entropy

MDPI

Article

An Approach for Determining the Number of Clusters in a Model-Based Cluster Analysis

Serkan Akogul [1,*] and Murat Erisoglu [2]

[1] Department of Statistics, Yildiz Technical University, 34220 Istanbul, Turkey
[2] Department of Statistics, Necmettin Erbakan University, 42090 Konya, Turkey; merisoglu@konya.edu.tr
* Correspondence: sakogul@yildiz.edu.tr; Tel.: +90-212-383-4438

Received: 12 July 2017; Accepted: 27 August 2017; Published: 29 August 2017

Abstract: To determine the number of clusters in the clustering analysis that has a broad range of applied sciences, such as physics, chemistry, biology, engineering, economics etc., many methods have been proposed in the literature. The aim of this paper is to determine the number of clusters of a dataset in a model-based clustering by using an Analytic Hierarchy Process (AHP). In this study, the AHP model has been created by using the information criteria Akaike's Information Criterion (AIC), Approximate Weight of Evidence (AWE), Bayesian Information Criterion (BIC), Classification Likelihood Criterion (CLC), and Kullback Information Criterion (KIC). The achievement of the proposed approach has been tested on common real and synthetic datasets. The proposed approach based on the corresponding information criteria has produced accurate results. The currently produced results have been seen to be more accurate than those corresponding to the information criteria.

Keywords: model-based clustering; cluster analysis; information criteria; analytic hierarchy process

1. Introduction

Many clustering algorithms have been encountered in the literature. The clustering algorithms can be categorized into centroid-based clustering, connectivity-based clustering, model-based clustering, and so on [1]. Since each one of the clustering algorithms is of great importance in its own application area, and thus model-based clustering has a very large application field, the present study focuses on the use of the model-based clustering one with the combination of the Analytic Hierarchy Process (AHP). Since the AHP is one of the most important multi-criteria decision making (MCDM) [2] and determining the number of clusters can be modeled as the MCDM problem [3,4], this work pays its attention to the consideration of the AHP in deciding the number of clusters of datasets in a combination way with model-based clustering.

Pearson [5] first introduced the idea of the mixture distribution model by studying a mixture of two univariate normal distributions with different means and variances. Later on, many related works [6–8] have been carried out. Model-based clustering based on a mixture of distributions has commonly been used in the clustering of datasets. Some of those who used the mixture of multivariate normal distributions in the cluster analysis are Wolfe [9,10], Day [11], and Binder [12]. To estimate parameters in the mixture distribution model, the Expectation-Maximization (EM) algorithm suggested by Dempster et al. [13] has been widely used [14,15].

Model-based clustering based on finite normal mixture models is the most commonly used approach [16–31]. In estimating the number of clusters in model-based clustering, the information criteria have widely been used [32–42]. Some of the common criteria in the literature are Akaike's Information Criterion (AIC) [32], Approximate Weight of Evidence (AWE) [37], Bayesian Information Criterion (BIC) [33], Classification Likelihood Criterion (CLC) [39], Kullback Information Criterion (KIC) [40], etc. These information criteria may give different results in estimating the number of

clusters of a dataset. For example, while the number of clusters in the Iris dataset [43] is 3, according to AIC, AWE, BIC, CLC, and KIC, the number of clusters in the set has been seen to be 4, 2, 2, 4, and 3, respectively (Table 7). The Iris is a multivariate dataset introduced by Ronald Fisher [43] and the dataset consists of 50 samples from each of three species: setosa, virginica, and versicolor. Each one of the criteria has been seen to possibly produce a different number of clusters for the same dataset. To overcome this problem, model-based clustering and the AHP have been combined. The AHP model has been created by using the information criteria AIC, AWE, BIC, CLC and KIC. For the first time, to the best of authors' knowledge, the number of clusters of a dataset in a model-based clustering has been determined by using the AHP. Thus, the common influence of those information criteria has been benefited. Satisfactory results have been obtained in terms of the suggested model.

The rest of the paper is organized as follows. Section 2 describes the model-based clustering, the AHP model, and the proposed approach. Section 3 presents details of the experimental study and analyses the results. Finally, Section 4 presents our conclusions and recommendation.

2. Materials and methods

2.1. The Model-Based Clustering

The model-based clustering assumes that a dataset to be clustered consists of various clusters with different distributions. The entire dataset is modeled by a mixture of these distributions. The clustering assumes a set of n p-dimensional vectors $y_1, \ldots, y_n$ of observations from a multivariate mixture of a finite number of g components or clusters each with some unknown mixing proportions or weights $\pi_1, \ldots, \pi_g$ [44]. The probability density function (PDF) of finite mixture distribution models can be given by

$$f\left(y_j; \Psi\right) = \sum_{i=1}^{g} \pi_i f_i\left(y_j; \theta_i\right) \tag{1}$$

where $f_i(y_j; \theta_i)$ is the PDF of the components. Here $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{g} \pi_i = 1$ ($i = 1, \ldots, g$ and $j = 1, 2, \ldots, n$). The parameter vector $\Psi = (\pi, \theta)$ contains all of the parameters of the mixture models. Here $\theta = (\theta_1, \theta_2, \ldots, \theta_g)$ denotes the unknown parameters of the PDF of the i-th components in the mixture models [17].

In the model-based clustering, the cluster analysis based on the mixture of multivariate normal distributions is the most commonly used. In this case, in Equation (1), $f_i(y_j; \theta_i)$'s are assumed to be multivariate normal distribution function of the form

$$f_i(y_j; \theta_i) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{\left\{-\frac{1}{2}(x_j - \mu_i)^T \Sigma_k^{-1} (x_j - \mu_i)\right\}} \tag{2}$$

where μ_i and Σ_i stand for the mean vector and the covariance matrix, respectively ($i = 1, 2, \ldots, g$). Here, θ_i stems from the mean compound vectors $\mu = (\mu_1, \mu_2, \ldots, \mu_g)$ and the compound covariance matrices $\Sigma = (\Sigma_1, \Sigma_2, \ldots, \Sigma_g)$ of the parameters of the compound PDF in the mixture distribution model [17].

The mixture likelihood approach has been used for estimating the parameters in the mixture models. This approach assumes that the PDF is the sum of the weighted component densities. If the mixture likelihood approach is used for clustering, the clustering problem comes out to be a problem of estimating the parameters of the assumed mixture distribution model. The maximum-likelihood function can then be given as follows [45],

$$L(\Psi) = \prod_{j=1}^{n} \sum_{i=1}^{g} \pi_i f_i\left(y_j \mid \theta_i\right). \tag{3}$$

The most widely used approach for parameter estimation is the EM algorithm [17].

Determination of the number of clusters is one of the most important problems in the cluster analysis. The information criteria for the number of clusters have often been used in model-based clustering. The criteria to be used in this study are given in Table 1.

Table 1. The information criteria for the model selection.

Criteria	Formula	Reference
AIC	$-2logL(\widehat{\Psi}) + 2d$	Akaike [32]
AWE	$-2logL_c + 2d(3/2 + logn)$	Banfierd and Raftery [37]
BIC	$-2logL(\widehat{\Psi}) + dlog(n)$	Schwarz [33]
CLC	$-2logL(\widehat{\Psi}) + 2EN(\widehat{\tau})$	Biernacki and Gnvaert [39]
KIC	$-2logL(\widehat{\Psi}) + 3(d+1)$	Cavanaugh [40]

n: The number of observations; d: The number of parameters in the model.

A model that gives the minimum of the values of the criteria AIC, AWE, BIC, CLC, and KIC are selected to be the best model. In Table 1, the log-likelihood function for the completed data is shown as $logL_c(\Psi) = logL(\Psi) + EN(\tau)$. Here, $EN(\tau) = -\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}log\tau_{ij}$ is the entropy of the related classification matrix [46].

2.2. The Analytic Hierarchy Process (AHP)

The AHP was developed by Saaty [2]. The AHP is one of the most widely used multiple criteria decision-making tools. The AHP is a method for structuring, measurement, and synthesis [47]. The AHP is a hierarchical structure consisting of goal, criteria, and alternatives [48]. The AHP chooses the best one among the alternatives, taking into account the goal and the criteria [49]. In addition, the AHP is a mathematical approach that evaluates qualitative and quantitative variables together. The literature [50] tells us that the AHP has been implemented in various fields of science such as selecting a best alternative, planning, optimization, etc. To solve decision-making problems using the AHP, the following steps are applied [48–52]:

Structuring: Initially goal, criteria, and alternatives are determined. Then, the hierarchy model is constructed at different levels according to the structure of the problem. A three level-hierarchy, that has k criteria and m alternatives, can be given in Figure 1.

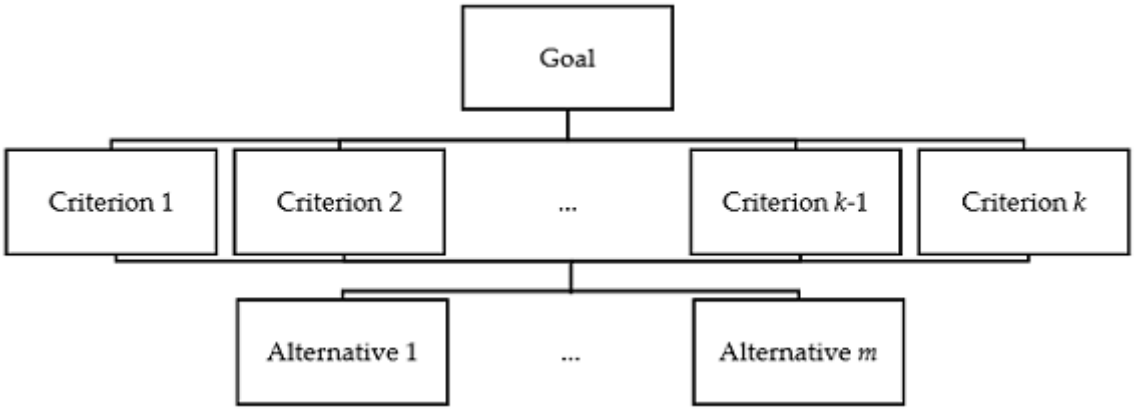

Figure 1. A three-level hierarchy.

Measurement: Firstly, a decision matrix is formed. The decision matrix involves the assessments of each alternative with respect to the decision criteria. The decision matrix has been given in Table 2. Here, element d_{ij} indicates the importance level of the i-th alternative with respect to the j-th criterion ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, k$).

Table 2. The decision matrix.

	Criterion 1	**Criterion 2**	…	**Criterion** k
Alternative 1	d_{11}	d_{12}	…	d_{1k}
Alternative 2	d_{21}	d_{22}	…	d_{2k}
⋮	⋮	⋮	⋮	⋮
Alternative m	d_{m1}	d_{m2}	…	d_{mk}

Secondly, the pairwise comparison matrices of the criteria and alternatives for each criterion have been produced. In general, the pairwise comparison matrix is constructed as in Table 3. Here, a_{ij} stands for the degree of preference of the i-th criterion/alternative over j-th criterion/alternative ($a_{ii} = 1$; $a_{ij} = 1/a_{ji}$), and Sum$t$ is summation of the t-th column the pairwise comparison matrix.

Table 3. The pairwise comparison matrix.

	X_1	X_2	…
X_1	a_{11}	a_{12}	…
X_2	a_{21}	a_{22}	…
…	…	…	…
Sum	Sum1	Sum2	…

X_t: t-th criterion ($t = 1, 2, \ldots, k$)/t-th alternative ($t = 1, 2, \ldots, m$).

Synthesis: To find the maximum eigenvalue (λ_{max}), consistency index (CI), consistency ratio (CR), and normalized eigenvector of each pairwise comparison matrix, the necessary calculations are performed. Note that CR = (CI/RI) is calculated for all pairwise comparison matrices. Here, CI = $(\lambda_{max} - r)/(r - 1)$ is the consistency index and RI is the random consistency index. As well-known from the literature [2,49,51,52], the average RI is calculated in terms of the dimension of the matrix, r. If the CR value is less than 0.10, it indicates that the matrices are consistent. As given in reference [49], if $\lambda_{max} = r$, then the pairwise comparison matrix is considered to be consistent.

After the consistency test, the following calculations are made. Firstly, the relative importance vector (RIV) of the criteria is determined using the pairwise comparison matrix. The row averages of the normalized matrix are represented by RIV = $[Avg1, Avg2, ...]^T$. To obtain the normalized matrix, the element of each column in the pairwise comparison matrix is divided by the column sum. The normalized matrix is then given in Table 4. The RIV of the alternatives for each criterion and the RIV of the criteria are calculated separately using the normalized matrices.

Table 4. The normalized matrix.

	X_1	X_2	…	**Average**
X_1	a_{11}/Sum1	a_{12}/Sum2	…	Avg1
X_2	a_{21}/Sum1	a_{22}/Sum2	…	Avg2
…	…	…	…	…

Finally, to calculate the composite relative importance vector (C-RIV), the matrix formed by the RIV of the alternatives for each criterion is multiplied by the RIV of the criteria. Thus, the C-RIV determines the overall ranking of the alternatives.

2.3. The Proposed Approach for the AHP Model and the Pairwise Comparison Matrix

The model-based clustering is currently a very popular statistical-model. The information criteria for determining the number of clusters in the model-based clustering have commonly been used [42].

A number of criteria have been proposed to determine the number of clusters in a dataset. The current study has proposed an approach for determining the number of clusters by using the combination of the model-based clustering with the AHP. The AHP model has been created by using the information criteria AIC, AWE, BIC, CLC, and KIC. In Figure 2, the proposed approach has been described for determining the number of clusters of a dataset in the model-based clustering.

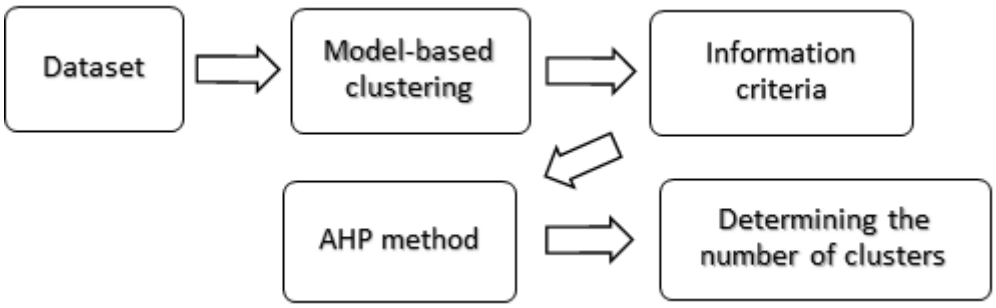

Figure 2. The proposed approach for determining the number of clusters.

The proposed approach is summarized in the following steps:

- **Step 1.** The hierarchical structure of the AHP has been created in Figure 3. In the figure, determination of the number of clusters is the goal, the AIC, AWE, BIC, CLC, KIC are the criteria, and 2, 3, 4, 5 are the alternatives.
- **Step 2.** The dataset has been modeled as the mixture of a multivariate normal distribution for the different number of clusters in the model-based clustering. The mean vectors, the covariance matrices, the mixture proportions, and the likelihood function have been estimated by the EM algorithm.
- **Step 3.** For each number of clusters, the values of the information criteria have been calculated. The decision matrix has been constructed using those values. Although a model that gives the lowest value of the information criteria in the model-based clustering is selected as the best model, in the AHP, the preferred case is the one with the highest value of the C-RIV. Therefore, the value of the information criteria has been reversed in the decision matrix; for example, the AIC is taken to be 1/AIC.
- **Step 4.** The pairwise comparison matrices have been obtained by using the decision matrix.
- **Step 5.** For each alternative, the C-RIV has been calculated.
- **Step 6.** The alternative having the highest C-RIV value is the optimal number of clusters for the dataset.

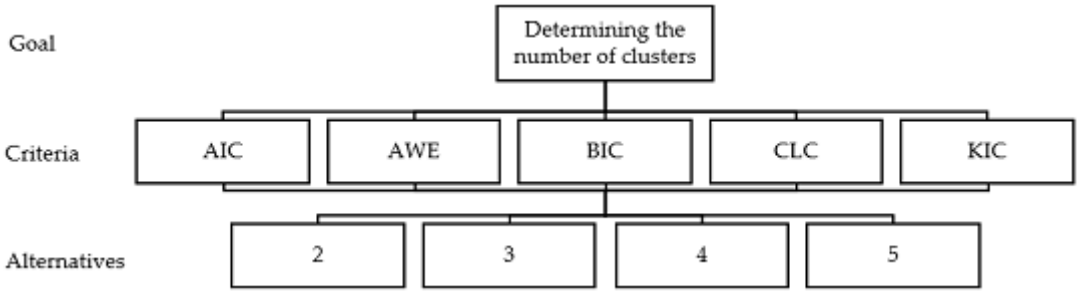

Figure 3. The hierarchical structure of the AHP for the proposed approach.

To form the pairwise comparison matrix of the criteria, the study of Akogul and Erisoglu [53] has been used. In their study [53], the efficiency of the information criteria was examined. They also

analyzed real datasets that are commonly used in clustering analysis. Those datasets have different characteristics such as sample size (i.e., 75, 150, 178, 345, 846, 2310 and 6435), number of clusters (i.e., 2, 3, 4, 6 and 7), and number of variables (i.e., 2, 4, 6, 13, 18, 19 and 36). The synthetic datasets were generated from the multivariate normal distribution by using the mean and covariance vectors of each dataset. Then, the number of clusters of the synthetic datasets were estimated by using the information criteria. This process was repeated 1000 times. Thus, the success of finding the right number of clusters in the dataset was computed for each information criterion. For all the synthetic datasets, in the corresponding study, the average of successes of the information criteria was given [53] as 43.6, 21.2, 47.4, 17.3, and 58.2 for AIC, AWE, BIC, CLC, and KIC, respectively.

In the work of Akogul and Erisoglu [53], the effectiveness of the information criteria was determined according to the success of finding right number of clusters. In the current study, those successes have been used to determine the importance level of the criteria in the AHP model. To produce the pairwise comparison matrix of the criteria, the average of the successes of the information criteria is considered. The average success is taken to be the degree of preference of a criterion over other criteria. The proposed pairwise comparison matrix of the criteria and the RIV of the criteria have been given in Table 5. For example, in Table 5, value 2.0566 can be interpreted as the degree of preference of the AIC over the AWE. The average of successes of the AIC is 43.6, while the AWE is of 21.2. Thus, the AIC is about two times more successful than the AWE.

Table 5. The proposed pairwise comparison matrix of the criteria and the RIV of the criteria.

Criteria	AIC	AWE	BIC	CLC	KIC	$RIV_{criteria}$ *
AIC	1	2.0566	0.9198	2.5202	0.7491	0.2323
AWE	0.4862	1	0.4473	1.2254	0.3643	0.1129
BIC	1.0872	2.2358	1	2.7399	0.8144	0.2525
CLC	0.3968	0.8160	0.3650	1	0.2973	0.0922
KIC	1.3349	2.7453	1.2278	3.3642	1	0.3101
Sum	4.3050	8.8538	3.9599	10.8497	3.2251	

* The relative importance vector (RIV) of the criteria.

3. Application and Results

3.1. Testing of the Proposed Approach for the Real Datasets

The achievement of the proposed approach has been tested on common real datasets, namely, Chemical Diabetes [54], Crab [55], Liver Disorders [56], Ionosphere [57], Iris [43], Wine [58], Ruspini [59], E.coli [60] and Vehicle Silhouettes [61]. They have been provided by the UCI machine learning repository [62] and the GitHub [63]. Their characteristics have been exhibited in Table 6.

Table 6. Descriptions of the real datasets.

Datasets	Sample Size	Number of Variables	Number of Clusters
Crab	200	5	2
Liver Disorders	345	6	2
Ionosphere	351	34	2
Chemical Diabetes	145	4	3
Iris	150	4	3
Wine	178	13	3
Ruspini	75	2	4
E.coli	336	8	4
Vehicle Silhouettes	846	18	4

In this section, to determine the number of clusters in the Iris dataset, all calculations have been produced step by step. For the other datasets, only the final results have been presented and the decision matrices have been represented in the Appendix. Their pairwise comparison matrices can be obtained by using the decision matrices.

The results of the information criteria in determining the number of clusters for the Iris dataset have been presented in Table 7. According to AIC, AWE, BIC, CLC, and KIC, the number of clusters in the Iris dataset has been estimated to be 4, 2, 2, 4, and 3, respectively.

Table 7. The results in determining the number of clusters for the Iris dataset.

Alternatives	AIC	AWE	BIC	CLC	KIC
2	487.11	806.74 *	574.42 *	429.12	519.11
3	449.15	944.15	581.61	371.21	496.15 *
4	448.86 *	1126.55	626.49	358.29 *	510.86
5	474.12	1378.81	696.90	415.24	551.12

* The minimum value of the information criteria.

To form the decision matrix, the values of the information criteria have been reversed (for example, AIC = 1/AIC). The decision matrix of the Iris dataset can be given in Table 8.

Table 8. The decision matrix for the Iris dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.2053	0.1240	0.1741	0.2330	0.1926
3	0.2226	0.1059	0.1719	0.2694	0.2016
4	0.2228	0.8877	0.1596	0.2791	0.1957
5	0.2109	0.7253	0.1435	0.2408	0.1814

The pairwise comparison matrix and the RIV of each criterion, which are obtained by using the decision matrix, have been seen in Table 9.

Table 9. The pairwise comparison matrix and the RIV of each criterion for the Iris dataset.

AIC	**2**	**3**	**4**	**5**	$\mathbf{RIV}_{AIC}$
2	1	0.9221	0.9215	0.9733	0.2383
3	1.0845	1	0.9994	1.0556	0.2584
4	1.0852	1.0006	1	1.0563	0.2586
5	1.0274	0.9473	0.9467	1	0.2448
Sum	4.1972	3.8700	3.8676	4.0852	
AWE	**2**	**3**	**4**	**5**	$\mathbf{RIV}_{AWE}$
2	1	1.1703	1.3964	1.7091	0.3169
3	0.8545	1	1.1932	1.4604	0.2708
4	0.7161	0.8381	1	1.2239	0.2269
5	0.5851	0.6848	0.8170	1	0.1854
Sum	3.1557	3.6932	4.4066	5.3934	
BIC	**2**	**3**	**4**	**5**	$\mathbf{RIV}_{BIC}$
2	1	1.0125	1.0906	1.2132	0.2682
3	0.9876	1	1.0772	1.1982	0.2649
4	0.9169	0.9284	1	1.1124	0.2459
5	0.8242	0.8346	0.8990	1	0.2211
Sum	3.7288	3.7755	4.0667	4.5239	

Table 9. *Cont.*

CLC	2	3	4	5	RIV_{CLC}
2	1	0.8651	0.8349	0.9676	0.2279
3	1.1560	1	0.9652	1.1186	0.2635
4	1.1977	1.0361	1	1.1589	0.2730
5	1.0334	0.8940	0.8629	1	0.2356
Sum	4.3871	3.7951	3.6630	4.2452	
KIC	**2**	**3**	**4**	**5**	RIV_{KIC}
2	1	0.9558	0.9841	1.0617	0.2497
3	1.0463	1	1.0297	1.1108	0.2613
4	1.0162	0.9712	1	1.0788	0.2538
5	0.9419	0.9003	0.9270	1	0.2352
Sum	4.0044	3.8272	3.9407	4.2512	

The C-RIV has been presented in Table 10. In the table, the alternative value three is the best alternative because it has the maximum value of 0.2628 for the C-RIV. Thus, the number of clusters for the Iris dataset has been seen to be determined correctly.

Table 10. The C-RIV for the Iris dataset.

Alternatives	RIV_{AIC}	RIV_{AWE}	RIV_{BIC}	RIV_{CLC}	RIV_{KIC}	C-RIV
2	0.2383	0.3169	0.2682	0.2279	0.2497	0.2573
3	0.2584	0.2708	0.2649	0.2635	0.2613	0.2628 *
4	0.2586	0.2269	0.2459	0.2730	0.2538	0.2516
5	0.2448	0.1854	0.2211	0.2356	0.2352	0.2283
$RIV_{criteria}$	0.2323	0.1129	0.2525	0.0922	0.3101	

* The maximum value of the C-RIV.

The C-RIV for the real datasets has also been presented in Table 11. In the table, the number of clusters for the real datasets has been estimated correctly by using the proposed approach.

Table 11. The C-RIV of the real datasets for the proposed approach.

	C-RIV								
	Crab	**Liver**	**Ionosphere**	**Diabetes**	**Iris**	**Wine**	**Ruspini**	**E.coli**	**Vehicle**
2	0.2517 *	0.2507 *	0.2841 *	0.2490	0.2573	0.2476	0.2436	0.0709	0.2451
3	0.2491	0.2502	0.2449	0.2513 *	0.2628 *	0.2578 *	0.2486	0.2989	0.2513
4	0.2481	0.2504	0.2559	0.2502	0.2516	0.2524	0.2541 *	0.3325 *	0.2534 *
5	0.2511	0.2487	0.2151	0.2496	0.2283	0.2421	0.2537	0.2977	0.2502

* The maximum value of the C-RIV for each dataset.

3.2. Testing of the Proposed Approach for the Synthetic Datasets

For the synthetic-1 dataset, we generate 1000 samples from a two-component bivariate normal mixture with the mixing proportions $\pi_1 = \pi_2 = 1/2$, the mean vectors $\mu_1 = [2, 4]^T$, $\mu_2 = [5, 6]^T$, and the covariance matrices $\Sigma_1 = [1, 0; 0, 1]$, $\Sigma_2 = [2, 0; 0, 0.5]$. Figure 4 shows the scatter plot and the PDF of the mixture model of the synthetic-1 dataset. The decision matrix of the synthetic-1 dataset has been presented in Table 12.

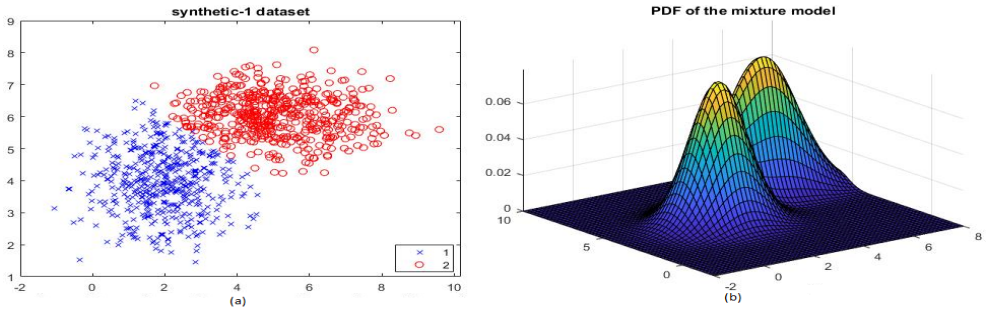

Figure 4. Synthetic-1 dataset: (**a**) the scatter plot; (**b**) the PDF of the mixture model.

Table 12. The decision matrix for the synthetic-1 dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.1460	0.1373	0.1449	0.1408	0.1457
3	0.1461	0.1254	0.1443	0.1301	0.1456
4	0.1457	0.1188	0.1434	0.1245	0.1452
5	0.1458	0.1118	0.1428	0.1182	0.1451

The RIV of the alternatives for each criterion and the RIV of the criteria have been given in Table 13. To identify the best alternative, the C-RIV has been calculated using the corresponding values. In the table, the alternative value two is the best alternative because it has the maximum value, 0.2561, for the C-RIV. That is, the number of clusters has been seen to be determined correctly for the synthetic-1 dataset.

Moreover, this operation has been repeated 1000 times. The success of finding right number of clusters in the synthetic-1 dataset has been computed for each information criterion. The proposed approach (100%) has been seen to be more accurate than the AIC (92%), the AWE (27%), the BIC (99%), the CLC (93%), and the KIC (98%).

Table 13. The C-RIV for the synthetic-1 dataset.

Alternatives	$\mathbf{RIV}_{AIC}$	$\mathbf{RIV}_{AWE}$	$\mathbf{RIV}_{BIC}$	$\mathbf{RIV}_{CLC}$	$\mathbf{RIV}_{KIC}$	C-RIV
2	0.2502	0.2782	0.2518	0.2741	0.2505	0.2561 *
3	0.2503	0.2543	0.2508	0.2533	0.2504	0.2512
4	0.2497	0.2409	0.2492	0.2424	0.2496	0.2479
5	0.2498	0.2266	0.2482	0.2302	0.2495	0.2449
$\mathbf{RIV}_{criteria}$	0.2323	0.1129	0.2525	0.0922	0.3101	

* The maximum value of the C-RIV.

For the synthetic-2 dataset, we generate 1000 samples from a three-component bivariate normal mixture with the mixing proportions $\pi_1 = \pi_2 = \pi_3 = 1/3$, the mean vectors $\mu_1 = [-1, 2]^T$, $\mu_2 = [1, 1]^T$, $\mu_3 = [0, -4]^T$, and the covariance matrices $\Sigma_1 = [1, 0; 0, 1]$, $\Sigma_2 = [0.5, -0.7; -0.7, 1.5]$, and $\Sigma_3 = [2, 0; 0, 2]$. Figure 5 shows the scatter plot and the PDF of the mixture model of the synthetic-2 dataset.

The decision matrix and the C-RIV of the synthetic-2 dataset have been given in Tables 14 and 15. In Table 15, the alternative value three is the best alternative. Namely, the number of clusters for the synthetic-2 dataset has been determined correctly. Similar to previous calculations, this operation has been repeated 1000 times. The success of finding right number of clusters in the synthetic-2 dataset has been computed for each information criterion. The proposed approach (93%) has been seen to be better than the AIC (74%), the AWE (10%), the BIC (92%), the CLC (31%), and the KIC (86%).

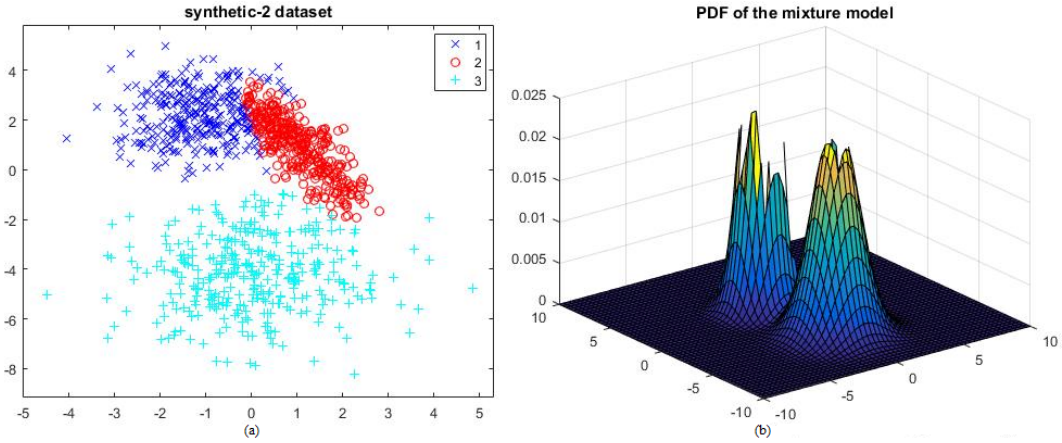

Figure 5. Synthetic-2 dataset: (**a**) the scatter plot; (**b**) the PDF of the mixture model.

Table 14. The decision matrix for the synthetic-2 dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.1275	0.1234	0.1266	0.1263	0.1273
3	0.1316	0.1226	0.1302	0.1270	0.1313
4	0.1316	0.1154	0.1296	0.1208	0.1311
5	0.1319	0.1170	0.1295	0.1241	0.1313

Table 15. The C-RIV for the synthetic-2 dataset.

Alternatives	$\mathbf{RIV}_{AIC}$	$\mathbf{RIV}_{AWE}$	$\mathbf{RIV}_{BIC}$	$\mathbf{RIV}_{CLC}$	$\mathbf{RIV}_{KIC}$	C-RIV
2	0.2440	0.2579	0.2455	0.2535	0.2443	0.2469
3	0.2519	0.2562	0.2524	0.2550	0.2520	0.2528 *
4	0.2518	0.2412	0.2513	0.2424	0.2517	0.2496
5	0.2524	0.2446	0.2509	0.2491	0.2521	0.2507
$\mathbf{RIV}_{criteria}$	0.2323	0.1129	0.2525	0.0922	0.3101	

* The maximum value of the C-RIV.

For the synthetic-3 dataset, we generate again 1000 samples from a four-component bivariate normal mixture with the mixing proportions $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$, the mean vectors $\mu_1 = \mu_2 = [-2, -2]^T$, $\mu_3 = [3, 1]^T$, $\mu_4 = [1, -3]^T$, and the covariance matrices $\Sigma_1 = [0.2, 0; 0, 0.2]$, $\Sigma_2 = [3, 2; 2, 7]$, $\Sigma_3 = [1, 0; 0, 4]$, and $\Sigma_4 = [1, 0; 0, 1]$. Figure 6 shows the scatter plot and the PDF of the mixture model of the synthetic-3 dataset.

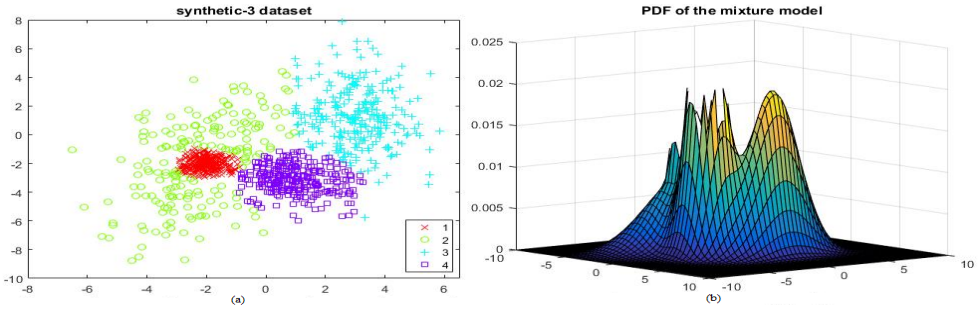

Figure 6. Synthetic-3 dataset: (**a**) the scatter plot; (**b**) the PDF of the mixture model.

The decision matrix and the C-RIV of the synthetic-3 dataset have been seen in Tables 16 and 17. In Table 17, the alternative value four is the best alternative because it has the maximum value, 0.2526,

of the C-RIV. The number of clusters for the synthetic-3 dataset has been determined correctly. Similarly, this operation has been repeated 1000 times. The success of finding the right number of clusters in the synthetic-3 dataset has been computed for each information criterion. The proposed approach (92%) has been seen to be better than the AIC (80%), the AWE (4%), the BIC (89%), the CLC (65%), and the KIC (89%).

Table 16. The decision matrix for the synthetic-3 dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.1204	0.1146	0.1196	0.1171	0.1202
3	0.1223	0.1106	0.1211	0.1142	0.1220
4	0.1244	0.1114	0.1227	0.1164	0.1240
5	0.1245	0.1080	0.1223	0.1140	0.1240

Table 17. The C-RIV for the synthetic-3 dataset.

Alternatives	$\mathbf{RIV}_{AIC}$	$\mathbf{RIV}_{AWE}$	$\mathbf{RIV}_{BIC}$	$\mathbf{RIV}_{CLC}$	$\mathbf{RIV}_{KIC}$	C-RIV
2	0.2449	0.2578	0.2463	0.2536	0.2452	0.2476
3	0.2488	0.2487	0.2493	0.2473	0.2489	0.2488
4	0.2531	0.2506	0.2527	0.2522	0.2530	0.2526 *
5	0.2532	0.2429	0.2518	0.2469	0.2529	0.2510
$\mathbf{RIV}_{criteria}$	0.2323	0.1129	0.2525	0.0922	0.3101	

* The maximum value of the C-RIV.

Table 18 summarizes the estimations of the number of clusters for all datasets produced by the information criteria and the proposed approach. The bottommost column has given the correct number of clusters for each dataset determined by the information criteria and the proposed approach.

Table 18. Results summary.

Datasets	#Cluster	AIC	AWE	BIC	CLC	KIC	Proposed Approach
Crab	2	5	2	2	5	5	2
Liver	2	4	2	2	5	4	2
Ionosphere	2	4	2	2	4	2	2
Diabetes	3	5	2	3	5	3	3
Iris	3	4	2	2	4	3	3
Wine	3	4	2	3	5	3	3
Ruspini	4	5	2	4	5	5	4
E.coli	4	4	3	3	5	4	4
Vehicle	4	4	2	4	4	4	4
Synthetic-1	2	3	2	2	2	2	2
Synthetic-2	3	5	2	3	3	5	3
Synthetic-3	4	5	2	4	2	4	4
Correct number		2	4	10	3	8	12

4. Conclusions and Recommendation

This paper has proposed to combine the AHP and some information criteria, namely AIC, AWE, BIC, CLC, and KIC, in determining the number of clusters of a dataset in model-based clustering. It has been concluded that the proposed approach has been seen to be more accurate than the corresponding information criteria. The approach has thus been realized to be capable of application to a widespread number of clustering algorithms. To carry out this study, the decision matrix has been created by using the information criteria values for each case. To increase the successes of the information criteria, a pairwise comparison matrix has been suggested in this study. Note that the proposed method is

strongly expected to be very effective in analyzing data come out in various of fields science such as economics, biology, engineering etc. For further studies, researchers can pay their attention to produce different decision and pairwise comparison matrices to deal with their problems.

Acknowledgments: This research has been supported by TUBITAK-BIDEB (2211) Ph.D. scholarship program. The author is grateful to anonymous referees for their constructive comments and valuable suggestions to improve this paper. The authors wish to thank Murat SARI (Yildiz Technical University, Istanbul) for reading the manuscript and providing many useful suggestions.

Author Contributions: Serkan Akogul and Murat Erisoglu conceived of the research and wrote the paper. Both authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The decision matrices of the real datasets are given in Tables A1–A8. Their pairwise comparison matrices can easily be obtained by using the decision matrices.

Table A1. The decision matrix for the Crab dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.3414	0.2878	0.3263	0.3428	0.3363
3	0.3424	0.2710	0.3200	0.3512	0.3349
4	0.3463	0.2554	0.3163	0.3588	0.3362
5	0.3550	0.2459	0.3165	0.3771	0.3420

Table A2. The decision matrix for the Liver dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.0678	0.0645	0.0668	0.0680	0.0675
3	0.0681	0.0630	0.0666	0.0683	0.0677
4	0.0685	0.0618	0.0665	0.0687	0.0679
5	0.0683	0.0603	0.0659	0.0688	0.0676

Table A3. The decision matrix for the Ionosphere dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.0816	0.0354	0.0584	0.1026	0.0739
3	0.0741	0.0267	0.0481	0.1030	0.0650
4	0.0829	0.0227	0.0459	0.1423	0.0686
5	0.0702	0.0184	0.0379	0.1260	0.0575

Table A4. The decision matrix for the Diabetes dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.1579	0.1501	0.1558	0.1591	0.1571
3	0.1602	0.1483	0.1569	0.1620	0.1590
4	0.1604	0.1444	0.1560	0.1623	0.1588
5	0.1607	0.1415	0.1552	0.1638	0.1587

Table A5. The decision matrix for the Wine dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.1915	0.1316	0.1699	0.2081	0.1840
3	0.2082	0.1193	0.1724	0.2389	0.1953
4	0.2104	0.1051	0.1643	0.2552	0.1932
5	0.2066	0.0926	0.1536	0.2635	0.1863

Table A6. The decision matrix for the Ruspini dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.7096	0.6600	0.6970	0.7209	0.7026
3	0.7300	0.6519	0.7096	0.7484	0.7195
4	0.7519	0.6442	0.7230	0.7784	0.7375
5	0.7562	0.6242	0.7196	0.7907	0.7383

Table A7. The decision matrix for the E.coli dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.2961	0.2335	0.2741	0.3082	0.2897
3	1.7566	0.5182	1.0228	2.7483	1.4721
4	2.1498	0.4388	0.9891	5.3666	1.6362
5	1.9438	0.3589	0.8349	6.0067	1.4359

Table A8. The decision matrix for the Vehicle dataset ($\times 10^{-3}$).

Alternatives	AIC	AWE	BIC	CLC	KIC
2	0.0124	0.0116	0.0121	0.0125	0.0123
3	0.0128	0.0116	0.0124	0.0130	0.0127
4	0.0130	0.0114	0.0124	0.0132	0.0129
5	0.0129	0.0110	0.0122	0.0132	0.0127

References

1. Yu, H.; Liu, Z.; Wang, G. An automatic method to determine the number of clusters using decision-theoretic rough set. *Int. J. Approx. Reason.* **2014**, *55*, 101–115.
2. Saaty, T.L. The Analytic Hierarchy Process. In *Cook WD and Seiford LM (1978). Priority Ranking and Consensus Formation, Management Science*; McGraw-Hill: New York, NY, USA, 1980; pp. 1721–1732.
3. Peng, Y.; Kou, G.; Wang, G.; Wu, W.; Shi, Y. Ensemble of software defect predictors: An AHP-based evaluation method. *Int. J. Inf. Technol. Decis. Mak.* **2011**, *10*, 187–206.
4. Peng, Y.; Zhang, Y.; Kou, G.; Shi, Y. A multicriteria decision making approach for estimating the number of clusters in a data set. *PLoS ONE* **2012**, *7*, e41713.
5. Pearson, K. Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. Lond. A* **1894**, *185*, 71–110.
6. Rao, C.R. The utilization of multiple measurements in problems of biological classification. *J. R. Stat. Soc. Ser. B* **1948**, *10*, 159–203.
7. Hasselblad, V. Estimation of parameters for a mixture of normal distributions. *Technometrics* **1966**, *8*, 431–444.
8. Hasselblad, V. Estimation of finite mixtures of distributions from the exponential family. *J. Am. Stat. Assoc.* **1969**, *64*, 1459–1471.
9. Wolfe, J.H. *A Computer Program for the Maximum Likelihood Analysis of Types*; Technical Report; Naval Personnel Research Activity: San Diego, CA, USA, 1965.
10. Wolfe, J.H. *NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions*; Technical Report; Naval Personnel Research Activity: San Diego, CA, USA, 1967.
11. Day, N.E. Estimating the components of a mixture of normal distributions. *Biometrika* **1969**, *56*, 463–474.

12. Binder, D.A. Bayesian cluster analysis. *Biometrika* **1978**, *65*, 31–38.
13. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38.
14. Everitt, B. Maximum likelihood estimation of the parameters in a mixture of two univariate normal distributions; a comparison of different algorithms. *Statistician* **1984**, *33*, 205–215.
15. McLachlan, G.J.; Peel, D.; Basford, K.E.; Adams, P. The EMMIX software for the fitting of mixtures of normal and t-components. *J. Stat. Softw.* **1999**, *4*, 1–14.
16. Celeux, G.; Govaert, G. Gaussian parsimonious clustering models. *Pat. Recognit.* **1995**, *28*, 781–793.
17. McLachlan, G.; Peel, D. *Finite Mixture Models*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
18. Yeung, K.Y.; Fraley, C.; Murua, A.; Raftery, A.E.; Ruzzo, W.L. Model-based clustering and data transformations for gene expression data. *Bioinformatics* **2001**, *17*, 977–987.
19. Meilă, M.; Heckerman, D. An experimental comparison of model-based clustering methods. *Mach. Learn.* **2001**, *42*, 9–29.
20. Fraley, C.; Raftery, A.E. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **2002**, *97*, 611–631.
21. Biernacki, C.; Celeux, G.; Govaert, G. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Stat. Data Anal.* **2003**, *41*, 561–575.
22. Pernkopf, F.; Bouchaffra, D. Genetic-based EM algorithm for learning Gaussian mixture models. *IEEE Trans. Pat. Anal. Mach. Intell.* **2005**, *27*, 1344–1348.
23. Raftery, A.E.; Dean, N. Variable selection for model-based clustering. *J. Am. Stat. Assoc.* **2006**, *101*, 168–178.
24. Jain, A.K. Data clustering: 50 years beyond K-means. *Pat. Recognit. Lett.* **2010**, *31*, 651–666.
25. Browne, R.P.; McNicholas, P.D.; Sparling, M.D. Model-based learning using a mixture of mixtures of Gaussian and uniform distributions. *IEEE Trans. Pat. Anal. Mach. Intell.* **2012**, *34*, 814–817.
26. Yang, M.S.; Lai, C.Y.; Lin, C.Y. A robust EM clustering algorithm for Gaussian mixture models. *Pat. Recognit.* **2012**, *45*, 3950–3961.
27. Lee, S.X.; McLachlan, G.J. Model-based clustering and classification with non-normal mixture distributions. *Stat. Methods Appl.* **2013**, *22*, 427–454.
28. Bouveyron, C.; Brunet-Saumard, C. Model-based clustering of high-dimensional data: A review. *Comput. Stat. Data Anal.* **2014**, *71*, 52–78.
29. Kwedlo, W. A new random approach for initialization of the multiple restart EM algorithm for Gaussian model-based clustering. *Pat. Anal. Appl.* **2015**, *18*, 757–770.
30. Malsiner-Walli, G.; Frühwirth-Schnatter, S.; Grün, B. Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **2016**, *26*, 303–324.
31. Marbac, M.; Biernacki, C.; Vandewalle, V. Model-based clustering of Gaussian copulas for mixed data. *Commun. Stat. Theory Methods* **2017**, doi:10.1080/03610926.2016.1277753.
32. Akaike, H. Information theory and an extension of the Maximum likelihood Principal. In Proceedings of the 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, 2–8 September 1971.
33. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
34. Bozdogan, H. On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Commun. Stat. Theory Methods* **1990**, *19*, 221–278.
35. Bozdogan, H. Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In *Information and Classification*; Springer: Berlin, Germany, 1993; pp. 40–54.
36. Bozdogan, H. Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*; Springer: Berlin, Germany, 1994; pp. 69–113.
37. Banfield, J.D.; Raftery, A.E. Model-based Gaussian and non-Gaussian clustering. *Biometrics* **1993**, *49*, 803–821.
38. Celeux, G.; Soromenho, G. An entropy criterion for assessing the number of clusters in a mixture model. *J. Classif.* **1996**, *13*, 195–212.
39. Biernacki, C.; Govaert, G. Using the classification likelihood to choose the number of clusters. *Comput. Sci. Stat.* **1997**, *29*, 451–457.
40. Cavanaugh, J.E. A large-sample model selection criterion based on Kullback's symmetric divergence. *Stat. Probab. Lett.* **1999**, *42*, 333–343.

41. Smyth, P. Model selection for probabilistic clustering using cross-validated likelihood. *Stat. Comput.* **2000**, *10*, 63–72.
42. Oliveira-Brochado, A.; Martins, F.V. *Assessing the Number of Components in Mixture Models: A Review*; Technical Report; Universidade do Porto, Faculdade de Economia do Porto: Porto, Portugal, 2005.
43. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Hum. Genet.* **1936**, *7*, 179–188.
44. Erol, H. A model selection algorithm for mixture model clustering of heterogeneous multivariate data. In Proceedings of the 2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA), Albena, Bulgaria, 19–21 June 2013; pp. 1–7.
45. Fraley, C. Algorithms for model-based Gaussian hierarchical clustering. *SIAM J. Sci. Comput.* **1998**, *20*, 270–281.
46. Hathaway, R.J. Another interpretation of the EM algorithm for mixture distributions. *Stat. Probab. Lett.* **1986**, *4*, 53–56.
47. Forman, E.H.; Gass, S.I. The analytic hierarchy process—An exposition. *Oper. Res.* **2001**, *49*, 469–486.
48. Saaty, R.W. The analytic hierarchy process—What it is and how it is used. *Math. Model.* **1987**, *9*, 161–176.
49. Saaty, T.L.; Vargas, L.G. *Models, Methods, Concepts & Applications of the Analytic Hierarchy Process*; Springer Science & Business Media: Berlin, Germany, 2012; Volume 175.
50. Vaidya, O.S.; Kumar, S. Analytic hierarchy process: An overview of applications. *Eur. J. Oper. Res.* **2006**, *169*, 1–29.
51. Saaty, T.L. How to make a decision: The analytic hierarchy process. *Eur. J. Oper. Res.* **1990**, *48*, 9–26.
52. Muralidhar, K.; Santhanam, R.; Wilson, R.L. Using the analytic hierarchy process for information system project selection. *Inf. Manag.* **1990**, *18*, 87–95.
53. Akogul, S.; Erisoglu, M. A Comparison of Information Criteria in Clustering Based on Mixture of Multivariate Normal Distributions. *Math. Comput. Appl.* **2016**, *21*, 34.
54. Reaven, G.; Miller, R. An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia* **1979**, *16*, 17–24.
55. Campbell, N.; Mahon, R. A multivariate study of variation in two species of rock crab of the genus Leptograpsus. *Aust. J. Zool.* **1974**, *22*, 417–425.
56. Forsyth, R. *PC/Beagle User's Guide*; BUPA Medical Research Ltd.: Hong Kong, China, 1990.
57. Sigillito, V.G.; Wing, S.P.; Hutton, L.V.; Baker, K.B. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Tech. Dig.* **1989**, *10*, 262–266.
58. Aeberhard, S.; Coomans, D.; De Vel, O. *Comparison of Classifiers in High Dimensional Settings*; Technical Report 92-02; Department of Computer Science and Department of Mathematics and Statistics, James Cook University of North Queensland: Townsville City, Australia, 1992.
59. Ruspini, E.H. Numerical methods for fuzzy clustering. *Inf. Sci.* **1970**, *2*, 319–350.
60. Horton, P.; Nakai, K. A probabilistic classification system for predicting the cellular localization sites of proteins. In Proceedings of the International Conference on Intelligent Systems for Molecular Biology, St. Louis, MO, USA, 12–15 June 1996; pp. 109–115.
61. Siebert, J.P. *Vehicle Recognition Using Rule Based Methods*; Turing Institute: London, UK, 1987.
62. Lichman, M. UCI Machine Learning Repository. Available online: http://archive.ics.uci.edu/ml (accessed on 17 June 2013).
63. GitHub. Available online: http://vincentarelbundock.github.io/Rdatasets/datasets.html/ (accessed on 17 June 2017).