A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 255, 2011

Ewa Witek[*]

THE COMPARISON OF MODEL-BASED CLUSTERING WITH HEURISTIC CLUSTERING METHODS

Abstract. Most clustering done in practice is based largely on heuristic but intuitively reasonable procedures. This article reviews the model-based approach to clustering, based on probability models and presents the comparison with well known hierarchical and iterative relocation clustering methods.

Key words: Model-based clustering (MBC), heuristic methods of clustering, probability models.

I. MODEL-BASED AND HEURISTIC METHODS OF CLUSTERING

Cluster analysis is the identification of groups of observations that are cohesive and separated from other groups. Most clustering done in practice is based largely on heuristic but intuitively reasonable procedures. One widely used class of methods involves hierarchical agglomerative clustering, in which two groups chosen to optimize some criterion are merged at each stage of the algorithm. Popular criteria include the sum of within group sums of squares (Ward method) and the shortest distance between groups, which underlies the single-linkage method. Another common class of methods is based on iterative relocation (also called iterative partitioning), in which data points are moved from one group to another until there is no further improvement in some criterion. Iterative relocation with the sum of squares criterion is often called *k*-means clustering. However, the statistical properties of these methods are generally unknown, precluding the possibility of formal inference.

It was realized that cluster analysis can also be based on probability models (see Bock 1996). It has also been shown that some of the most popular heuristic clustering methods are approximate estimation methods for certain probability models (see Fraley 2002, Witek 2009 a). Using mixture models we can provide a principled statistical approach to the practical questions that arise in applying clustering methods. In finite mixture models, each component probability

[*] Ph.D student, Department of Statistics, The Karol Adamiecki University of Economics, Katowice.

distribution corresponds to a cluster. The problems of determining the number of clusters and of choosing an clustering method can be recast as statistical model choice problems, and models that differ in numbers of component distributions can be compared.

II. PARAMETER ESTIMATION AND MODEL SELECTION

In model-based clustering, individual clusters are described by multivariate normal distributions, where the class labels, parameters and proportions are unknown. The data $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^{\mathrm{T}}$ are assumed to be generated by a mixture with density:

$$f(\mathbf{x}) = \prod_{i=1}^{n} \sum_{s=1}^{u} \tau_s f_s(\mathbf{x_i} | \mathbf{\Theta}_s), \tag{1}$$

where $f_s(\mathbf{x_i} | \mathbf{\Theta}_s)$ is a probability distribution with parameters $\mathbf{\Theta}_s$, and τ_s is the probability of belonging to the *s*th component. The parameters of the model are usually estimated by maximum likelihood using the Expectation-Maximization (EM) algorithm (Dempster et al. 1977). Each EM iteration consist of two steps – an E-step and an M-step. Given an initial guess for the cluster means $\boldsymbol{\mu}_s$, covariances $\mathbf{\Sigma}_s$ and proportions τ_s, the E-step calculates the conditional probability that object *i* belongs to the *s*th component:

$$\hat{\mathbf{z}}_{is} = \frac{\hat{\tau}_s f_s(\mathbf{x}_i | \hat{\boldsymbol{\mu}}_s, \hat{\mathbf{\Sigma}}_s)}{\sum_{r=1}^{u} \hat{\tau}_r f_r(\mathbf{x}_i | \hat{\boldsymbol{\mu}}_r, \hat{\mathbf{\Sigma}}_r)} \tag{2}$$

The maximization step (M-step) consists of estimating the parameters from the data and the conditional probabilities $\hat{\mathbf{z}}_{is}$. The E- and M-steps iterate until convergence. Finally, each object is classified in the class in which it has the highest conditional or posterior probability. The results of the EM are highly dependent on the initial values, model-based hierarchical clustering can be a good solution (Fraley and Raftery 1998; Dasgupta and Raftery 1998).

In order to select the optimal clustering model, several measures have been proposed (McLachlan and Peel [2000]). Three information criteria are available in *flexmix* package of **R**: BIC (*Bayesian Information Criterion*), AIC (*Akaike Information Criterion*) and ICL (*Integrated Completed Likelihood*). The criteria are defined as

$$BIC_s = -2\log p(\mathbf{x}, y|\hat{\boldsymbol{\Theta}}_s, M_s) + v_s \log(n), \quad (3)$$

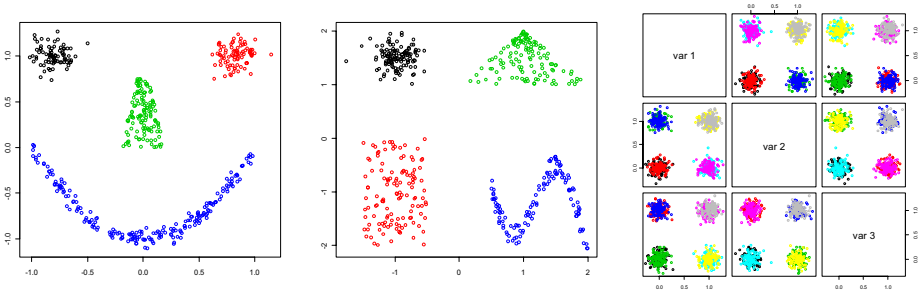
$$AIC_s = -2\log p(\mathbf{x}, y|\hat{\boldsymbol{\Theta}}_s, M_s) + 2v_s, \quad (4)$$

$$ICL_s = -2\log p(\mathbf{x}, y, \mathbf{z}|\hat{\boldsymbol{\Theta}}_s, M_s) - \frac{v_s}{2}\log(n), \quad (5)$$

where $\log p(\mathbf{x}, y, \mathbf{z}|\hat{\boldsymbol{\Theta}}_s, M_s)$ is the maximized loglikelihood for the model M_s, v_s is the number of parameters to be estimated in the mixture model and n is the sample size. See model-based clustering for more details in i.e. Fraley (2002), Witek (2009 a, b).

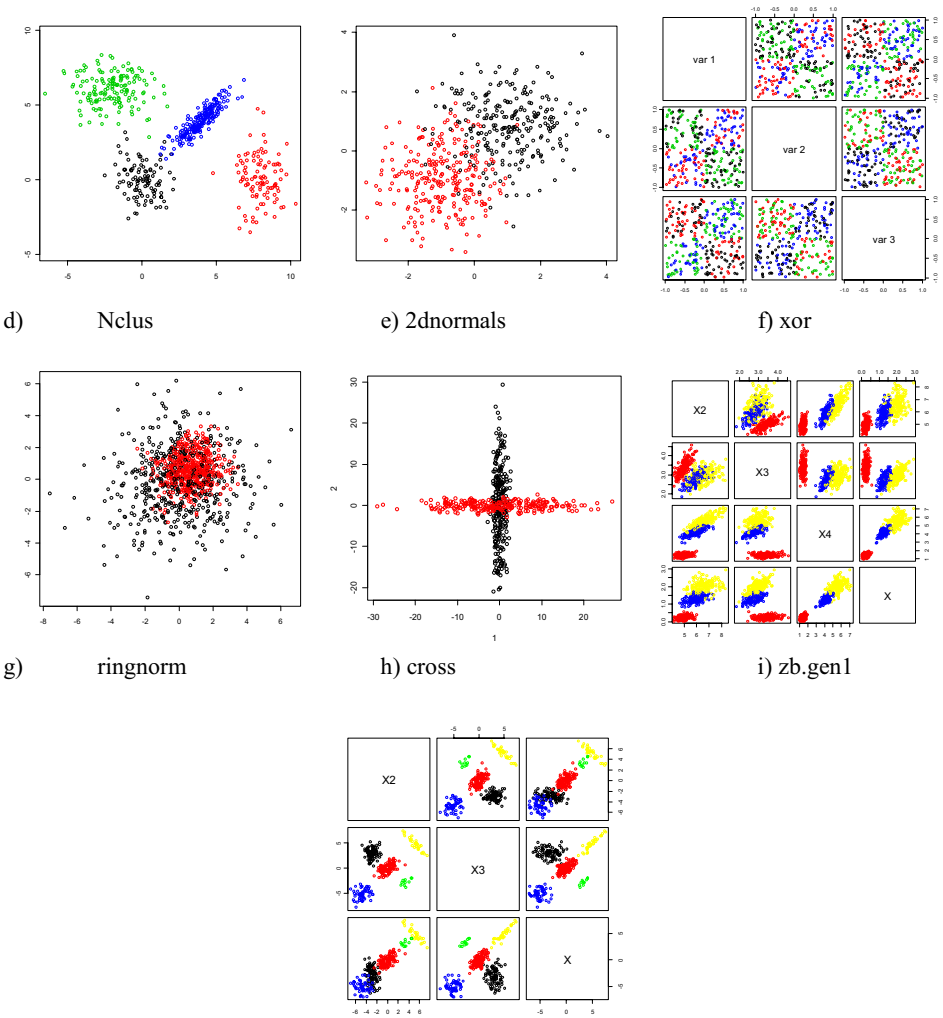
III. EXAMPLE

The main goal of this example is to compare different methods of clusters estimation and the quality of partitions for heuristic and model-based clustering. We analyzed 10 artificially generated datasets[1]. There is the graphical presentation of the data given in the Fig. 1. All computations and graphics in this paper were done in cluster, flexclust, clusterSim, flexmix, mclust and mlbench packages of R (version 2.8).



a) smiley b) shapes c) corners

[1] The data are available in `mlbench`, `flexmix`, `mclust` packages, zb_gen1 and zb_gen2 was generated by `cluster.Gen` function (`cluster.Sim` package of **R**).



d) Nclus e) 2dnormals f) xor

g) ringnorm h) cross i) zb.gen1

j) zb.gen2

Figure 1. The graphical characteristic of datasets
Source: Own research.

We can see that the data is characterized by different number of clusters and different level of separation. In the first part of our analysis we investigated different clusters selection measures for model-based and heuristic methods of clustering. To determine the number of clusters for well-known heuristic methods mostly Davies and Bouldin, Calinski and Harabasz, Baker and Hubert, Hubert and Levine, Gap, Krzanowski and Lai, Silhouette indices are proposed.

We calculated each of them for k-means, k-medoids, Ward and single linkage methods of clustering. Table 1 reports just Silhouette index results, aggregated over 4 heuristic methods of clustering. Firstly we aggregated the number of clusters for each index and all deterministic clustering methods on the basis on mode. Secondly we chosen index which was given the true number of clusters most frequently (we could compare the different cluster selection measures because the true number of datasets is known). Table 1 presents the BIC criterion- model-based measure of clusters estimation as well. The comparison of AIC, BIC and ICL criterion results was shown that overall BIC performs best, as was found in other studies - model choice based on BIC has given good results in a range of applications (i.e. Kass and Raftery 1995, Keribin 2000). We can see in Table 1 that BIC criterion 8 times, while the Silhouette index only 5 times suggested the right number of clusters.

Table 1 Different methods of clusters estimation- results

Dataset	BIC	Silhouette	Clusters
smiley	8	6	4
shapes	8	4	4
corners	8	8	8
Nclus	4	4	4
2dnormals	2	2	2
xor	4	4	4
ringnorm	2	3	2
cross	2	5	2
zb.gen.1	3	2	3
zb.gen.2	5	4	5

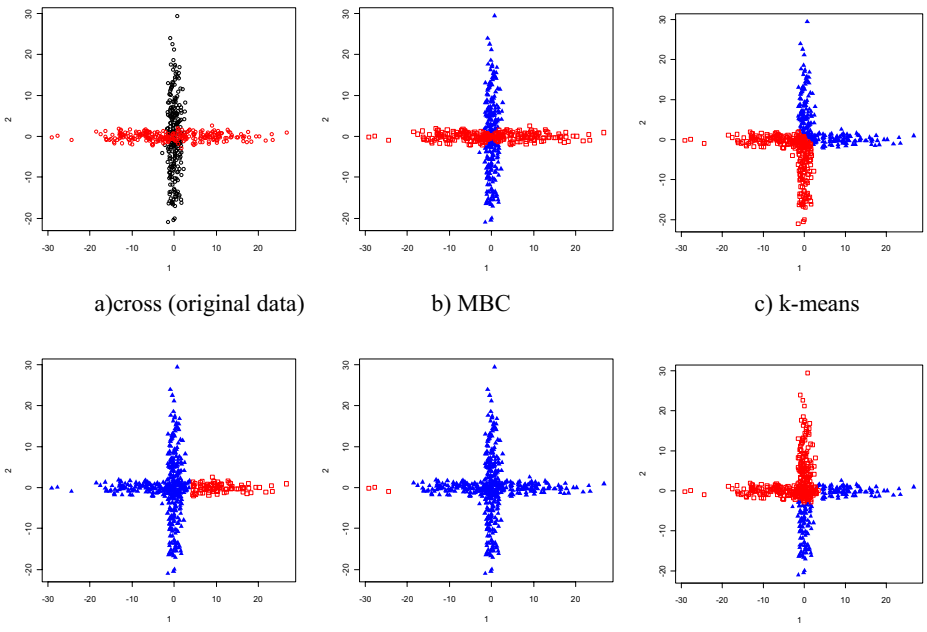
Source: Own research.

In the second part of our analysis we compared the quality of partitions for heuristic and model-based clustering approach. Since we know the group memberships in advance, we can check the clustering quality measured by the adjusted Rand index. The results are presented in Table 2. We can see that for well-separated datasets (i.e. smiley, shapes, corners) quality of partitions is very good for almost each clustering methods (Rand index equals to one). In the other side, for low-separated datasets the quality is rather poor for each of them. However, in general the model-based clustering gives equal or the highest quality of partition for each datasets in comparison with heuristic methods of clustering. It is worth to notice that the model-based clustering gives significantly better result for overlapping clusters datasets i.e. cross dataset (see Table 2 and Fig. 3).

Table 2 Rand index measures for heuristic and MBC (model-based clustering) methods

Dataset	k-means	k-medoids	single linkage	Ward	MBC
smiley	0,5848	0,8597	**1**	**1**	**1**
shapes	**1**	**1**	**1**	**1**	**1**
corners	0,8314	**1**	**1**	**1**	**1**
Nclus	0,9399	0,9350	0,2621	0,9661	**0,9796**
2dnormals	0,6265	0,6202	0	0,5408	**0,6329**
xor	0,1486	0,1780	0,0003	0,1745	**0,2326**
ringnorm	0,1259	0,1203	0	0,1275	**0,2573**
cross	0,0016	0,1173	0,0001	0,0018	**0,7322**
zb.gen.1	0,7597	0,7721	0,5902	0,8545	**0,9542**
zb.gen.2	0,6483	0,4797	0,2275	0,4687	**1**

Source: Own research.



a)cross (original data) b) MBC c) k-means

d) k-medoids e) single linkage f) Ward

Figure 2. Clustering results for cross dataset

Source: Own research.

IV. CONCLUSIONS

We compared different methods of clusters estimation and quality of partitions for heuristic and model-based clustering methods. Although we have received much better results for BIC than Silhouette index measure, on the basis on the real datasets analysis, we would like to stress that there is no dominating "good" method for the number of clusters estimation. Some are good only in some specific simulations or examples. However it is worth to stress that a difficulty of some of the more heuristic clustering algorithms is the lack of a statistically principled method for determining the number of clusters. Model–based clustering is an inferentially based procedure- use model selection methods to make this decision. As far as the quality of partition is concerned we can say that for the true number of clusters, model-based clustering gives much better results.

REFERENCES

Bock H.H., (1996*), Probabilistic models in cluster analysis*, Computational Statistics and Data Analysis, 23, 5-28.

Dasgupta A., Raftery A.E. (1998), *Detecting features in spatial point processes with clutter via model-based clustering*, „Journal of the American Statistical Association", 93, 294-302.

Dempster A.P., Laird N.M., Rubin D.B. (1977), *Maximum likelihood for incomplete data via the EM algorithm (with discussion)*, „Journal of the Royal Statistical Society", ser. B, 39, 1–38.

Fraley C., Raftery A.E. (1998), *How many clusters? Which clustering method? Answers via model-based cluster analysis*, „The Computer Journal", 41, 577-588.

Fraley C., Raftery A.E. (2002), *Model-based clustering, discriminant analysis, and density estimation*, „Journal of the American Statistical Association", 97, 611–631.

Kass R.E., Raftery A.E. (1995), Bayes Factors, *Journal of the American Statistical Association*, 90, 928–934.

Keribin, C., Consistent estimation of the order of mixture models. Sankhya Indian J. Stat. v. 62. 49–66.

McLachlan G.J., Peel D. (2000), *Finite mixture models*, Wiley, New York.

Witek E., 2009 (a), *Analiza skupień-podejście modelowe*, in: M. Walesiak, E. Gatnar (ed.), *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa, s. 434–462.

Witek E., 2009 (b), *On an improvement of the model-based clustering method,* in: Cz. Domański, J. Białek (ed.), *Multivariate statistical analysis statistical inference, statistical models and applications,* Wydawnictwo Uniwersytetu Łódzkiego, s. 229–235.

Schwarz G. (1978), *Estimating the dimension of a model*, „The Annals of Statistics", 6, 461–464.

Ewa Witek

PORÓWNANIE ANALIZY SKUPIEŃ OPARTEJ NA MODELACH Z KLASYCZNYMI METODAMI TAKSONOMICZNYMI

Najczęściej w różnych analizach statystycznych wykorzystywane są klasyczne metody analizy skupień, opierające się na podejściu heurystycznym. W referacie zaprezentowane zostanie podejście modelowe w analizie skupień (*model-based clustering*), bazujące na modelach probabilistycznych. W części empirycznej referatu podejście to zostanie porównane z klasycznymi metodami taksonomicznymi (metodami hierarchicznymi oraz metodami iteracyjno-aglomeracyjnymi).