Individual approximate clusters: methods, properties, applications

Boris Mirkin

Abstract A least-squares data approximation approach to finding individual clusters is advocated. A simple local optimization algorithm leads to suboptimal clusters satisfying some natural tightness criteria. Three versions of an iterative extraction approach are considered, leading to a portrayal of the cluster structure of the data. Of these, probably most promising is what is referred to as the incjunctive clustering approach. Applications are considered to the analysis of semantics, to integrating different knowledge aspects and consensus clustering.

1 Individual clusters in graph theory and clustering

In spite of the ubiquitous use of partitions and hierarchies as the only two cluster structures of interest (see, for example, [8]), individual clusters are prominent in the analysis of similarity data from the start. Intuitively, cluster is a set of highly similar entities that are dissimilar from entities outside of the cluster.

Currently, the most popular format for similarity data is of square matrix $A = (a_{ij})$ of pair-wise indices a_{ij} expressing similarity between entities $i, j \in I$. The greater the value of a_{ij}, the greater the similarity between i and j. Some examples of similarity data are (i) individual judgements of similarity expressed using a fixed range, (2) correlation coefficients between variables or time series, (3) graphs represented by 1/0-similarity matrices, (4) weighted graphs, or networks, (5) probabilities of common ancestry, especially in proteomics, (6) affinnity data obtained by transformation of distances using a Gaussian or another kernel function. Consider an example of a data set of this type.

Eurovision song contest scoring

Boris Mirkin
Division of Applied Mathematics, Higher School of Economics, Moscow, RF, and Department of Computer Science and Information Systems, Birkbeck, University of London, UK, e-mail: bmirkin@hse.ru

Table 1 presents the average scores given by each country to her 10 top choices at the Eurovision song contests (up to and including year 2011). I compiled this using public data at http://www.escstats.com/ (visited 28/2/2013). Each row of the table corresponds to one out of selected nineteen European countries, and assigns a non-zero score to those of the other eighteen that have been among the 10 best choices. The cluster structure of the table should quantify to what extent the gossip of the effects of cultural and ethnical links on voting is justified, because the quality of songs and performances may be considered random from year to year, so that in the ideal case when no cultural preferences are involved at evaluations, the similarity matrix should be of a random structure too.

Table 1 Eurovision scoring: Each row contains the average score given by the row country to the column country in Eurovision song contests (multiplied by 10).

Country	Az	Be	Bu	Es	Fr	Ge	Gr	Is	It	Ne	Pol	Por	Ro	Ru	Se	Sp	Sw	Ukr	UK
1 Azerbajan	0	0	0	0	0	0	61	48	0	0	0	0	50	65	0	0	0	90	0
2 Belgium	38	0	0	0	0	39	40	0	0	47	0	0	0	0	0	34	0	0	42
3 Bulgaria	67	0	0	0	0	0	93	0	0	0	0	0	0	48	60	0	0	44	0
4 Estonia	41	0	0	0	0	0	0	0	43	0	0	0	0	88	0	0	0	43	0
5 France	0	37	43	0	0	0	0	56	47	0	0	54	0	0	80	0	0	0	41
6 Germany	0	0	0	0	34	0	37	35	0	0	55	0	0	0	70	0	0	0	42
7 Greece	54	0	80	0	41	0	0	0	0	0	0	0	40	0	80	44	0	38	0
8 Israel	50	0	0	0	0	0	0	0	0	43	0	0	66	74	50	0	0	62	43
9 Italy	0	0	100	0	54	0	0	0	0	0	0	0	120	0	0	0	0	65	52
10 Netherlands	39	46	0	0	0	38	0	45	0	0	0	0	0	0	70	0	0	0	0
11 Poland	84	43	0	39	0	0	0	0	90	0	0	0	0	0	0	0	0	82	0
12 Portugal	0	35	0	0	0	45	0	41	81	0	0	0	52	0	57	42	0	74	43
13 Romania	52	0	0	0	0	0	82	0	60	0	0	0	0	49	80	0	0	35	0
14 Russia	99	0	0	0	0	0	37	36	0	0	0	0	0	0	80	0	0	77	0
15 Serbia	0	0	53	0	0	0	73	0	0	0	0	0	0	44	0	0	0	44	0
16 Spain	0	0	78	0	0	51	45	0	74	0	0	43	79	0	47	0	0	46	0
17 Switzerland	0	0	0	0	44	0	0	42	47	0	0	0	0	0	106	41	0	0	41
18 Ukraine	111	0	0	0	0	0	0	0	0	0	60	0	0	98	90	0	0	0	0
19 UK	0	0	0	36	0	39	38	0	0	0	0	0	0	0	37	0	0	0	0

There are several individual cluster related graph-theoretic concepts: (a) *connected component* (a maximal subset of nodes in which there is a path connecting each pair of nodes), (b) *bicomponent* (a maximal subset of nodes in which each pair of nodes belongs to a cycle), and (c) *clique* (a maximal subset of nodes in which each pair of nodes is connected by an edge). Even more relevant is a more recent concept of (d) the *maximum density subgraph* [5]. The density $g(S)$ of a subgraph $S \subset I$ is the ratio of the number of edges in S to the number of elements $|S|$. For an edge weighted graph with weights specified by the matrix $A = (a_{ij})$, the density of a subgraph on $S \subseteq I$ $g(S)$ is defined by the *Rayleigh quotient* $s^T As/s^T s$, where $s = (s_i)$ is the characteristic vector of S, viz. $s_i = 1$ if $i \in S$ and $s_i = 0$ otherwise. The maximum value of the Raleigh quotient of a symmetric matrix over any real vector s is equal to the maximum eigenvalue and is attained at an eigenvector corresponding to this eigenvalue. This gives rise to the so-called (e) *spectral clustering*.

Cluster-specific individual cluster concepts include those of B-cluster [7] and Apresian's cluster [1].

2 Approximation models for summary and semi-average criteria

2.1 Least-squares approximation

The idea is to find such a subset $S \subseteq I$ that its binary matrix $s = (s_{ij})$ approximates a given symmetric similarity matrix A as close as possible. To take into account the difference in the unit of measurement of the similarity as well as for its zero point, matrix s should be also supplied with (adjustable) scale shift and rescaling coefficients, say λ and μ. That would mean that the approaximation is sought in the set of all binary $\lambda + \mu$ / μ matrices $\lambda s + \mu$ with $\lambda > 0$. Unfortunately, such an approximation, at least when follows the least squares approach, would have little value as a tool for producing a cluster, because the optimal values for λ and μ would not separate the optimal S from the rest [10, 11]. This is why this author uses only one parameter λ, change of the unit of measurement, in formulating approximation problems in clustering. The issue of adjustment of similarity zero point, in such a setting, is moved out of the modeling stage to the data pre-processing stage. This amounts to substraction of a similarity shift value from all the similarity values before doing data analysis. Choice of the similarity shift value may affect the clustering results, which the user can take advantage of to differently contrast within- and between- cluster similarities. In the remainder, it is assumed that a similarity shift value has been subtracted from all the similarity entries. Another assumption, for the sake of simplicity, is that the diagonal entries a_{ii} are all zero (after the preprocessing step). From now on, S is represented by a vector $s = (s_i)$ such that $s_i = 1$ if $i \in S$ and $s_i = 0$, otherwise. Our approximation model is

$$a_{ij} = \lambda s_i s_j + e_{ij} \tag{1}$$

where a_{ij} are the preprocessed similarity values, $s = (s_i)$ is the unknown cluster belongingness vector and λ, the rescaling value, also referred to as the cluster intensity value. To fit the model (1), only the least squares criterion $L^2 = \sum_{i,j \in I} e_{ij}^2$ is considered here.

2.1.1 Pre-specified intensity

We first consider the case in which the intensity λ of the cluster to be found is prespecified. Since $s_i^2 = s_i$ for any 0/1 variable s_i, the least squares criterion can be expressed as

$$L^2(S,\lambda) = \sum_{i,j\in I}(a_{ij} - \lambda s_i s_j)^2 = \sum_{i,j\in I} a_{ij}^2 - 2\lambda \sum_{i,j\in I}(a_{ij} - \lambda/2)s_i s_j \quad (2)$$

Since $\sum_{i,j} a_{ij}^2$ is constant, for $\lambda > 0$, minimizing (2) is equivalent to maximizing the summary within-cluster similarity after subtracting the threshold value $\pi = \lambda/2$, i.e.,

$$f(S,\pi) = \sum_{i,j\in I}(a_{ij} - \pi)s_i s_j = \sum_{i,j\in S}(a_{ij} - \pi). \quad (3)$$

This is the so-called summary similarity criterion which satisfies the following properties:

Statement 1 *A cluster S optimizes criterion (3) over similarity matrix A if and only if S optimizes it over symmetric similarity matrix $A + A'$.*

Statement 2 *The optimal cluster size according to criterion (3) can only decrease when π grows.*

One more property of the criterion is that it leads to provably tight clusters. Let us refer to cluster S as suboptimal if, for any entity i, the value of criterion (3) can only decrease if i changes its state in respect to S. Entity i changes its state in respect to S if it is added to S, in the case that $i \notin S$, or removed from S if $i \in S$.

Statement 3 *If S is a suboptimal cluster, then the average similarity $a(i,S)$ of i with other entities in S is greater than π if $i \in S$, or less than π if $i \notin S$.*

An algorithm for producing a suboptimal cluster S starting from any entity i by adding/removing a single entity can be drawn using property:

$$\Delta(S,k) = f(S \pm k) - f(S) = -2z_k \sum_{i\in S} a_{ik}, \quad (4)$$

under the assumption that the diagonal similarities a_{ij} are not considered and z_k in (4) corresponds to S, that is, taken before the change of sign.

2.1.2 Optimal intensity

When λ in (2) is not fixed but can be adjusted to further minimize the criterion, it is easy to prove that the optimal λ is

$$\lambda = a(S) = s^T A s / [s^T s]^2, \quad (5)$$

where $a(S)$ is the average within cluster S similarity.

By putting this equation in the least-squares criterion (2), one can prove:

$$L^2(S) = (A,A) - [s^T A s / s^T s]^2, \quad (6)$$

which implies that the optimal cluster S is a maximizer of

$$g^2(S) = [s^T As/s^T s]^2 = a^2(S)|S|^2 \quad (7)$$

According to (7), the maximum of $g^2(S)$ may correspond to either positive or negative value of $a(S)$. The focus here is on maximizing (7) only for positive $a(S)$. This is equivalent to maximizing its square root, that is the Rayleigh quotient,

$$g(S) = s^T As/s^T s = a(S)|S| \quad (8)$$

This criterion is a form of the so-called semi-average clustering criterion which has a number of properties similar to those of the summary similarity criterion. In particular a cluster tightness property is:

Statement 4 *If S is a suboptimal cluster, then the average similarity $a(i,S)$ of i with other entities in S is greater than $a(S)/2i$ if $i \in S$, or less than $a(S)/2$ if $i \notin S$.*

An algorithm for producing a suboptimal cluster S starting from any $i \in I$ can be drawn by selecting such an entity i whose adding to S if $i \notin S$ or removal from S if $i \in S$ makes the greatest increment of criterion (8).

2.2 Partitional, additive and incjunctive clusters: iterative extraction

The approximation model can be extended to a set of (not necessarily disjoint) similarity clusters $S_1, S_2, ..., S_K$:

$$a_{ij} = \biguplus_{k=1}^{K} \lambda_k s_i^k s_j^k + e_{ij}, \quad \text{for } i, j \in I, \quad (9)$$

where $s^k = (s_i^k)$ and λ_k are k-th cluster belongingness vector and the intensity. The symbol $\biguplus$ denotes an operation of integration of the binary values together with their intensities. We consider three versions of the operation: (a) additive clusters: $\biguplus$ is just summation; (b) partitional clusters: $\biguplus$ denotes the fact that clusters are disjunct, no overlapping; (c) incjunctive clusters: $\biguplus$ is maximum over $k = 1, 2, ..., K$, that is, operation of inclusive disjunction.

The goal is to minimize the residuals e_{ij} with respect to the unknown relations R^k and intensities λ_k.

Additive cluster model was introduced, in the English language literature, by Shepard and Arabie in [18], and independently, and even earlier, in a more general form embracing other cluster structures as well, by the author in mid-seventies in Russian ([10], see references in [11]). Incjunctive clusters have not been considered in the literature, to our knowledge.

We maintain that cluster structures frequently are similar to that of the Solar system so that clusters hidden in data much differ with respect to their "contributions". We proposed an iterative extraction method [10] to find clusters one by one (see also

[11, 13]). Depending on the setting, that is, meaning of $\biguplus$ in (9), one may use the following options:

i **Additive clusters.** The iterative extraction works as this:

   a. Initialization. Given a preprocessed similarity matrix A, compute the data scatter $T = (A,A)$. Put $k = 0$.
   b. General step. Add 1 to k. Find cluster S (locally) maximizing criterion $g(S)$ in (8). Output that as S_k, the intensity of this cluster, the within-cluster average $a(S)$ as λ_k, and its contribution to the data scatter, $w_k = a(S)^2|S|^2$.
   c. Test. Check a stopping condition. If it does hold, assign $K = k$ and halt. Otherwise, compute the residual similarity matrix as $A - \lambda_k s_k s_k^T$ and go back to General step with the residual matrix as A.

   The stopping condition can be either reaching a prespecified number of clusters or contribution of the individual cluster has become too small or the total contribution of the so far found clusters has become too large. The individual cluster contributiona are additive in this process. Moreover, the residual matrix in this process tends to 0 when k increases [10, 11].

ii **Partitional clusters** This method works almost like the iterative extraction at the additive clustering model, except that here no residual matrix is considered, but rather the found clusters are removed from the set of entities.

iii **Incjunctive clusters.** Make a loop over $i \in I$. Run the semi-average criterion suboptimal algorithm at $S = \{i\}$ for each i. Remove those of the found clusters that overlap with others too much. This can be done by applying the same algorithm to the cluster-to-cluster similarity matrix; entries in this matrix are defined as proportional to the overlap values. The individual cluster over this matrix contains those clusters that overlap too much - only one of them shoild be left.

For an example, let us apply each of these three strategies to the Eurovision matrix, preliminarily made symmetric with zeroed diagonal entries.

a Additive clusters one by one: With the condition to stop when the contribution of an individual cluster becomes less than 1.5% of the total data scatter, the algorithm found, in addition to the universal cluster I with the intensity equal to the similarity average, six more clusters (see Table 2). We can see that, say, pair

Table 2 Additive clusters found at the Eurovision song contest dataset.

n.	Cluster	Intensity	Contribution, %
1	Azerbajan, Bulgaria, Greece, Russia, Serbia, Ukraine	70.0	21.43
2	Azerbajan, Israel, Romania, Russia, Ukraine	49.5	7.13
3	Bulgaria, Greece, Italy, Romania, Spain	46.8	6.38
4	Azerbajan, Poland, Ukraine	66.8	3.90
5	Italy, Portugal, Romania	53.0	2.46
6	Greece, Romania, Serbia	43.7	1.67

Azerbajan and Ukraine belong to three of the clusters and contribute, therefore,

the summary intensity value 70.0+49.5+66.8=186.3 as the "model" similarity between them (the summary similarity between them in Table 1 is 201).

b Partitional clusters one by one. Here the algorithm is run on the entities remaining unclustered after the previous step (see Table 3).

Table 3 Partitional clusters found one-by-one at the Eurovision song contest dataset.

n.	Cluster	Intensity	Contribution, %
1	Azerbajan, Bulgaria, Greece, Russia, Serbia, Ukraine	70.0	21.43
2	Italy, Portugal, Romania, Spain	56.1	5.50
3	Belgium, Netherlands	57.3	0.96
4	Germany, UK	45.3	0.60
5	France, Israel, Switzerland	11.6	0.12
6	Estonia, Poland	3.3	0.00

There are only two meaningful clusters, East European and Latin South European, in Table 3; the other four contribute too little. The first of the clusters is just a replica of that in the additive clustering computation. Yet the second cluster combines clusters 3 and 5 cleaned from the Balkans in the additive clusters results Table 2.

c Incjunctive clusters from every entity. The semi-average algorithm has been applied starting from $S = \{i\}$ for every $i \in I$. Most of the final clusters coincide with each other, so that there are very few different clusters (see Table 4).

Table 4 All four different incjunctive clusters found at the Eurovision song contest dataset starting from every entity.

	Cluster	Intensity	Contribution, %
1.	Azerbajan, Bulgaria, Greece, Russia, Serbia, Ukraine	70.0	21.43
2.	Belgium, Netherlands	57.3	0.96
3.	Bulgaria, Greece, Serbia	110.6	10.7
4.	Italy, Portugal, Romania, Spain	56.1	5.50

According to the data recovery model, these clusters lead to a recovered similarity matrix as follows: first of all, the subtracted average value, 35.72, should be put at every entry. Then the two entries of Belgium/Netherlands link are to be increased by the intensity of cluster 2, 57.3. Similarly, the intensities of clusters 1 and 4 are to be added for any pair of entities within each. Then entries for pairs from cluster 3 are to be changed for 35.7+110.6=146.3.

This is an example at which the local nature of the algorithm is of an advantage rather than a drawback. Clusters in Table 4 reflect cultural interrelations rather than anything else.

3 Applications

3.1 Semantics of domain-specific nouns

The idea that semantics of domain-specific nouns lies in their relation to specific situations, functions, etc., a few decades back was not that obvious in cognitive sciences as it is now. In the absence of Internet, the researchers used the so-called sorting experiments to shed light on semantics of domain specific nouns [16, 4]. In a sorting experiment, a set of domain-specific words is specified and written down, each on a small card; a respondent is asked then to partition cards into any number of groups according to their perceived similarity among the nouns. Then. a similarity matrix between the words can be drawn so that the similarity score between two words is defined as the number of respondents who put them together in the same cluster. A cognitive scientist may think that behind the similarity matrix can be some "additive" elementary meanings. In the analysis of similarities between 72 kitchenware terms, the iterative one-by-one extraction with the semi-average similarity suboptimal algorithm found that the clusters related to the usage only: (i) a cooking process, such as frying or boiling; (ii) a common consumption use, such as drinking or eating, and (iii) a common situation such as a banket [4]. In contrast to expectations, none of the clusters reflected logical or structural similarities between the kitchenware items.

3.2 Determining similarity threshold by combining knowledge

In [14] partitional clusters of protein families in herpes viruses are found. The similarity between them is derived from alignments of protein amino acid sequences and similarity neigbourhoods. At different similarity shifts, different numbers of clusters can be obtained, from 99 non-singleton clusters (of 740 entities) at the zero similarity shift to only 29 non-singleton clusters at the shift equal to 0.97 [14]. To choose a proper value of the shift, external information can be used – of functional activities of the proteins under consideration in [14]. Although function of most proteins under consideration was unknown, the set of pairs of functionally annotated proteins can be used to shed light onto potentially admissible values of the similarity shift. In each pair, the proteins can be synonymous (sharing the same function) or not. Because of a high simplicity of virus genomes, the synonymous proteins should belong in the same aggregate protein family, whereas proteins of different functions should belong in different protein families. The similarity shift value should be taken as that between the sets of similarity values for synonymous and nonsynonymous proteins. Then, after subtraction of this value, similarities between not synonymous HPFs get negative while those between synonymous HPFs remain positive. In [14] no nonsynonymous pair has a greater *mbc* similarity than 0.66, which should imply that the shift value 0.67 confers specificity for the production of aggregate protein families.

Unfortunately, the situation is less clear cut for synonymous proteins: although the similarities between them indeed are somewhat higher, 24% pairs is less than 0.67. To choose a similarity shift that minimizes the error in assigning negative and positive similarity values, one needs to compare the distribution of similarity values in the set of synonymous pairs with that in the set of non-synonymous pairs and derive the intersection point similarity value (see details in [14]).

3.3 Consensus clustering

Consensus clustering is an activity of summarizing a set of clusterings into a single clustering. This has become popular recently because after applying different clustering algorithms, or the same algorithm at different parameter settings, on a data set, one gets a number of different solutions. Consensus clustering seeks a unified cluster structure behind the solutions found (see, for example, [21, 13]). Here some results of applying an approach from Mirkin and Muchnik [15] in the current setting will be reported (see also [13]).

Consider a partition $S = \{S_1, ..., S_K\}$ on I and corresponding binary membership $N \times K$ matrix $Z = (z_{ik})$ where $z_{ik} = 1$ if $i \in S_k$ and $z_{ik} = 0$, otherwise ($i = 1, ..., N, k = 1, ..., K$). Obviously, $Z^T Z$ is a diagonal $K \times K$ matrix in which (k,k)-th entry is equal to the cardinality of S_k, $N_k = |S_k|$. On the other hand, $ZZ^T = (s_{ij})$ is a binary $N \times N$ matrix in which $s_{ij} = 1$ if i and j belong to the same class of S, and $s_{ij} = 0$, otherwise. Therefore, $(Z^T Z)^{-1}$ is a diagonal matrix of the reciprocals $1/N_k$ and $P_Z = Z(Z^T Z)^{-1} Z^T = (p_{ij})$ is an $N \times N$ matrix in which $p_{ij} = 1/N_k$ if both i and j belong to the same class S_k, and $p_{ij} = 0$, otherwise. Matrix P_Z represents the operation of orthogonal projection of any N-dimensional vector x onto the linear subspace $L(Z)$ spanning the columns of matrix Z.

A set of partitions R^u, $u = 1, 2, ..., U$, along with the corresponding binary membership $N \times L_t$ matrices X^u, found with various clustering procedures, can be thought of as proxies for a hidden partition S, along with its binary membership matrix Z. Each of the partitions can be considered as related to the hidden partition S by equations

$$x^u_{il} = \sum_{k=1}^{K} c^u_{kl} z_{ik} + e^u_{ik} \tag{10}$$

where coefficients c^u_{kl} and matrix z_{ik} are to be chosen to minimize the residuals e^u_{ik}.

By accepting the sum of squared errors $E^2 = \sum_{i,k,u} (e^u_{ik})^2$ as the criterion to minimize, one immediately arrives at the optimal coefficients being orthogonal projections of the columns of matrices X^u onto the linear subspace spanning the hidden matrix Z. More precisely, at a given Z, the optimal $K \times L_u$ matrices $C^u = (c^u_{kl})$ are determined by equations $C^u = Z(Z^T Z)^{-1} X^u$. By substituting these in equations (10), the square error criterion can be reformulated as:

$$E^2 = \sum_{u=1}^{U} ||X^u - P_Z X^u||^2 \tag{11}$$

where $||.||^2$ denotes the sum of squares of the matrix elements. It is not difficult to show that the criterion can be reformulated in terms of the so-called consensus similarity matrix. To this end, let us form $N \times L$ matrix $X = (X^1 X^2 ... X^U)$ where $L = \sum_{u=1}^{U} L_u$. The columns of this matrix correspond to clusters R_l that are present in partitions $R^1, ..., R^U$. Then the least squares criterion can be expressed as $E^2 = ||X - P_Z X||^2$, or equivalently, as $E^2 = Tr((X - P_Z X)(X - P_Z X)^T)$ where Tr denotes the trace of $N \times N$ matrix, that is, the sum of its diagonal elements, and ^{T}, the transpose. By opening the parentheses in the latter expression, one can derive that $E^2 = Tr(XX^T - P_Z XX^T)$. Let us denote $A = XX^T$ and take a look at (i, j)-th element of this matrix $a_{ij} = \sum_l x_{il} x_{jl}$ where summation goes over all clusters R_l of all partitions $R^1, R^2, ..., R^U$. Obviously, a_{ij} equals the number of those partitions $R^1, R^2, ..., R^U$ at which i and j are in the same class. This matrix is referred to in the literature as the consensus matrix. The latter expression can be reformulated thus as

$$E^2 = NU - \sum_{k=1}^{K} \sum_{i,j \in S_k} a_{ij}/N_k.$$

This leads us to the following statement.

Statement 5 *A partition $S = \{S_1, ..., S_K\}$ is an ensemble consensus clustering if and only if it maximizes criterion*

$$g(S) = \sum_{k=1}^{K} \sum_{i,j \in S_k} a_{ij}/N_k \tag{12}$$

where $A = (a_{ij})$ is the consensus matrix.

Criterion (12) is but the sum of semi-average criteria for clusters S_1, ..., S_K. Therefore, the iterative extraction algorithm in its partitional clusters format is applicable here. We compared the performances of this algorithm and a number of up-to-date algorithms of consensus clustering (see Table 5) [19].

Table 5 Consensus clustering methods involved in the experiments.

n.	Method	Author(s)	Reference
1	Bayes	Wang et al.	[23]
2	Vote	Dimitriadi et al.	[3]
3	CVote	Ayad, Kamel	[2]
4	Borda	Sevillano et al.	[17]
5	Fusion	Guenoche	[6]
6	CSPA	Strehl, Ghosh	[21]
7	MCLA	Strehl, Ghosh	[21]

These algoritms have been compared with two versions of the iterative extraction partitional clusters method above differing by the condition whether the option of zeroing all the diagonal entries of the similarity matrix has been utilized or not (Lsc1 and Lsc2). Three types of datasets have been used: (a) datasets from the Irvine Data Repository, (b) generated synthetic datasets, and (c) specially drawn artificial 2D shapes. Here we present only results of applying the algorithms to the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from UCI Data Repository (569 entities, 30 features, two classes) (see Figure 1). The results are more or less similar to each other, although the superiority of our algorithm is expressed more clearly on the other datasets [19].

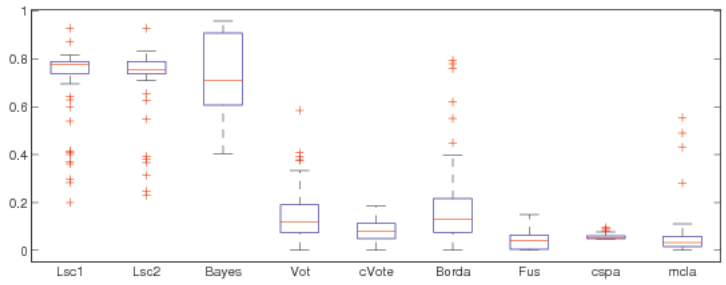

Fig. 1 Comparison of the accuracy of consensus clustering algorithms at WDBC dataset.

Conclusion

The paper describes least squares approximation approaches for finding individual similarity clusters which can be useful in several perspectives - summary similarity criterion, semi-average criterion, spectral clustering criterion and approximation criterion. The clustering criterion involves, in different forms, the concept of similarity threshold, or similarity shift - a value subtracted from all the similarity matrix entries. The threshold can be used for bridging different aspects of the phenomenon under study together. This is demonstrated in section 3.2, in which the final choice of clustering involves the protein function and gene arrangement in the genomic circle, in addition to the original similarity derived from protein sequences.

The criterion leads to nice properties of the clusters: they are quite tight over average similarities of individual entities with them. Also, unlike methods for finding global optima, the one starting from an entity leads to recovery of the local cluster structure of the data, probably a single most important innovation proposed in this paper.

This work was partially supported by the International Laboratory of Decision Choice and Analysis at NRU HSE (headed by F. Aleskerov) and the Laboratory of Algorithms and Technologies for Network Analysis NRU HSE Nizhny Novgorod by means of RF goverment grant ag. 11.G34.31.0057 (headed by V. Kalyagin).

References

1. Y.D. Apresian (1966) An algorithm for finding clusters by a distance matrix, *Computer. Translation and Applied Linguistics*, 9, 7279 (in Russian).
2. H. Ayad, M. Kamel (2010) On voting-based consensus of cluster ensembles, *Pattern Recognition*, 1943-1953.
3. E. Dimitriadou, A. Weingessel and K. Hornik (2002) A Combination Scheme for Fuzzy Clustering, *Journal of Pattern Recognition and Artificial Intelligence*, 332-338.
4. R. Frumkina, B. Mirkin (1986) Semantics of domain-specific nouns: a psycho-linguistic approach, Notices of Russian Academy of Science: Language and Literature, 45(1), 12-22 (in Russian).
5. G. Gallo, M.D. Grigoriadis, and R.E. Tarjan (1989) A fast parametric maximum flow algorithm and applications, *SIAM Journal on Computing, 18*, 30-55.
6. A. Guenoche (2011) Consensus of partitions : a constructive approach, *Adv. Data Analysis and Classification*, 5, pp. 215-229.
7. K.J. Holzinger and H.H. Harman (1941) *Factor Analysis*, University of Chicago Press, Chicago.
8. A.K. Jain and R.C. Dubes (1988) *Algorithms for Clustering Data*, Englewood Cliffs, NJ: Prentice Hall.
9. Hideya Kawaji, Yoichi Takenaka, Hideo Matsuda (2004) Graph-based clustering for finding distant relationships in a large set of protein sequences, *Bioinformatics*, 20(2), 243-252.
10. B. Mirkin (1976) *Analysis of Categorical Features*, Finansy i Statistika Publishers, Moscow, 166 p. (In Russian)
11. B. Mirkin (1987) Additive clustering and qualitative factor analysis methods for similarity matrices, *Journal of Classification, 4*, 7-31; Erratum (1989), *6*, 271-272.
12. B. Mirkin (1990) A sequential fitting procedure for linear data analysis models, *Journal of Classification, 7*, 167-195.
13. B. Mirkin (2012) *Clustering: A Data Recovery Approach*, 2nd Edition, Chapman and Hall, Boca Raton.
14. B.G. Mirkin, R. Camargo, T. Fenner, G. Loizou and P. Kellam (2010) Similarity clustering of proteins using substantive knowledge and reconstruction of evolutionary gene histories in herpesvirus, *Theoretical Chemistry Accounts: Theory, Computation, and Modeling*, 125(3-6), 569-581.
15. B. Mirkin and I. Muchnik (1981) Geometric interpretation of clustering criteria, in B. Mirkin (Ed.) Methods for Analysis of Multidimensional Economics Data, Nauka Publishers (Siberian Branch), Novosibirsk, 3-11 (in Russian).
16. S Rosenberg, MP Kim (1975) The method of sorting as a data-gathering procedure in multivariate research, *Multivariate Behavioral Research*, 10, 489-502.
17. X. Sevillano Dominguez, J. C. Socoro Carrie and F. Alias Pujol (2009) Fuzzy clusterers combination by positional voting for robust document clustering, *Procesamiento del lenguaje natural*, 43, pp. 245-253.
18. R.N. Shepard and P. Arabie (1979) Additive clustering: representation of similarities as combinations of overlapping properties, *Psychological Review, 86*, 87-123.
19. A. Shestakov, B. Mirkin (2013) Least squares consensus clustering applied to k-means results (in progress).
20. M. Smid, L.C.J. Dorssers and G. Jenster (2003) Venn Mapping: clustering of heterologous microarray data based on the number of co-occurring differentially expressed genes, *Bioinformatics*, 19, no. 16, 2065-2071.
21. A. Strehl, J. Ghosh (2002) Cluster ensembles - a knowledge reuse framework for combining multiple partitions, *Journal on Machine Learning Research*, 583-617.
22. S. Swift, A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu and P. Kellam (2004) Consensus clustering and functional interpretation of gene expression data, *Genome Biology*, 5:R94.
23. H. Wang, H. Shan, A. Banerjee (2009) Bayesian cluster ensembles. In: *Proceedings of the Ninth SIAM International Conference on Data Mining*, 211-222.