Fuzzy Cluster Validation Using the Partition Negentropy Criterion

Luis F. Lago-Fernández[1], Manuel Sánchez-Montañés[1], and
Fernando Corbacho[2]

[1] Departamento de Ingeniería Informática, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Madrid, Spain
[2] Cognodata Consulting, Calle Caracas 23, 28010 Madrid, Spain

Abstract. We introduce the Partition Negentropy Criterion (PNC) for cluster validation. It is a cluster validity index that rewards the average normality of the clusters, measured by means of the negentropy, and penalizes the overlap, measured by the partition entropy. The PNC is aimed at finding well separated clusters whose shape is approximately Gaussian. We use the new index to validate fuzzy partitions in a set of synthetic clustering problems, and compare the results to those obtained by the AIC, BIC and ICL criteria. The partitions are obtained by fitting a Gaussian Mixture Model to the data using the EM algorithm. We show that, when the real clusters are normally distributed, all the criteria are able to correctly assess the number of components, with AIC and BIC allowing a higher cluster overlap. However, when the real cluster distributions are not Gaussian (i.e. the distribution assumed by the mixture model) the PNC outperforms the other indices, being able to correctly evaluate the number of clusters while the other criteria (specially AIC and BIC) tend to overestimate it.

Key words: Clustering, Cluster validation, Mixture Model, Negentropy, EM algorithm

1 Introduction

Cluster analysis [1] deals with the automatic partition of a data set into a finite number of natural structures, or clusters. The elements inside a cluster must be similar, while those belonging to different clusters must not. Clustering algorithms are usually divided into crisp and fuzzy. In crisp clustering, each data point is uniquely assigned to a single cluster. On the contrary, fuzzy clustering allows each point to belong to any of the clusters with a certain degree of membership.

A standard approach in fuzzy clustering is model clustering, where it is assumed that the observed data are generated from a mixture of probability distributions (clusters or components). The mathematical structure of these distributions is assumed to be of a certain type (usually Gaussians) but the specific parameters (e.g. means and covariances) must be found. Once the number of components and their parameters have been selected following some strategy, the degree of membership of the point x with respect to the cluster c is usually related to the probability $p(c|x)$.

There exist different methodologies to select the parameters of the mixture model, the most popular being the Expectation-Maximization (EM) algorithm [2]. However, a common problem is how to determine the correct number of components in the mixture. A different but closely related problem is how to measure the validity of the outcomes of a particular fuzzy clustering method. This is the subject of cluster validation [3], whose objective is to provide a quality measure, or validity index, that allows to evaluate the results obtained by a clustering algorithm. Many cluster validity indices have been proposed in the literature, including geometric [4–6], probabilistic [7–9], graph theoretic [10], and visual [11, 12] approaches.

In the context of density estimation using mixture models, different strategies have been explored to automatically select the number of mixture components [13–16], the most popular being the Akaike's Information Criterion (AIC) [17] and the Bayesian Inference Criterion (BIC) [18, 19]. The last two approaches are based on the maximization of criteria that combine a term based on the likelihood of the observations and a term that penalizes the complexity of the model. In principle, they could also be used as validity indices in model clustering, to compare models with different number of components and thus automatically select the number of clusters. This possibility has been deeply explored in the literature, and different indices based on these and similar criteria have been proposed to validate clustering partitions and assess the number of components in clustering problems using mixture models [7, 8, 20]. In particular, the Information Completed Likelihood (ICL) [21, 22], which is essentially the BIC criterion penalized by subtraction of the estimated partition mean entropy, has been shown to outperform AIC and BIC when the focus is clustering rather than density estimation. The AIC and BIC criteria do not explicitly penalize the overlap amongst the clusters, and so they tend to overestimate the number of cluster components when the kind of distribution followed by the real clusters does not match the distribution assumed by the mixture model [21].

In this article we present the *Partition Negentropy Criterion* (PNC), a new cluster validity index which combines a term that rewards the average normality of the clusters and a term that penalizes the average overlap. The normality of a cluster is computed using the negentropy, while the average overlap is measured by the partition entropy. Here we use the new index to validate Gaussian mixtures that are fitted with the EM algorithm, and compare its performance to the AIC, BIC and ICL criteria in a set of synthetic clustering problems. We first check that the PNC is able to assess the correct number of components in problems where the underlying cluster distributions are Gaussian. When the Gaussian clusters are highly overlapped the AIC and BIC criteria obtain a slightly better detection rate, that is they are able to assess the number of components under higher overlap, than the ICL and the PNC. However, in situations where the underlying cluster distributions are not purely Gaussian, the AIC and BIC criteria systematically overestimate the number of components, regardless of the separation amongst the clusters. In these cases only the ICL and the PNC obtain satisfactory results. In particular, the proposed PNC index provided a high detection rate and the lowest overestimation rate in all the tests performed.

2 The Partition Negentropy Criterion

In this section we develop the Partition Negentropy Criterion, a cluster validity index whose aim is to find well separated clusters as normally distributed as possible. The normality of a cluster is characterized by means of its negentropy, a standard measure of distance to normality which computes the difference between the cluster's entropy and the entropy of a Gaussian distribution with the same covariance matrix [23]. The negentropy of a continuous random variable $\mathbf{X}$ is defined as:

$$J(\mathbf{X}) = \hat{H}(\mathbf{X}) - H(\mathbf{X}) \tag{1}$$

where $H(\mathbf{X})$ is the differential entropy of $\mathbf{X}$ and $\hat{H}(\mathbf{X})$ is the differential entropy of a normal distribution with the same covariance matrix. The Gaussian distribution maximizes the differential entropy for a given covariance matrix [24], so the negentropy is always non-negative, being zero if and only if $\mathbf{X}$ is normally distributed. The maximum entropy property associated to the normal distribution also provides a hint on why normality is a desired property of any cluster. Maximum entropy, or equivalently maximum uncertainty, implies minimum structure, and so a normally distributed cluster can not be expected to contain other substructures.

Let us consider a set of data points $\{\mathbf{x}_i\}$ and a fuzzy partition $p(c|\mathbf{x}_i)$ into a set of clusters C. Our goal is to obtain a cluster validity index that rewards partitions into well separated Gaussian clusters. We will measure the quality of the partition by:

$$H(C|\mathbf{X}) + J(\mathbf{X}|C) \tag{2}$$

The first term measures the average degree of overlap amongst the clusters, while the second corresponds to the average negentropy of the clusters, which measures how distant they are from a Gaussian distribution. So minimization of this expression will favour partitions that consist of well separated normally distributed clusters. We can write $J(\mathbf{X}|C)$ as:

$$J(\mathbf{X}|C) = \hat{H}(\mathbf{X}|C) - H(\mathbf{X}|C) \tag{3}$$

And, using basic properties of the conditional entropy [24], rewrite it as:

$$J(\mathbf{X}|C) = \hat{H}(\mathbf{X}|C) + H(C) - H(\mathbf{X}) - H(C|\mathbf{X}) \tag{4}$$

Note that the term $H(\mathbf{X})$ is constant for the problem, and so it can be ignored when minimizing the expression in 2. The term $\hat{H}(\mathbf{X}|C)$ can be expressed as:

$$\hat{H}(\mathbf{X}|C) = \sum_{c=1}^{n_c} p(c)\hat{H}(\mathbf{X}|c) \tag{5}$$

where the sum extends to all the n_c clusters in C, $p(c)$ is the a-priori probability of cluster c, and $\hat{H}(\mathbf{X}|c)$ is the differential entropy of the cluster c assuming normality, that is:

$$\hat{H}(\mathbf{X}|c) = \frac{1}{2}\log|\Sigma_c| + \frac{d}{2}\log 2\pi e \tag{6}$$

where Σ_c is the covariance matrix of cluster c and d is the dimension of $\mathbf{X}$. If we substitute equations 4, 5 and 6 into expression 2, and neglect terms that do not depend on the partition C, we obtain the Partition Negentropy Criterion as:

$$PNC(C) = \frac{1}{2}\sum_{c=1}^{n_c} p(c)\log|\Sigma_c| - \sum_{c=1}^{n_c} p(c)\log p(c) \tag{7}$$

Given different partitions of a data set, we will select that with a lower PNC.

3 Evaluation of the PNC

To test the new cluster validity criterion we use the PNC to validate fuzzy partitions resulting from the application of the EM algorithm to a set of synthetic problems, and compare the results to those obtained with the AIC, the BIC, and the ICL criteria. For every problem we follow the same approach. First, the EM algorithm is used to fit a set of Gaussian mixtures with different number of components. Then, for each index (AIC, BIC, ICL, PNC) we select the mixture which provides the best index value. We compute the PNC by directly substituting the covariance matrices and the prior probabilities given by the EM algorithm into equation 7.

3.1 Two Simple Examples

We will first illustrate the PNC with two simple examples consisting of three well separated clusters in two dimensions. In the first case each cluster consists of 1000 points drawn from a normal distribution with covariance matrix equal to the identity matrix, $\Sigma = \mathbf{I}$, and centered at $\mu_1 = (0,0)$, $\mu_2 = (5,0)$, and $\mu_3 = (5,5)$ respectively (see figure 1). For this problem we have run the EM algorithm to fit a mixture of n_c Gaussians, with $n_c \in \{1,2,3,4,5\}$. The algorithm has been run 10 times for each n_c. The best partitions according to each of the four validity criteria are shown in figures 1A (AIC), 1B (BIC), 1C (ICL), and 1D (PNC). The solid lines represent the contours of the Gaussian components. Note that for well separated normal clusters all the criteria select partitions with the correct number of clusters ($n_c = 3$).

The second problem consists of three non-Gaussian clusters of 1000 points each. In polar coordinates, the clusters follow a gamma distribution in the radius

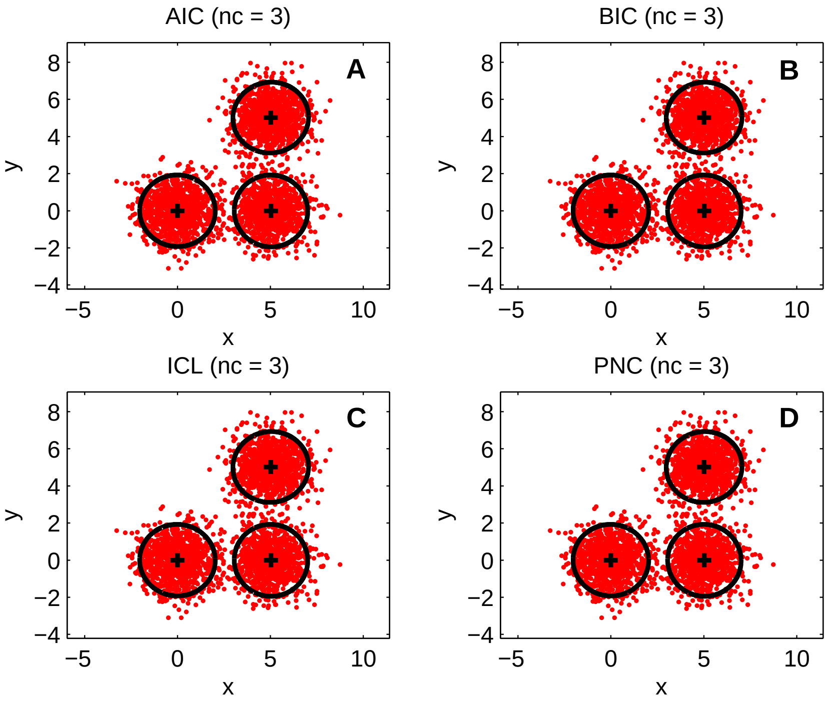

Fig. 1. Assessment of the number of clusters in a problem consisting of a mixture of three Gaussian distributions in 2D. The EM algorithm has been run 10 times for each number of components (ranging from 1 to 5), and the best solution according to the four criteria is selected. **A.** Akaike's criterion (AIC). **B.** Bayesian inference criterion (BIC). **C.** Information Completed Likelihood (ICL). **D.** Partition Negentropy Criterion (PNC).

and a uniform distribution in the angle. The gamma distribution has a shape parameter $k = 2$ and a scale parameter $\theta = 1.5$. The three clusters are centered at $\mu_1 = (0, 0)$, $\mu_2 = (15, 0)$, and $\mu_3 = (15, 15)$ respectively. As before, we have run the EM algorithm 10 times for each n_c, and we have selected the best partitions according to the four validity criteria. The results are shown in figures 2A (AIC), 2B (BIC), 2C (ICL), and 2D (PNC). Note that, although the clusters are easily separable, only the PNC is able to correctly assess the number of clusters ($n_c = 3$). The other criteria overestimate the number of components. The AIC and BIC select partitions with $n_c = 5$ clusters, while the ICL selects a partition with $n_c = 4$ clusters.

These examples show that standard criteria such as AIC, BIC or ICL can perform poorly as cluster validity indices when the underlying data distribution does not match the assumed mixture model. On the other hand, the PNC shows a good performance even when the cluster distributions are not pure Gaussians.

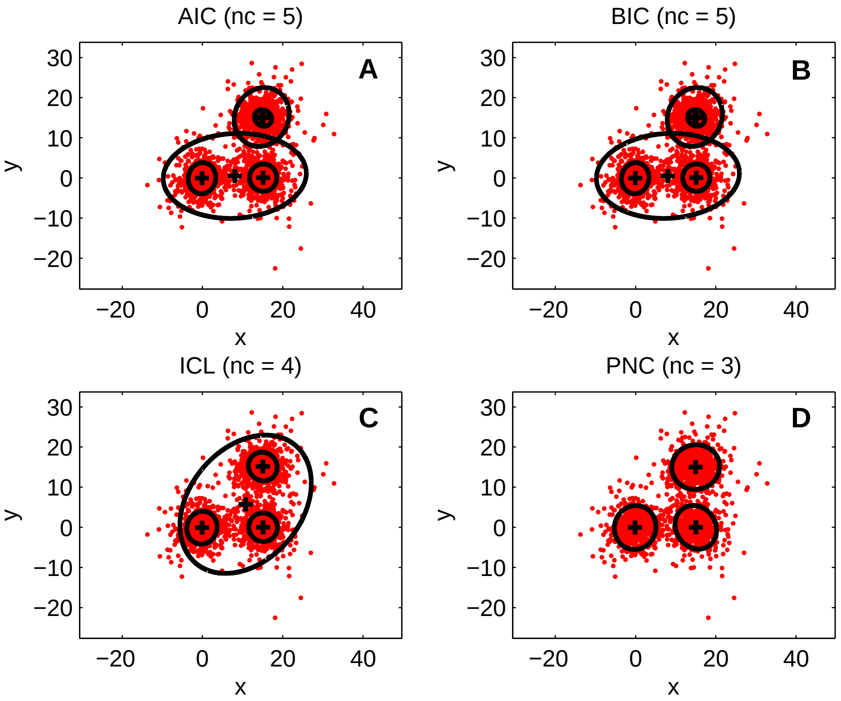

Fig. 2. Assessment of the number of clusters in a problem consisting of a mixture of three non-Gaussian distributions in 2D. The EM algorithm has been run 10 times for each number of components (ranging from 1 to 5), and the best solution according to the four criteria is selected. **A.** Akaike's criterion (AIC). **B.** Bayesian inference criterion (BIC). **C.** Information Completed Likelihood (ICL). **D.** Partition Negentropy Criterion (PNC).

3.2 Number of Detected Clusters versus Inter-Cluster Distance

As a second test we consider the assessment of the number of components in problems consisting of two spherical clusters with covariance matrices equal to the identity matrix, $\varSigma = \mathbf{I}$, and 1000 points each in two dimensions. The distance between the cluster centers is varied in order to obtain problems with different degree of overlap. We want to study the performance of the PNC and the other criteria as a function of the overlap for different cluster shapes. The shape of the clusters is modeled according to the following four kinds of probability distributions:

1. *Pure normal distribution.* Its covariance matrix is equal to the identity matrix, $\varSigma = \mathbf{I}$.
2. *Truncated normal distribution.* Points are generated using a pure normal distribution, and those whose distance to the mean exceeds 1.8 times the

standard deviation are discarded. A scaling factor is applied to ensure that the resulting distribution has a covariance matrix $\Sigma = \mathbf{I}$.

3. *Uniform distribution inside a circle.* The circle radius is selected such that the covariance matrix is $\Sigma = \mathbf{I}$.
4. *Gamma-uniform distribution.* The same distribution as in the second example of section 3.1 is used. The shape parameter of the gamma is $k = 2$, and the scale parameter is selected to ensure that the resulting distribution has a covariance matrix $\Sigma = \mathbf{I}$.

In all the cases the first cluster is centered at $\mu_1 = (0, 0)$ and the second one at $\mu_2 = (d, 0)$, where d is the inter-cluster distance. For each problem and for values of d ranging between 0 and 5, we use the EM algorithm to fit a Gaussian mixture with a number of components $n_c \in \{1, 2, 3, 4, 5\}$. As before, the EM algorithm is run 10 times for each n_c and the best partition according to each of the four criteria is selected. A total of 40 different problems are generated for each d in order to average. In figure 3 we plot the average number of clusters in the best partition selected by each criterion versus the inter-cluster distance. When the clusters follow a Gaussian distribution (figure 3A), the average number of clusters selected by all the criteria is between 1 and 2. As we can see AIC and BIC can stand a higher overlap than the other indices. On the other hand, when the underlying cluster distribution is not Gaussian (figures 3B, 3C and 3D) AIC and BIC tend to overestimate the number of clusters for any degree of overlap. In these cases only the ICL and PNC criteria provide satisfactory results. The ICL criterion admits a slightly higher overlap, but it also overestimates the number of components for gamma-uniform distributed clusters at high d. At the price of a higher tendency to merge overlapping clusters, the proposed PNC is the only index that correctly assesses the number of clusters for all the considered cluster shapes when the clusters are well separated.

3.3 Number of Clusters in Randomly Generated Problems

Finally, we make the last test using 1000 randomly generated problems. Each problem contains three clusters in two dimensions, but the shape, orientation, position and scale of the clusters are selected randomly. The 4 cluster shapes of section 3.2 are considered. As for previous tests, the EM algorithm is used to fit a Gaussian mixture model to each problem. We try different number of components, $n_c \in \{1, 2, 3, 4, 5\}$, and the algorithm is run 5 times for each n_c. Then the four validity criteria are used to select a preferred partition. In table 1 we show the percentage of problems for which each criterion selects a partition with 1, 2, 3, 4 or 5 clusters. Note that the PNC is the one which selects the correct number of clusters ($n_c = 3$) with a higher probability (76% of the problems). In addition, it only overestimates the number of clusters for the 9% of the problems. The ICL follows closely (72.4% of correct partitions and 15.5% of overestimation), but AIC and BIC perform very poorly and they only select the correct number of clusters for the 11.6% and the 26% of the problems respectively.

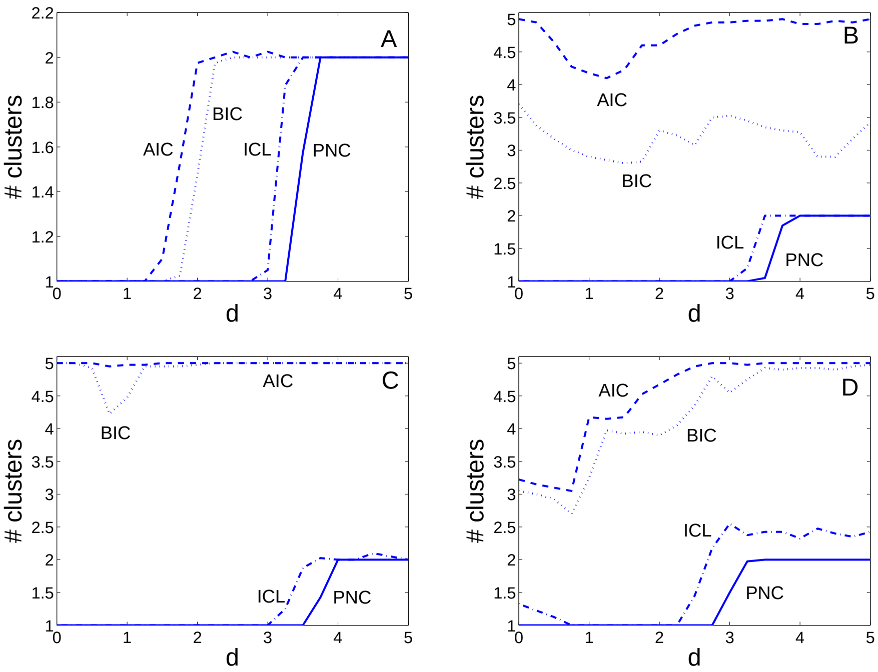

Fig. 3. Average number of clusters selected by the four considered validity criteria versus inter-cluster distance. **A.** The two clusters are normally distributed. **B.** The clusters are Gaussians without tails. **C.** The clusters are uniformly distributed inside a circle. **D.** The clusters follow a gamma-uniform distribution.

4 Discussion

In this article we have presented the Partition Negentropy Criterion (PNC), a cluster validity index whose aim is to find clearly separated clusters whose shape is as Gaussian as possible. The index measures the normality of the clusters using the negentropy, and it measures the cluster separation using the partition entropy. We investigated the performance of the PNC using synthetic clustering problems, where the ability to assess the number of clusters was compared to the AIC, BIC and ICL criteria. We first checked that the PNC is able to determine the correct number of clusters in problems whose data points are generated from a mixture of well separated Gaussian distributions. We observed, however, that the performance degrades when the Gaussians are overlapped. In these cases the PNC obtains a similar performance to ICL, but slightly worst than AIC and BIC. On the other hand, we showed that the PNC provides good results even when the data are generated from mixtures of non-Gaussian distributions. In such situations the PNC is still able to detect the correct number of clusters, outperforming the other criteria, which are more prone to overestimate this number.

Table 1. Percentage of data sets for which each criterion predicts 1, 2, 3, 4 or 5 clusters.

n_c	AIC	BIC	ICL	PNC
1	0.0	0.0	0.2	0.6
2	0.3	0.5	11.9	14.4
3	11.6	26.0	72.4	76.0
4	27.7	38.7	14.1	5.2
5	60.4	34.8	1.4	3.8

Although the results here presented are promising, future work using real datasets is needed in order to validate the proposed criterion. On the other hand, the mathematical simplicity of the PNC, whose evaluation involves just the computation of the determinants of the covariance matrices of each cluster, may permit an analytical study of its performance. This is specially interesting as it would allow to compare the PNC to other indices at the mathematical level.

In the present work, the calculation of the mixture model parameters via EM is conceptually separated from the validation step, in which the validity indices are used to evaluate the outcomes of a particular run of the algorithm. Given the mathematical simplicity of the PNC, we believe that it could be possible to integrate it into an EM algorithm, thus obtaining an iterative procedure that simultaneously searches for the number of clusters and the mixture model parameters.

Finally, we will investigate how the PNC could be extended to deal with variables which are not continuous.

Acknowledgments. This work has been partially supported with funds from MEC BFU2006-07902/BFI, CAM S-SEM-0255-2006 and CAM/UAM project CCG08-UAM/TIC-4428.

References

1. Everitt, B., Landau, S., Leese, M.: Cluster Analysis. Hodder Arnold, London (2001)
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. J. Royal Statistical Soc. B 39, 1-38 (1977)
3. Gordon, A.D.: Cluster Validation. In: Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.H., Baba, Y. (eds.) Data Science, Classification and Related Methods, pp. 22-39. Springer-Verlag, New York (1998)
4. Bezdek, J.C., Pal, R.N.: Some New Indexes of Cluster Validity. IEEE Trans. Systems, Man and Cybernetics B 28, 3, 301-315 (1998)
5. Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: Validity Index for Crisp and Fuzzy Clusters. Pattern Recognition 37, 3, 487-501 (2004)

6. Bouguessa, M., Wang, S., Sun, H.: An Objective Approach to Cluster Validation. Pattern Recognition Letters 27, 13, 1419-1430 (2006)
7. Bozdogan, H.: Choosing the Number of Component Clusters in the Mixture-Model Using a New Information Complexity Criterion of the Inverse-Fisher Information Matrix. In: Opitz. O., Lausen, B., Klar, R. (eds.) Data Analysis and Knowledge Organization, pp. 40-54. Springer-Verlag, Heidelberg (1993)
8. Biernacki, C., Celeux, G., Govaert, G.: An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture Model. Pattern Recognition Letters 20, 3, 267-272 (1999)
9. Geva, A.B., Steinberg, Y., Bruckmair, S., Nahum, G.: A Comparison of Cluster Validity Criteria for a Mixture of Normal Distributed Data. Pattern Recognition Letters 21, 6-7, 511-529 (2000)
10. Pal, N.R., Biswas, J.: Cluster Validation Using Graph Theoretic Concepts. Pattern Recognition 30, 6, 847-857 (1997)
11. Hathaway, R.J., Bezdek, J.C.: Visual Cluster Validity for Prototype Generator Clustering Models. Pattern Recognition Letters 24, 9-10, 1563-1569 (2003)
12. Ding, Y., Harrison, R.F.: Relational Visual Cluster Validity (RVCV). Pattern Recognition Letters 28, 15, 2071-2079 (2007)
13. Richardson, S., Green, P.: On Bayesian Analysis of Mixtures with Unknown Number of Components. J. Royal Statistical Soc. B 59, 731-792 (1997)
14. Rasmussen, C.: The Infinite Gaussian Mixture Model. In: Solla, S., Leen, T., Müller, K.-R. (eds.) Advances in Neural Information Processing Systems 12, pp. 554-560. MIT Press (2000)
15. Neal, R.M.: Markov Chain Sampling Methods for Dirichlet Process Mixture Models. J. Computational and Graphical Statistics 9, 2, 249-265 (2000)
16. Figueiredo, M.A.T., Jain, A.K.: Unsupervised Learning of Finite Mixture Models. IEEE Trans. Pattern Analysis and Machine Intelligence 24, 3, 381-396 (2002)
17. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automatic Control 19, 716-23 (1974)
18. Schwartz, G.: Estimating the Dimension of a Model. Annals of Statistics 6, 461-464 (1978)
19. Fraley, C., Raftery, A.: How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. Technical Report 329, Dept. Statistics, Univ. Washington, Seattle, WA (1998)
20. Bezdek, J.C., Li, W.Q., Attikiouzel, Y., Windham, M.: A Geometric Approach to Cluster Validity for Normal Mixtures. Soft Computing 1, 166-179 (1997)
21. Biernacki, C., Celeux, G., Govaert, G.: Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. IEEE Trans. Pattern Analysis Machine Intelligence 22, 7, 719-725 (2000)
22. Samé, A., Ambroise, C., Govaert, G.: An Online Classification EM Algorithm Based on the Mixture Model. Stat. Comput. 17, 209-218 (2007)
23. Comon, P.: Independent Component Analysis, a New Concept? Signal Processing 36, 3, 287-314 (1994)
24. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley, New York (1991)