/0.01	0.69/0.01	0.68/0.01
	#	3/0	3/0	3/0	3.2/0.45	12.2/3.63	13.4/2.88	4/0	4/0
9	ARI	0.16/0.01	0.16/0.02	0.17/0.01	0.17/0.02	0.79/0.01	0.79/0.01	0.72/0.01	0.72/0.01
	#	3/0	3.6/0.55	3/0	3.6/0.55	23.4/0.89	23.6/0.89	9/0	9/0
15	ARI	0.10/0.01	0.10/0.01	0.10/0.01	0.11/0.01	0.83/0.01	0.83/0.01	0.77/0.01	0.77/0.01
	#	3/0	4.4/0.89	3.4/0.55	4.6/0.55	30.0/1.9	31.6/2.7	15/0	15/0

Upon generation of a partition ensemble and computing the corresponding consensus matrix, it was processes with either of 8 processing options depending on the matrix transformation option (modularity shifting or scale shift), the criterion utilized (summary or semi-average), and the algorithm used (agglomeration or Louvain). The results are presented in Tables 6, 7, 8, and 9, depending on the sizes of data, N, and partition ensembles, M.

These tables clearly show the following:

1. The results at 1000 and 3000 objects almost coincide, which means that the number of objects under consideration has little effect at the consensus solutions.

2. In contrast to expectations, the results at two different normalizations, the modular one and the scale shift, are much similar, so that the question of which one to use gets irrelevant in consensus clustering.

3. The cluster recovery is always better with the semi-average criterion rather than with the summary one. The larger the number of clusters, the greater the difference.

4. The greater the size of the ensemble, the better: the data recovery results at $M = 10$ are considerably worse that at $M = 40$. Especially devastating is the effect at the summary criterion at which the cluster recovery level is kept, on average, at ARI equal to 0.4 at $K = 4, 0.2$ at $K = 9$, and 0.1 at $K = 15$.

5. At $M = 40$, Agglomeration at the semi-average criterion leads to perfect results at $K = 9, 15$, and almost perfect results at $K = 4$. The Louvain algorithm at the semi-average criterion reaches almost as good results in cluster recovery. However, it fails with respect to the number of clusters. In contrast, at $M = 10$, the Louvain algorithm always wins at cluster recovery, although still overestimating the number of clusters.

6. There is a method among those under consideration, that always recovers the number of clusters correctly: the Agglomeration for the Semi-average criterion. This works even when the ARI decreases to the order of 0.7. The only case at which this might fail, however slightly, is the case of $K = 15$, $M = 10$, that is, $M < K$, at a smaller number of objects ($N = 1000$, but not at $N = 3000$).

6. DISCUSSION

Some of the empirical results above can be explained by theoretical considerations involving the general concept of consensus clustering based on indexes of distance between partitions. Given an index $d(R, S)$ scoring dissimilarity between any partitions R and S of I, one can define the concept of consensus partition as follows. Given an ensemble of partitions $R_1, R_2, \ldots, R_M$ of I, a consensus partition is any partition R of I which minimizes the summary distance $D(R) = \sum_{m=1}^{M} d(R_m, R)$. Usually, distance $d(R_m, R)$ is defined as the mismatch, or Mirkin's, distance between corresponding $N \times N$ binary matrices r_m and r whose elements $r_m(i, j)$ or $r(i, j)$ are 1 if i and j are in the same part of R_m or R, respectively; otherwise, they are 0. Mirkin's distance is the number of inconsistent pairs (i, j) such that i and j are in a same part of one of the partitions while i and j belong to different parts in the other partition [10]. Obviously, that is half of the $L1$ distance between partitions' binary matrices. It is not difficult to prove that the consensus partition with Mirkin's distance is that one maximizing the summary criterion

$$F(R) = \sum_{k=1}^{K} \sum_{i,j \in R_k} \left(a_{ij} - \frac{M}{2} \right), \tag{7}$$

where a_{ij} are elements of the consensus matrix for ensemble $R_1, R_2, \ldots, R_M$.

Indeed,

$$D(R) = \sum_{m=1}^{M} d(R_m, R) = \sum_{m=1}^{M} \sum_{i,j=1}^{N} |r_m(i, j) - r(i, j)|/2 = \sum_{i,j=1}^{N} \sum_{m=1}^{M} |r_m(i, j) - r(i, j)|/2.$$

It is not difficult to see that the internal sum is

$$\sum_{m=1}^{M} |r_m(i,j) - r(i,j)| = \sum_{m=1}^{M} |r_m(i,j) - r(i,j)|^2$$

$$= \sum_{m=1}^{M} (r_m(i,j) + r(i,j) - 2r(i,j)\, r_m(i,j)) = a_{ij} + Mr(i,j) - 2a_{ij} r(i,j),$$

since $\sum_{m=1}^{M} r_m(i,j) = a_{ij}$. These manipulations are correct because values of items $r(i,j)$ and $r_m(i,j)$ here are zero or one, so that the quadratic operation leaves them invariant.

By leaving aside the first item, a_{ij}, which is constant here, and multiplying the remainder by $-1/2$, one can see that the problem of minimization of $D(R)$ is equivalent to the problem of maximization of $F(R)$ in (7) indeed.

One can see that criterion (7) indeed is the within-cluster summary criterion (3) with the preliminarily shifting similarities by $M/2$. Unfortunately, there is a shortcoming of thus defined consensus partition: it fails the so-called Muchnik's test [11]. The test requires to check, for any partition $T = \{T_1, T_2, \dots, T_K\}$ on I, whether the T is a consensus partition for the ensemble of its K dichotomous representations $T^k = \{T_k, I - T_k\}$ $(k = 1, 2, \dots, K)$. If yes, the distance passes the test; if not, the distance fails the test.

Let us see whether Mirkin's distance consensus satisfies the test. Take a partition $T = \{T_1, T_2, \dots, T_7\}$ with $K = 7$ parts, so that $K/2 = 3.5$. Consider consensus matrix a_{ij} entries in this case. Assume, first, that i and j belong to the same part of T. Then they must belong to the same part in every T^m because each part of T is contained in either part of T^m. That means $a_{ij} = 7$ for such i and j. Consider now that i belongs, say, to T_1, and j to another part, say T_2. Then i and j belong to different parts in both T^1 and T^2. However, they belong to the same part $I - T_3$ in T^3 because $I - T_3$ contains both T_1 and T_2. Similarly, these i and j belong to the same part in T^m for every other $m = 4, 5, 6, 7$. Thus, $a_{ij} = 5$ for these i and j. Therefore, all entries in the consensus matrix here are either 5 or 7, both being greater than $K/2 = 3.5$. Thus, it is beneficial to maximize the criterion (7) by collecting all the entities in the universal partition {I} consisting of the only part, I itself, but not the partition T. Therefore, Muchnik's test is failed indeed.

There exists another distance measure, which we refer to as a projective distance [12, 11]. Consider a nominal feature over an entity set I represented by partition $S = \{S_l\}$, and another nominal feature represented by partition $R = \{R_k\}$. Let us define an $N \times L$ dummy matrix Y corresponding to partition S, the partition incidence matrix, by assigning each category S_l in S with a binary variable y_l, a dummy, which is just a $1/0$ N-dimensional vector whose elements $y_{il} = 1$ if $i \in S_l$ and $y_{il} = 0$, otherwise $(l = 1, \dots, L)$. Similarly define an $N \times K$ incidence matrix X whose columns x_k are 0/1-vectors corresponding to categories S_k of S. The projective distance is defined as the summary quadratic difference between Y and its orthogonal projection onto the linear space spanning columns of X [12, 11]. Using symbol $\| \|^2$ for denoting the sum of squares (the squared norm), the projective distance between R and S is defined by the formula $\delta(X,Y) = \|Y - P_X Y\|^2$ where P_X is the orthogonal projector $P_X = X(X^T X)^{-1} X^T$ to the linear space spanning the columns of X. Note that this distance measure is asymmetrical. An exact meaning of distance $\delta(X,Y)$ is analyzed in [11, p. 319]. Here we are going to focus on the summary distance $\triangle(R) = \sum_{m=1}^{M} \delta(X, Y_m)$ to be minimized with respect to unknown partition R represented by matrix X, to obtain a projective distance consensus clustering. Matrices Y_m represent here partitions R_m of the given partition ensemble.

It appears this problem is equivalent to the problem of maximization of the semi-average criterion $g(R)$ in (4). To prove this, consider the incidence matrices X and Y_m of partitions R

and R_m, respectively. These binary matrices mark by 1 the belongingness of objects (rows) to clusters. Let us denote the total number of clusters in all the ensemble partitions $(m = 1, 2, \ldots, M)$ by L and form $N \times L$ matrix $Y = (Y_1 Y_2 \ldots Y_M)$ consisting of all the L columns in these matrices. The columns of Y correspond to all the clusters in partitions $R_1, R_2, \ldots, R_M$. Then the criterion $\triangle(R) = \sum_{m=1}^{M} \delta(X, Y_m)$ can be reformulated as $\Delta(X) = \|Y - P_X Y\|^2$, or equivalently, as $\Delta(X) = Tr((Y - P_X Y)(Y - P_X Y)^T)$ where Tr denotes the trace of a square matrix, that is, the sum of its diagonal elements. By opening the parentheses in the latter expression, we have $\Delta(Y) = Tr(YY^T - P_X YY^T - YY^T P_X + P_X YY^T P_X) = Tr(YY^T - P_X YY^T)$. Indeed, the operation Tr is commutative, so that $Tr(P_X YY^T) = Tr(YY^T P_X)$ and $Tr(P_X YY^T P_X) = Tr(P_X P_X YY^T) = Tr(P_X YY^T)$. The last equation follows from the fact that $P_X P_X = P_X$, which is easy to prove directly. Notice now that matrix YY^T is equal to the consensus matrix A. Obviously, $a_{ii} = L$ for all $i \in I$, so that $Tr(YY^T) = NL$. On the other hand, the (i, i)th diagonal element of matrix $P_X A$ equals to the sum of products $p_{ij} a_{ij}$ where p_{ij} is either 0, if i and j are in different clusters, or $1/N_k$ if both i and j belong to the same cluster S_k. This completes the proof.

Now we can prove that the consensus partition defined using the projective distance does pass the Muchnik's test. Consider again a partition $T = \{T_1, T_2, \ldots, T_K\}$ on I and the ensemble of its K dichotomous representations $T^k = \{T_k, I - T_k\}$ $(k = 1, 2, \ldots, K)$. The consensus matrix A here consists of $a_{ij} = K$, if both i and j belong to some T_k $(k = 1, 2, \ldots, K)$, or $a_{ij} = K - 2$, if i and j belong to different parts of T. Consider the semi-average criterion (4) for a partition $R = \{R_1, R_2, \ldots, R_m\}$. Denote the average similarity within R_k by a_k. Then the value of (4) is obviously equal to the sum of $N_k a_k$ where N_k is the number of objects in R_k. The maximum value of a_k in this case is K, and it is reached when R_k is part of a part of T because in this case all the within-cluster values $a_{ij} = K$. If, in contrast, R_k intersects several parts of T, some within-cluster a_{ij} values will be equal to $K - 2$, so that $a_k < K$. This proves that the maximum value of criterion (4) in the case under consideration is NK (as the sum of all $N_k K$ values), and it is reached at any R either coinciding with T or being T's more granular version obtained by dividing some of its parts.

The proven facts can be considered a substantiation of the reported results: the semi-average criterion (4) embodies a good concept of consensus clustering using the projective distance between partitions, whereas the summary criterion (3) relates to a poor concept of consensus clustering using the Mirkin's distance. This is why the criterion (4) in our experiments is overwhelmingly superior to the criterion (3).

We are yet to explain two other empirically observed facts:

1. Why so much different data transformations, as the modularity shift and scale shift, lead to very similar consensus clustering results?

2. Why the heuristic of constant zeroing of the diagonal entries is so much efficient in determining the right number of clusters with the semi-average criterion?

It should be noted that the visible "contradiction" between the high ARI values and wrong numbers of clusters (see Louvain results for the semi-average criterion in Tables 6 and 7 above) can be easily explained by the insensitivity of ARI index to superfluous small clusters. Take, for example, a partition R of a 1000-strong set in two equal-sized parts. Make 20 singleton clusters out of one of the parts and denote thus obtained partition of I in 22 clusters by S. The ARI index between R and S is 0.96, not that far from unity.

7. CONCLUSION

The main goal of this paper is to bring forth a semi-average consensus clustering criterion (4) modified with the constant zeroing of the main diagonal heuristic as that which should be used for consensus clustering to recover both the hidden partition and the number of clusters in it.

We point out that this criterion emerges as that of consensus clustering with a specially designed dissimilarity between partitions score system, the projective distance. Unlike the conventionally used mismatch or Mirkin's distance between partitions (see, for example, in [5]), the projective distance is shown to pass a natural validity test (the Muchnik's test). In the reported experiments, the agglomerative clustering with criterion (4) shows a very strong tendency to reconstruct both the hidden partition and the number of clusters in it. We compare the performance of this method with that by a most popular clustering method, the modularity clustering. Unfortunately, the modularity clustering appears to perform less than satisfactory and should not be applied as a consensus clustering tool. Another contribution of this paper is a novel design of computational experiments with consensus clustering methods. Instead of conventional approaches developing partition ensembles as those mediated by datasets and clustering methods applied, we propose a simple probabilistic mutation mechanism to generate a representative partition ensemble, diversity of which is controlled by the value of mutation probability. We do not include real-world datasets such as those in the celebrated UC Irvine Machine Learning repository in our experiments, because there is no direct evidence that features in these datasets do relate to the ground truth partitions. Future work should involve explanation of the oddities observed, developing more realistic mutation mechanisms, and adaptation of the approach to big data sets. An interesting direction can be FCA-related approaches [16].

REFERENCES

1. de Amorim, R.C., Shestakov, A., Mirkin, B., et al., The Minkowski central partition as a pointer to a suitable distance exponent and consensus partitioning, *Pattern Recognition*, 2017, pp. 62–72.
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., et al., Fast unfolding of communities in large networks, *J. Statist. Mechan.: Theory Experiment*, 2008, no. 10, pp. 10008–10016.
3. Brandes, U., Delling, D., Gaertler, M., et al. On modularity clustering, *IEEE Transactions on Knowledge and Data Engineering*, 2007, vol. 20, no. 2, pp. 172–188.
4. Fern, X. and Lin, W., Cluster ensemble selection, *Statist. Anal. Data Mining: The ASA Data Sci. J.*, 2008, no. 1, pp. 128–141. https://doi.org/10.1002/sam.10008
5. Guénoche, A., Consensus of partitions: a constructive approach, *Advances in Data Analysis and Classification*, 2011, no. 5(3), pp. 215–229.
6. Hubert, L.J. and Arabie, P., Comparing partitions, *J. Classifikat.*, 1985, no. 2, pp. 193–218.
7. Kovaleva, E.V. and Mirkin, B.G., Bisecting K-means and 1D projection divisive clustering: A unified framework and experimental comparison, *J. Classifikat.*, 2015, vol. 32, no. 2, pp. 414–442.
8. Lancichinetti, A. and Fortunato, S., Consensus clustering in complex networks, *Scientific Reports*, 2012, no. 2(1), pp. 1–7.
9. Liu, P., Zhang, K., Wang, P., et al., A clustering-and maximum consensus-based model for social network large-scale group decision making with linguistic distribution, *Inform. Sci.*, 2022, pp. 269–297.
10. Mirkin, B., An approach to the analysis of non-numerical data, in *Matematicheskir metody modelirovaniya i resheniya ekonomicheskikh zadach* (Mathematical Methods for Modeling and Solving Economic Problems), Bagrinovski, K., Ed., Novosibirsk: Institute of Economics, Siberian Branch of the USSR's Academy of Sciences, 1969, pp. 141–150.
11. Mirkin, B., Clustering: A Data Recovery Approach, 2012, vol. 19, New York: Chapman and Hall/CRC. https://doi.org/10.1201/9781420034912
12. Mirkin, B. and Muchnik, I., Geometric interpretation of clustering criteria, in *Metody analiza mnogomernoi ekonomicheskoi informatsii* (Methods for Analysis of Multidimensional Economics Data), Mirkin, B., Ed., Novosibirsk: Nauka, Sib. otd., 1981, pp. 3–11.

13. Murtagh, F. and Contreras, P., Algorithms for hierarchical clustering: an overview, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2012, no. 32, pp. 86–97.

14. Newman, M.E., Modularity and community structure in networks, *Proc. Nation. Acad. Sci.*, 2006, vol. 103, no. 23, pp. 8577–8582.

15. Pividori, M., Stegmayer, G., and Milone, D.H., Diversity control for improving the analysis of consensus clustering, *Inform. Sci.*, 2016, no. 361, pp. 120–134.

16. Gnatyshak, D., Ignatov, D.I., Mirkin, B.G., et al., A Lattice-based Consensus Clustering Algorithm, *CLA. CEUR Workshop Proceedings*, 2016, vol. 1624, pp. 45–56.

This paper was recommended for publication by A.A. Galyaev, a member of the Editorial Board