

# Choosing the number of clusters

Boris Mirkin

Department of Computer Science & Information Systems, Birkbeck University of London, London, UK and  
Department of Data Analysis and Machine Intelligence, Higher School of Economics, Moscow, RF

Email: mirkin@dcs.bbk.ac.uk, bmirkin@hse.ru

## Abstract

The issue of determining “the right number of clusters” is attracting ever growing interest. The paper reviews published work on the issue with respect to mixture of distributions, partition, especially in K-Means clustering, and hierarchical cluster structures. Some perspective directions for further developments are outlined.

## 1. Introduction

It is tempting to consider the problem of determining “the right number of clusters” as akin to the problem of estimation of the complexity of a classifier, which depends on the complexity of the structure of the set of admissible rules, assuming the worst-case scenario for the structure of the dataset (Vapnik 2006). Indeed there are several different clustering formats – such as partitioning, hierarchical clustering, one cluster clustering, fuzzy clustering, Kohonen self-organising maps, and mixtures of distributions – that should involve different criteria. Yet the worst-case scenario paradigm seems rather tangential to clustering, because the worst case here is probably a uniform noise: there is no widely acceptable theoretical framework for systematization of data structures available as yet.

The remainder reviews the bulk of published work on the issue followed by a systematic discussion of directions for further developments.

## 2. Review of published work

Most efforts in determining the number of clusters have been channelled along the following three cluster structures:

- (I) mixture of distributions;
- (II) binary hierarchy; and
- (III) partition.

We present them accordingly.

### 2I. Number of components in a mixture of distributions.

Under the mixture of distributions model, the data is assumed to be a random independent sample from a distribution of density

$$f(x | \theta) = \sum_{k=1}^K p_k f_k(x | \theta_k)$$

where unimodal density  $f_k(x|\theta_k)$  and positive  $p_k$  represent, in respect,  $k$ -th cluster and its probability ( $k=1, \dots, K$ ). Given a sample  $X=\{x_i\}$ , a maximum likelihood estimate of the parameters  $\theta=\{\theta_k\}$  can be derived by maximizing the log-likelihood  $L(g,l)=\sum_i \log f(x_i|\theta) = \sum_{ik} g_{ik}(l_{ik}-\log g_{ik})$  where  $g_{ik} = p_k f(x_i|\theta_k) / \sum_k p_k f(x_i|\theta_k)$  are posterior membership probabilities and  $l_{ik} = \log(p_k f(x_i|\theta_k))$ . A somewhat more complex version of  $L$  involving a random assignment of the observations to the clusters has received attention too (see, for example, Biernacki et al. 2000). Until recently, most research has been concentrated on representing the clusters by Gaussian distributions; currently, the choice is extended to, first of all, Dirichlet distributions that enjoy the advantage of having both a bounded density support and less parameters (for a review, see for example, Bouguila and Ziou 2007).

There have been four approaches to deriving recipes for choosing the right  $K$  under the mixture of distributions model:

(21a) Hypothesis testing for deciding between different hypotheses on the number of clusters (see Bock 1996 for a review), that mainly produced recipes for testing one number against another (Bock 1996, Lo, Mendell and Rubin 2001); this can be used for testing one cluster hypothesis against the two-cluster hypothesis to decide of stopping in hierarchical clustering.

(21b) Additional constraints to the maximization of the log-likelihood, expressed with the help of a parsimony principle such as the principles of the minimum description length or minimum message length; these generated a number of useful criteria to utilize additive terms penalizing  $L$  for the number of parameters to be estimated versus the number of observations, such are Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC) (for reviews, see Bozdogan 1994, McLachlan and Peel 2000 and Figueiredo and Jain 2002); the latter, along with a convenient data normalization, has been advocated by Yeung et al 2001, though Hu and Xu 2004, as well as other authors, present less favourable evidence that BIC tends to overestimate the number of clusters – yet the issue of data transformation should be handled more seriously. The trouble with these criteria is that they tend to be monotone over  $K$  when a cluster structure is not well manifested (Windham and Cutler 1992).

(21c) Collateral statistics that should be minimal at the “right” number  $K$  such as, for example, the information ratio (Windham and Cutler 1992) and the normalized entropy (Celeux and Soromento

1996). Both of these can be evaluated as side products of the ordinary expectation-maximization EM algorithm at each pre-specified  $K$  so that no additional efforts are needed to compute them. Yet, there is not enough evidence for using the latter while there is some disappointing evidence against the former.

(2Id) Adjusting the number of clusters while fitting the model incrementally, by penalizing the rival seeds, which has been shaped recently into a weighted likelihood criterion to maximize (see Cheung 2005 and references therein).

## **2II. Stopping in hierarchic clustering.**

Hierarchic clustering is an activity of building a hierarchy in a divisive or agglomerative way by sequentially splitting a cluster in two parts, in the former, or merging two clusters, in the latter. This is often used for determining a partition with a convenient number of clusters  $K$  in either of two ways:

(2IIa) Stopping the process of splitting or merging according to a criterion. This was the initial way of doing exemplified by the Duda and Hart (1973) ratio of the square error of a two-cluster solution over that for one cluster solution so that the process of division stops if the ratio is not small enough or a merger proceeds if this is the case. Similarly testing of a two-cluster solution versus the one-cluster solution is done by using a more complex statistic such as the log-likelihood or even BIC criterion of a mixture of densities model. The idea of testing individual splits with BIC criterion was picked up by Pelleg and Moore (2000) and extended by Ishioka (2005) and Feng and Hamerly (2006); these employ a divisive approach with splitting clusters by using 2-Means method.

(2IIb) Choosing a cut-off level in a completed hierarchy. In the general case, so far only heuristics have been proposed, starting from Mojena (1977) who proposed calculating values of the optimization criterion  $f(K)$  for all  $K$ -cluster partitions found by cutting the hierarchy and then defining a proper  $K$  such that  $f(K+1)$  exceeds the “three sigma” over the average  $f(K)$ . Obviously every partition-based rule can be applied too. Thirty different criteria were included by Milligan (1981) in his seminal study of rules for cutting cluster hierarchies, though this simulation study included rather small numbers of clusters and data sizes, and the clusters were well separated and equal-sized. Use of BIC criterion is advocated by Fraley and Raftery (1998). For the situations at which clustering is based on similarity data, a good way to go is subtracting some “ground” or “expected” similarity matrix so that the residual similarity leads to negative values in criteria conventionally involving sums of similarity values when the number of clusters is too small (see Mirkin’s (1996, p. 256) uniform threshold criterion or Newman’s modularity function (Newman and Girvan 2004)).

## **2III. Partition structure.**

Partition, especially K-Means clustering, attracts a most considerable interest in the literature, and the issue of properly defining or determining number of clusters  $K$  is attacked by dozens of researchers. Most of them accept the square error criterion that is alternatingly minimized by the so-called Batch version of K-Means algorithm:

$$W(S, \mathbf{C}) = \sum_{k=1}^K \sum_{i \in S_k} \|y_i - c_k\|^2 \quad (1)$$

where  $S$  is a  $K$ -cluster partition of the entity set represented by vectors  $y_i$  ( $i \in I$ ) in the  $M$ -dimensional feature space, consisting of non-empty non-overlapping clusters  $S_k$ , each with a centroid  $c_k$  ( $k=1,2,\dots,K$ ), and  $\| \cdot \|$  the Euclidean norm. A common opinion is that the criterion is an implementation of the maximum likelihood principle under the assumption of a mixture of the “spherical” Gaussian distributions with a constant variance model (Hartigan 1975, Banfield and Raftery 1993, McLachlan and Peel 2000) thus postulating all the clusters  $S_k$  to have the same radius. There exists, though, a somewhat lighter interpretation of (1) in terms of an extension of the SVD decomposition-like model for the Principal Component Analysis in which (1) is but the least-squares criterion for approximation of the data with a data recovery clustering model (Mirkin 1990, 2005, Steinley 2006) justifying the use of K-Means for clusters of different radii.

Two approaches to choosing the number of clusters in a partition should be distinguished: (i) post-processing multiple runs of K-Means at random initializations at different  $K$ , and (ii) a simple pre-analysis of a set of potential centroids in data. We consider them in turn:

### **(2IIIi) Post-processing.**

A number of different proposals in the literature can be categorized in several categories (Chiang and Mirkin 2010):

- (a) Variance based approach: using intuitive or model based functions of criterion (1) which should get extreme or “elbow” values at a correct  $K$ ;
- (b) Structural approach: comparing within-cluster cohesion versus between-cluster separation at different  $K$ ;
- (c) Combining multiple clusterings: choosing  $K$  according to stability of multiple clustering results at different  $K$ ;
- (d) Resampling approach: choosing  $K$  according to the similarity of clustering results on randomly perturbed or sampled data.

Currently, approaches involving simultaneously several such criteria are being developed too (see, for example, Saha and Bandyopadhyaya 2010).

### **(2IIIa) Variance based approach**

Let us denote the minimum of (1) at a specified number of clusters  $K$  by  $W_K$ . Since  $W_K$  cannot be used for the purpose because it monotonically decreases when  $K$  increases, there have been other  $W_K$  based indices proposed to estimate the  $K$ . For example, two heuristic measures have been experimentally approved by Milligan and Cooper (1985):

- A Fisher-wise criterion by Calinski and Harabasz (1974) finds  $K$  maximizing  $CH = ((T - W_K) / (K - 1)) / (W_K / (N - K))$ , where  $T = \sum_{i \in I} \sum_{v \in I'} y_{iv}^2$  is the data scatter. This criterion showed the best performance in the experiments by Milligan and Cooper (1985), and was subsequently utilized by other authors (for example, Casillas et al 2003).

- A “rule of thumb” by Hartigan (Hartigan 1975) utilizes the intuition that when there are  $K^*$  well separated clusters, then, for  $K < K^*$ , an optimal  $(K+1)$ -cluster partition should be a  $K$ -cluster partition with one of its clusters split in two, which would drastically decrease  $W_K$  because the split parts are well separated. On the contrary,  $W_K$  should not much change at  $K \geq K^*$ . Hartigan’s statistic  $H = (W_K / W_{K+1} - 1) / (N - K - 1)$ , where  $N$  is the number of entities, is computed while increasing  $K$ , so that the very first  $K$  at which  $H$  decreases to 10 or less is taken as the estimate of  $K^*$ . The Hartigan’s rule indirectly, via related Duda and Hart (1973) criterion, was supported by Milligan and Cooper (1985) findings. Also, it did surprisingly well in experiments of Chiang and Mirkin (2010), including the cases of overlapping non-spherical clusters generated.

A popular more recent recommendation involves the so-called Gap statistic (Tibshirani, Walther and Hastie 2001). This method compares the value of (1) with its expectation under the uniform distribution and utilizes the logarithm of the average ratio of  $W_K$  values at the observed and multiple-generated random data. The estimate of  $K^*$  is the smallest  $K$  at which the difference between these ratios at  $K$  and  $K+1$  is greater than its standard deviation (see also Yan and Ye 2007). Some other model-based statistics have been proposed by other authors such as Krzhanowski and Lai (1985) and Sugar and James (2003).

### **(2IIIb) Within-cluster cohesion versus between-cluster separation**

A number of approaches utilize indexes comparing within-cluster distances with between cluster distances: the greater the difference the better the fit; many of them are mentioned in Milligan and Cooper (1985). More recent work in using indexes relating within- and between-cluster distances is described in Kim et al. (2001), Shen et al. (2005), Bel Mufti et al. (2006) and Saha and

Bandyopadhyaya (2010); this latter paper utilizes a criterion involving three factors:  $K$ , the maximum distance between centroids and an averaged measure of internal cohesion.

A well-balanced coefficient, the Silhouette width involving the difference between the within-cluster tightness and separation from the rest, promoted by Kaufman and Rousseeuw (1990), has shown good performance in experiments (Pollard and van der Laan 2002). The largest average silhouette width, over different  $K$ , indicates the best number of clusters.

### **(2IIIc) Combining multiple clusterings**

The idea of combining results of multiple clusterings sporadically emerges here or there. The intuition is that results of applying the same algorithm at different settings or some different algorithms should be more similar to each other at the right  $K$  because a “wrong”  $K$  introduces more arbitrariness into the process of partitioning; this can be caught with an average characteristic of distances between partitions (Chae et al. 2006, Kuncheva and Vetrov 2005).

A different characteristic of similarity among multiple clusterings is the consensus entity-to-entity matrix whose  $(i,j)$ -th entry is the proportion of those clustering runs in which the entities  $i, j \in I$  are in the same cluster. This matrix typically is much better structured than the original data. An ideal situation is when the matrix contains 0's and 1's only: this is the case when all the runs lead to the same clustering; this can be caught by using characteristics of the distribution of consensus indices (the area under the cumulative distribution (Monti et al. 2003) or the difference between the expected and observed variances of the distribution, which is equal to the average distance between the multiple partitions (Chiang and Mirkin 2010)).

### **(2IIIid) Resampling methods**

Resampling can be interpreted as using many randomly produced “copies” of the data for assessing statistical properties of a method in question. Among approaches for producing the random copies are:

- ( $\alpha$ ) random sub-sampling in the data set;
- ( $\beta$ ) random splitting the data set into “training” and “testing” subsets,
- ( $\gamma$ ) bootstrapping, that is, randomly sampling entities with replacement, usually to their original numbers, and
- ( $\delta$ ) adding random noise to the data entries.

All four have been tried for finding the number of clusters based on the intuition that different copies should lead to more similar results at the right number of clusters: see, for example, Minaei-Bidgoli,

Topchy and Punch (2004) for ( $\alpha$ ), Dudoit, Fridlyand (2002) for ( $\beta$ ), McLachlan, Khan (2004) for ( $\gamma$ ), and Bel Mufti, Bertrand, Moubarki (2005) for ( $\delta$ ).

Of these perhaps most popular is approach by Dudoit and Fridlyand (2002) (see also Roth, Lange, Braun, Buhmann 2004 and Tibshirani, Walther 2005). For each  $K$ , a number of the following operations is performed: the set is split into non-overlapping training and testing sets, after which both the training and testing part are partitioned in  $K$  parts; then a classifier is trained on the training set clusters and applied for predicting clusters on the testing set entities. The predicted partition of the testing set is compared with that found on the testing set by using an index of (dis)similarity between partitions. Then the average or median value of the index is compared with that found at randomly generated datasets: the greater the difference, the better the  $K$ . Yet on the same data generator, the Dudoit and Fridlyand (2002) setting was outperformed by a model-based statistic as reported by MacLachlan and Khan (2004).

### **(2IIIii) Pre-processing.**

An approach going back to sixties is to recursively build a series of the farthest neighbors starting from the two most distant points in the dataset. The intuition is that when the number of neighbors becomes that of the number of clusters, the distance to the next farthest neighbor should drop considerably. This method as is does not work quite well, especially when there is a significant overlap between clusters and/or their radii considerably differ. Another implementation of this idea, involving a lot of averaging which probably smoothes out both causes of the failure, is the so-called  $iK$ -Means method (Mirkin 2005). This method extracts “anomalous patterns” that are averaged versions of the farthest neighbors (from an unvaried “reference point”) also one-by-one. In a set of experiments with generated Gaussian clusters of different between- and within-cluster spreads (Chiang and Mirkin 2010),  $iK$ -Means has appeared rather successful, especially if coupled with the Hartigan’s rule described above, and outperformed many of the post-processing approaches in terms of both cluster recovery and centroid recovery.

### **3. Directions for development**

Let us focus on the conventional wisdom, that the number of clusters is a matter of both data structure and the user’s eye. Currently the data side prevails in clustering research and we discuss it first.

#### **(3I) Data side.**

This basically amounts to two lines of development: (i) benchmark data for experimental verification and (ii) models of clusters.

##### **(3Ii) Benchmark data.**

The data for experimental comparisons can be taken from real-world applications or generated artificially. Either or both ways have been experimented with: real-world data sets only by Casillas et al. (2003), Minael-Bidgoli et al (2005), Shen et al. (2005), generated data only by Chiang and Mirkin (2009), Hand and Krzanowski (2005), Hardy (1996), Ishioka (2005), Milligan and Cooper (1985), Steinley and Brusco (2007), and both by Chae et al. 2006, Dudoit and Fridlyand (2002), Feng and Hamerly (2005), Kuncheva and Vetrov (2005).

There is nothing to report of research in the real data structures yet. As to the case of generated data, this involves choices of:

(A) Data sizes (growth can be observed from a hundred entities two dozen years ago such as Milligan and Cooper (1985) to a few thousands by now such as Feng and Hamerly (2006), Steinley and Brusco (2007), Chiang and Mirkin (2010));

(B) Number of clusters (from 2 to 4 two dozen years ago to a dozen or even couple of dozens currently);

(C) Cluster sizes (here the fashion moves from equal sized clusters to different and even random sizes, see Steinley and Brusco (2007), Chiang and Mirkin (2010), who both claim that the difference in the numbers of points in clusters does not change the results;

(D) Cluster shapes (mostly spherical and elliptical are considered);

(E) Cluster intermix (from well separated clusters to recognition of that as a most affecting factor (Steinley and Brusco 2007));

(F) Presence of irrelevant features (this is considered rather rarely (Dudoit and Fridlyand 2002, Kuncheva and Vetrova 2005, Milligan and Cooper 1985); and

(G) Data standardization (here the fashion moves from the conventional normalization by the standard deviation to normalization by a range related value – such a shift should not frustrate anybody because normalizing by the standard deviation may distort the data structure: it downplays the role of features with multimodal distributions, which is counterintuitive (Chiang and Mirkin 2010).

The data structures involved in the experiments reported in the literature are rather arbitrary and not comprehensive. It could be guessed that to advance this line of research one needs to apply such a type of cluster structure which is both simple and able to reasonably approximate any data – could it be a mix of the spherical distributions of different variance? – the evaluations of data complexity, which should emerge from such an approach, could perhaps be used for setting a standard for experiments.

(3Iii) Cluster modeling.



Cluster modeling in cluster analysis research is in a rather embryonic state too: Gaussian, or even spherical, clusters prevail to express a rather naïve view that clusters are sets around their centroids, and popular algorithms such as K-Means or EM clustering reflect that too. Experimental evidence suggests that the multiple initializations with centroids in randomly chosen entities outperform other initialization methods in such a setting (Pena et al. 1999, Hand and Krzanowski 2005, Steinley and Brusco 2007). With respect to modeling of cluster formation and evolution, and dynamic or evolving clusters in general, including such natural phenomena as star clusters or public opinion patterns, one can refer to a massive research effort going on in physics (for recent reviews, see Castellano, Fortunato, Loreto 2009 and Springel et al. 2005) – yet there is a huge gap between these and data clustering issues. Even a recent rather simple idea of the network organization as what is called “small world” which seems rather fruitful in analyzing network regularities (see, for example, Newman et al. 2006) has not been picked up in data clustering algorithms. Probably evolutionary or on-line computations can be a good medium for such modelling cluster structures (see some attempts in Jonnalagadda and Srinivasan 2009, Shubert and Sedenbladh 2005). Perhaps the paper by Evanno et al. 2005 involving a genetic multi-loci population model, can be counted as an attempt in the right direction.

### **(3II) User side**

Regarding the “user’s eye”, one can distinguish between at least these lines of development:

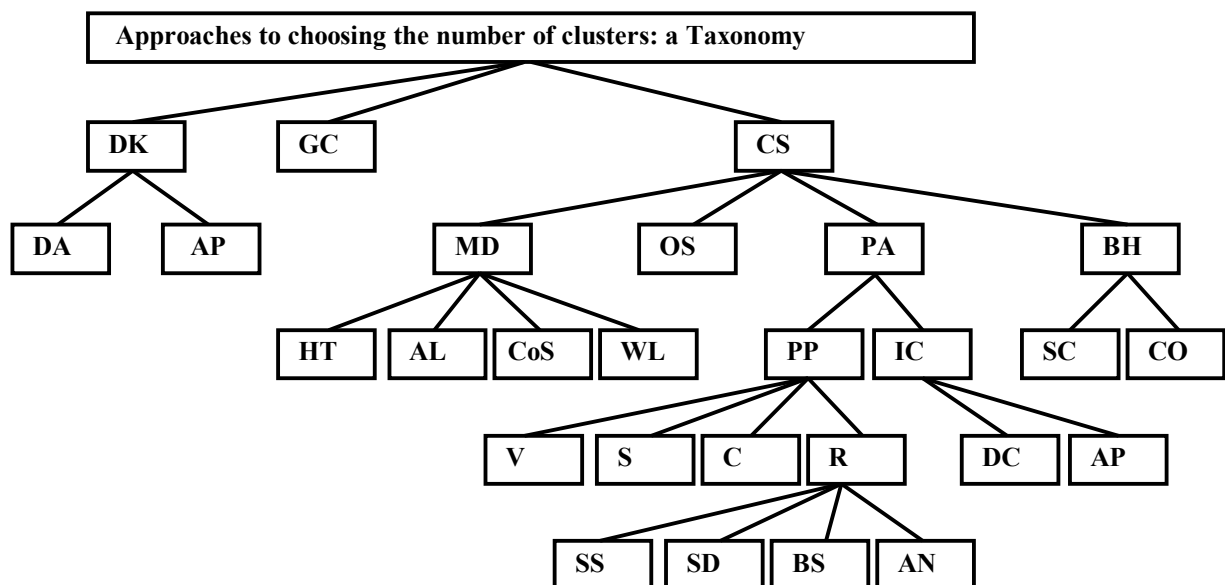
- (i) cognitive classification structures,
- (ii) preferred extent of granularity,
- (iii) additional knowledge of the domain.

Of the item (i), one cannot deny that cluster algorithms do model such structures as typology (K-Means), taxonomy (agglomerative algorithms) or conceptual classification (classification trees). Yet no meaningful formalizations of the classification structures themselves have been developed so far to include, say, “real” types, “ideal” types and social “strata” within the same concept of typology. A deeper line of development should include a dual “archetype” structure potentially leading to formalization of the process of generation of useful features from the raw data such as sequence motifs or attribute subsets – a matter of future developments going far beyond the scope of this paper. Of the item (ii), the number of clusters itself characterizes the extent of granularity; the issue is in simultaneously utilizing other characteristics of granularity, such as the minimal size at which a set can be considered a cluster rather than outlier, and the proportion of the data variance taken into account by a cluster – these can be combined within a data recovery and spectral clustering frameworks (see Mirkin 2005, for the former, and von Luxburg 2007, for the latter). Yet there

should be other more intuitive characteristics of granularity developed and combined with those above. Item (iii) suggests a formalized view of a most important clustering criterion, quite well known to all practitioners in the field – consistency between clusters and other aspects of the phenomenon in question. Long considered as a purely intuitive concept, this emerges as a powerful device, first of all in bioinformatics studies. Dotan-Cohen et al. (2009) and Freudenberg et al. (2009) show how the knowledge of biomolecular functions embedded in the so-called Gene Ontology can be used to cut functionally meaningful clusters out of a hierarchical cluster structure. More conventional approaches along the same line are explored in Lottaz et al. (2007) and Mirkin et al. (2010). The more knowledge of different aspects of real-world phenomena emerges, the greater importance of the consistency criterion in deciding of the right number of clusters.

#### 4. Conclusion

Figure 1 presents most approaches to the issue of choosing the right number of clusters mentioned above as a taxonomy tree.



**Figure 1:** A taxonomy for the approaches to choosing the number of clusters as described in the paper. Here DK is Domain Knowledge in which DA is Direct Adjustment of algorithm and AP is Adjustment by Post-processing of clustering results, GC is modelling of the process of Generation of Clusters, and CS is Cluster Structures. This last item is further divided in MD, that is, Mixture of Distributions, P, Partitions, BH, Binary Hierarchies, and OS, Other cluster Structures. MD involves HT which is Hypothesis Testing, AL, Additional terms in the Likelihood criterion, CoS, Collateral Statistics, and WL, Weighted Likelihood; BH involves SC, Stopping according to a Criterion, and CO, using a cut-off level over a completed tree. Item PA covers PP, Partition Post-processing, and IC, pre-processing by Initialization of Centroids. The latter involves DC, Distant Centroids, and AP, Anomalous Patterns. PP involves V, Variance based approach, S, Structure based approach, C, Combining clusterings approach, and R, Resampling based approach. The latter is composed of SS, Sub-sampling, SD, Splitting the Data, BS, Bootstrapping, and AN, Adding Noise.

Among the current cluster-structure-centered approaches, those involving plurality in either data

resampling or choosing different initializations or using different algorithms or even utilizing various criteria seem to have an edge over the others and should be further pursued and practiced to arrive at more or less rigid rules for generating and combining multiple clusterings.

Even more important seem directions such as developing a comprehensive model of data involving cluster structures, modelling clusters as real world phenomena, and pursuing the consistency criterion in multifaceted data sets.

### Acknowledgments

The author is grateful to the anonymous referees and editors for useful comments. Partial financial support of the Laboratory for Decision Choice and Analysis, Higher School of Economics, Moscow RF, is acknowledged.

### References

- Banfield JD, Raftery AE Model-based Gaussian and non-Gaussian clustering, *Biometrics*, 1993 49: 803-821.
- Bel Mufti G, Bertrand P, El Moubarki L Determining the number of groups from measures of cluster stability, In: *Proceedings of International Symposium on Applied Stochastic Models and Data Analysis*, 2005, 404-412.
- Biernacki C, Celeux G, Govaert G Assessing a mixture model for clustering with the integrated completed likelihood, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, 22(7):719-725.
- Bock HH Probability models and hypotheses testing in partitioning cluster analysis. In Arabie P, Hubert L. and De Soete G eds. *Clustering and Classification*. World Scientific Publ., River Edge NJ; 1996 377-453.
- Bouguila N, Ziou D High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on Minimum Message Length, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2007, 29(10):1716-1731.
- Bozdogan H Mixture-model cluster analysis and choosing the number of clusters using a new informational complexity ICOMP, AIC, and MDL model-selection criteria, In H. Bozdogan, S. Sclove, A. Gupta et al. (Eds.), *Multivariate Statistical Modeling* (vol. II), Dordrecht: Kluwer, 1994 69-113.
- Calinski T, Harabasz J A dendrite method for cluster analysis, *Communications in Statistics*, 1974 3(1): 1-27.
- Casillas A, Gonzales De Lena, MT, Martinez H Document clustering into an unknown number of clusters using a Genetic algorithm, *Text, Speech and Dialogue: 6th International Conference*, 2003 43-49.
- Castellano C, Fortunato S, Loreto V Statistical physics of social dynamics, *Reviews of Modern Physics*, 2009 81:591-646.
- Celeux G, Soromento M An entropy criterion for assessing the number of clusters in a mixture model, *Journal of Classification*, 1996 13:195-212.
- Chae SS, Dubien JL, Warde WD A method of predicting the number of clusters using Rand's statistic, *Computational Statistics and Data Analysis*, 2006 50(12): 3531-3546.
- Cheung Y Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection, *IEEE Transactions on Knowledge and Data Engineering* 2005 17(6):750-761.

- Chiang M-T, Mirkin B Intelligent choice of the number of clusters in K-means clustering: an experimental study with different cluster spreads, *Journal of Classification* 2010, 28:, doi: 10.1007/s00357-010- (to appear).
- Dotan-Cohen D, Kasif S, Melkman AA Seeing the forest for the trees: using the Gene Ontology to restructure hierarchical clustering *Bioinformatics* 2009 25(14):1789-95.
- Duda RO, Hart PE *Pattern Classification and Scene Analysis*, 1973 New York, Wiley.
- Dudoit S, Fridlyand J A prediction-based resampling method for estimating the number of clusters in a dataset, *Genome Biology*, 2002 3(7): research0036.1-0036.21.
- Evanno G, Regnaut S, Goudet J Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, *Molecular Ecology*, 2005, 14: 2611-2620.
- Feng Y, Hamerly G PG-means: learning the number of clusters in data, *Advances in Neural Information Processing Systems, 19 (NIPS Proceeding)*, MIT Press, 2006 393-400.
- Figueiredo M, Jain AK Unsupervised learning of finite mixture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002 24(3): 381-396.
- Fraley C, Raftery AE How many clusters? Which clustering method? Answers via model-based cluster analysis, *The Computer Journal*, 1988 41: 578-588.
- Freudenberg JM, Joshi VK, Hu Z, Medvedovic M CLEAN: CLustering Enrichment Analysis, *BMC Bioinformatics* 2009 10:234.
- Hand D, Krzhanowski WJ Optimising k-means clustering results with standard software packages, *Computational Statistics and Data Analysis*, 2005 49: 969-973.
- Hardy A On the number of clusters, *Computational Statistics & Data Analysis* 1996 23: 83-96
- Hartigan JA *Clustering Algorithms*, 1975 New York: J. Wiley & Sons.
- Hu X, Xu L Investigation of several model selection criteria for determining the number of cluster, *Neural Information Processing*, 2004 4(1):
- Ishioka T An expansion of X-means for automatically determining the optimal number of clusters, *Proceedings of International Conference on Computational Intelligence*, 2005 91-96.
- Jain AK, Dubes RC *Algorithms for Clustering Data*, 1988 Prentice Hall.
- Jonnalagadda S, Srinivasan R NIFTI: An evolutionary approach for finding number of clusters in microarray data, *BMC Bioinformatics*, 2009 10:40, <http://www.biomedcentral.com/1471-2105-10-40>.
- Kaufman L, Rousseeuw P *Finding Groups in Data: An Introduction to Cluster Analysis*, 1990 New York: J. Wiley & Son.
- Kim D, Park Y, Park D A novel validity index for determination of the optimal number of clusters, *IEICE Tans. Inf. And Systems* 2001, E84(2): 281-285.
- Krzhanowski W, Lai Y A criterion for determining the number of groups in a dataset using sum of squares clustering, *Biometrics*, 1985 44: 23-34.
- Kuncheva LI, Vetrov DP Evaluation of stability of k-means cluster ensembles with respect to random initialization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005 28: 1798-1808.
- Lo Y, Mendell NR, Rubin DB Testing the number of components in a normal mixture, *Biometrika* 2001 88(3):767-778.
- Lottaz C, Toedling J, Spang R Annotation-based distance measures for patient subgroup discovery in clinical microarray studies, *Bioinformatics*, 2007 23(17): 2256-2264.
- von Luxburg U A tutorial on spectral clustering, *Statistics and Computing*, 2007 17: 395-416.
- McLachlan GJ, Khan N On a resampling approach for tests on the number of clusters with mixture model-based clustering of tissue samples, *Journal of Multivariate Analysis*, 2004 90: 90-1005.
- McLachlan GJ, Peel D *Finite Mixture Models*, 2000, New York: Wiley.
- Milligan GW A Monte-Carlo study of thirty internal criterion measures for cluster analysis, *Psychometrika*, 1981 46: 187-199.

- Milligan GW, Cooper MC An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, 1985 50: 159-179.
- Minaei-Bidgoli B, Topchy A, Punch WF A comparison of resampling methods for clustering ensembles, *International conference on Machine Learning; Models, Technologies and Application (MLMTA04)*, 2004 Las Vegas, Nevada, pp. 939-945.
- Mirkin B Sequential fitting procedures for linear data aggregation model, *Journal of Classification*, 1990 7: 167-195.
- Mirkin B *Mathematical Classification and Clustering*, 1996 Kluwer, New York.
- Mirkin B *Clustering for Data Mining: A Data Recovery Approach*, 2005 Boca Raton Fl., Chapman and Hall/CRC.
- Mirkin B, Camargo R, Fenner T, Loizou G, Kellam P Similarity clustering of proteins using substantive knowledge and reconstruction of evolutionary gene histories in herpesvirus, *Theor Chemistry Accounts* 2010 125:569–581.
- Monti S, Tamayo P, Mesirov J, Golub T Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data, *Machine Learning*, 2003 52: 91-118.
- Mojena R Hierarchical grouping methods and stopping rules: an evaluation, *The Computer Journal*, 1977, 20(4): 359-363.
- Newman MEJ, Girvan M Finding and evaluating community structure in networks, *Physical Review E* 2004, 69: 026113.
- Newman MEJ, Barabasi AL, Watts DJ *Structure and Dynamics of Networks*, 2006 Princeton University Press.
- Pelleg D, Moore A X-means: extending k-means with efficient estimation of the number of clusters, *Proceedings of 17<sup>th</sup> International Conference on Machine Learning*, 2000 San-Francisco, Morgan Kaufmann, 727-734.
- Pena JM, Lozano JA, Larranga P An empirical comparison of four initialization methods for K-Means algorithm, *Pattern Recognition Letters*, 1999 20: 1027-1040.
- Pollard KS, Van Der Laan MJ A method to identify significant clusters in gene expression data, *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2002 107.
- Roth V, Lange V, Braun M, Buhmann J Stability-based validation of clustering solutions, *Neural Computation*, 2004 16: 1299–1323.
- Saha S and Bandyopadhyaya S A symmetry based multiobjective clustering technique for automatic evolution of clusters, *Pattern Recognition*, 2010 43:738-751.
- Shen J, Chang SI, Lee ES, Deng Y, Brown SJ Determination of cluster number in clustering microarray data, *Applied Mathematics and Computation*, 2005 169: 1172-1185.
- Shubert J, Sedenbladh H Sequential clustering with particle filters – Estimating the number of clusters from data, *Proceedings of 8<sup>th</sup> International Conference on Information Fusion*, IEEE, Piscataway NJ, 2005 A4-3, 1-8.
- Springel V, White SDM, Jenkins A, Frenk CS, Yoshida N, Gao L, Navarro J, Thacker R, Croton D, Helly J, Peacock JA, Cole S, Thomas P, Couchman H, Evrard A, Colberg J, Pearce F Simulations of the formation, evolution and clustering of galaxies and quasars, *Nature*, 2005 435(2): 629-636.
- Steinley D K-means clustering: A half-century synthesis, *British Journal of Mathematical and Statistical Psychology*, 2006 59: 1-34.
- Steinley D, Brusco M. Initializing K-Means batch clustering: A critical evaluation of several techniques, *Journal of Classification*, 2007 24: 99-121.
- Sugar CA, James GM Finding the number of clusters in a data set: An information-theoretic approach, *Journal of American Statistical Association*, 2003 98: 750-778.
- Tibshirani R, Walther G, Hastie T Estimating the number of clusters in a dataset via the Gap statistics, *Journal of the Royal Statistical Society B*, 2001 63: 411-423.

- Tibshirani R, Walther G Cluster validation by prediction strength, *Journal of Computational and Graphical Statistics*, 2005 14: 511-528.
- Vapnik V *Estimation of Dependences Based on Empirical Data*, Springer Science+ Business Media Inc., 2d edition 2006.
- Windham MP, Cutler A Information ratios for validating mixture analyses, *Journal of the American Statistical Association*, 1992 87: 1188-1192.
- Yan M, Ye K Determining the number of clusters using the weighted gap statistic, *Biometrics*, 2007 63:1031-1037.
- Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL Model-based clustering and data transformations for gene expression data, *Bioinformatics*, 2001, 17:977-987.