Parallel Evaluation of Large Scale Hierarchical Clustering Results

David Cruz-Rodriguez
Computer Engineering
Dr. Luis M. Vicente, Ph.D.
Electrical & Computer Engineering
Polytechnic University of Puerto Rico

Abstract *— Data clustering refers to the automatic grouping of object based on their similarity, i.e., similar objects should be in the same group and dissimilar objects should be in different groups. In particular, for hierarchical clustering algorithms there is also the notion of a hierarchy in which the objects and the cluster fit. Clustering is a fundamental task in data mining, machine learning, information retrieval, bioinformatics, and image analysis, among others. It is important to evaluate the result of clustering algorithms. However most evaluations approaches are geared towards non-hierarchical clustering approaches; this research explores how to use traditional validity measures to evaluate and assess hierarchical clustering results.*

Key Terms *— Clustering, Data Clustering, Hierarchical Clustering, and Validity Measures.*

DEVELOPMENT OF THE PROBLEM

Clustering is a fundamental task in data mining, machine learning, information retrieval, bioinformatics, and image analysis, among others. The challenge to evaluate the goodness of the clustering validity is equally important as the result obtained by the generated algorithm. The most common approach to solve this challenge is using non-hierarchical evaluations. This research explores a different approach; using traditional non-hierarchical validity indexes to assess the hierarchical approach.

Background

Cluster is the organization of similar objects that are inter-similar to each other and intra-similar to different clusters.

On other words, points within a cluster are more similar to each other than points belonging to other clusters. An example of clustering is shown in next Figure 1. Here, the points belonging to the same color are given the same label, meaning that they are in the same cluster.

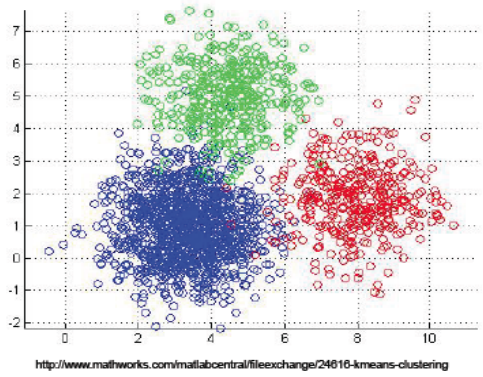

Figure 1
K-means Clustering

Since, there is a huge amount of clustering approaches to solve or model the same problem. Imagine that you are given a large data-set, a challenge that you may confront is to choose the perfect clustering model that could be more adequate to better describe the data-set. Another important question that may arise is how to confirm the strong fit of the chosen clustering model. They are known as cluster validity methods.

Huge amounts of validity indexes have been proposed in literature to address the problem of validity fit. For example, several of the approaches are based on counting pairs [1], [2], [3], [4], [5]. The most important criterion dealing with counting pairs of points, when comparing the clustering, is determining if the points on which the two clustering's agree or disagree [2]. Others examples are based on entropy and purity, calculating the quality of a set of clusters among many others [1].

Significance of the Study

This research is to surveys the core concepts and techniques in the large variety of cluster

validation for non-hierarchical to be applied into hierarchical evaluation.

General speaking, there are three types of clustering validation techniques: external, internal and relative criteria [6], [7]. On this research the attention should be focused on the external criteria.

REVIEW OF RELEVANT LITERATURE

In this section, there is an overview of the clustering evaluation and validation indexes.

Flat vs. Hierarchical Clustering

The purpose of the flat clustering is to set clusters without any explicit structure between the labels of each cluster. Again, a good example is the k-mean algorithm result from the Figure 1. As is illustrated in the depicted figure: the data is cluster and labeled with a color to symbolize the cluster. From the exploration standpoint is easily clear how similar are the data between one to the other. The algorithm is quite simple and effective, it minimize the sum of squares and the corresponding cluster centroid. Perhaps, that is why the k-means is a well-known flat clustering algorithm between academics. Also, the expectation maximation is one of the most popular ones.

Between the flat clustering you can make a second distinction, described below.

- **Hard Clustering:** Based on each label belong to one cluster or they don't.
- **Soft Clustering:** Based on each label is distributed over all clusters.

The k-means [8] is a non-hierarchical clustering and the most important flat clustering algorithm. According to [8], this application is not to find some unique, definitive grouping, but rather to simply aid the investigator in obtaining qualitative and quantitative understanding of large amounts of N-dimensional data by providing him good similarity groups.

By looking a hierarchical clustering (see Figure 2) output it is cleared that is more informative than the unstructured set of clusters reviewed in Flat Clustering.

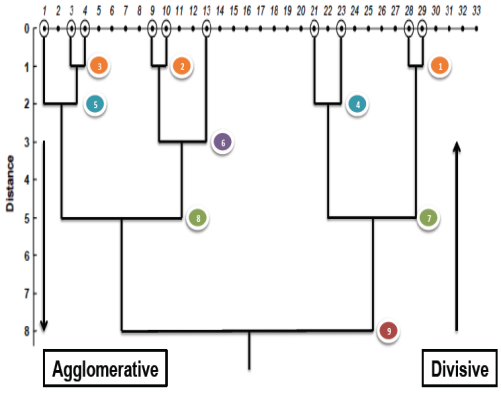

Figure 2
Hierarchical Clustering

Imagine that we have a dispersion of 10 points in 2-D and by applying the hierarchical algorithm obtained the output depicted on Figure 2. By quick inspection it can be determined a quick similarity. That each level of the hierarchy has at least a couple of merging, forming a cluster. Except the label number 6, identified as the color violet, that only has one merging. Additionally, the cutting point will give the numbers of clusters in each level of the hierarchy and the vertical axis will provide the similarity measure between clusters. In other words, the hierarchical method is more informative than the unstructured representation of the flat clustering. Also, the hierarchical clustering is very suitable because it does not require the number of clusters, as expected in flat clustering.

Hierarchical clustering can be general grouped in to two main categories, described below.

- **Agglomerative:** Based on the "bottom up" approach. By looking Figure 2., the agglomerative approach starts in each number as a cluster (1, 3, 4, 9, 10, 13, 21, 23, 28, and 29). Then merges into a cluster, depending on the selected parameter, until one cluster is left.
- **Divisive:** Based on the "top down" approach. Using the same example, but this time starting with all the numbers in on cluster, and splits the cluster until a single number cluster are left.

The depiction of a hierarchical clustering is called a dendogram. The dendogram is the hierarchy tree representation that shows the structure of the clusters as depicted in Figure 2.

Cluster Validity Methods

Generally speaking, there are three types of clustering validation techniques external, internal, and relative criteria [6], [7].

- **External Criteria:** Based on the priori knowledge about the data. This means that we evaluate the data based on previous set of clusters or results of a clustering algorithm. Is usually referred as "partition" (P).
- **Internal Criteria:** Based on the vectors of the data set alone. This case is very different from the external criteria, because the clustering results are evaluated from the data clustering results themselves.
- **Relative Criteria:** Based on the evaluation of clustering structuring by comparing it to other clustering schemes, a comparison with the different algorithm but same data inputs.

Therefore, if only the attention is given to the external validate-cluster data-sets, is important to keep in mind that the external criteria is required to possess a prior knowledge of the data-sets. For example, lets imagine that you possess a supervised learning and unsupervised learning results. In order to use the external indexes you need the prior knowledge of the data sets that is recognized as the supervised learning. This is the data that a human impose. Therefore, the external criteria compares between prior information with the generated by the clustering results. Also, important to keep in mind that in the real world usually there is no prior information of the data sets.

On previous works there are various measures which are to measure the strong fit of the data set produced by clustering algorithm [1], [6], [7], [9]. The first thing to start with is to review the well-known external indexes: the Rand index [4], Adjusted Rand index [5], Jaccard index [1], and Folkes & Mallows index [3], which are based on counting the pair of points on which two points agree or disagree. The Entropy index measures the quality of the cluster in each single class labels, which according to [1], entropy technique have also been defined as "variation of information" [2], among many others. The Purity index measures the frequency of the most common labels into each cluster.

VALIDATION INDEXES

In this section, there is an overview of validation indexes for external criterion. Refer to the Table 1 below; to see the notation meaning of each individual validity indices.

Table 1
Notation in Validity Indices

Notation	Meaning
M	Maximum Number of Pairs
N	Total Number of Points
n_{ij}	Number of elements in i^{th} partitions j^{th} clusters
n_i	Abstract and Key Terms
n_j	Body Text
E_j	Section Headings
p_{ij}	Section Sub-headings
k	Endnote
P_j	Equations

Metric Based on Counting Pairs

One of the approaches to evaluate metrics for clustering is considering statistics over pairs of items [2], [6], [7]. The most important concept criterion dealing with counting pairs of points, when comparing the clustering, is determining the points on which the two clustering agree or disagree.

For example, imagine that a given set of n objects $D = \{O_1, \dots, O_n\}$, suppose that $C = \{C_1, \dots, C_k\}$ (C is our clustering result) and $P = \{P_1, \dots, P_{k'}\}$ (P is our external criterion or partition). Both represent two different partitions of the objects D (D is our data). Having this in mind, we can create the contingency table or confusion matrix [10]; and [2] between our partition and the cluster (see Figure 3).

Partition/Cluster	P_1	P_2	…	$P_{k'}$	Sums
C_1	n_{11}	n_{12}	…	n_{1P}	$n_{1.}$
C_2	n_{21}	n_{22}	…	n_{1P}	$n_{2.}$
⋮	⋮	⋮		⋮	⋮
C_k	n_{k1}	n_{k2}	…	$n_{kk'}$	$n_{k.}$
Sums	$n_{.1}$	$n_{.2}$	…	$n_{.P}$	$n_{..} = n$

Figure 3
Confusion Matrix

The counting pairs matrix is the overlapping between the pair of points that can only fall in less

than one of four cases described below (see Figure 4): SS is the number of pairs of items belonging to the same cluster and partition; SD is the number of pairs belonging to the same cluster and different partition; DS is the number of pairs belonging to different cluster and the same partition; DD is the number of pairs belonging to different cluster and partition.

	P	P'
C	SS	SD
C'	DS	DD

Figure 4

Counting Pairs Matrix

The four counts always satisfy $M = DD + SD + DS + SS = n(n-1)/2$ (meaning M is the maximum number of all pairs and where n is the total number of points between C and P). The quantities between SS & DD can be interpreted as agreement "good choices" and SD & DS as disagreements "bad choices".

Some of the measures to define similarity between counting pairs such as:

- **Rand Index:** The Rand index (see Equation (1)) or Rand measure (Equation (2)) [4] in statistics, and in particular in data clustering, is a measure of agreement between the partitions; which [11], recommended as "This measure appears to be one of the most popular alternatives for comparing partitions…"(p. 193 – 194).

$$Rand\ index\ (R) = \frac{(SS+DD)}{(SS+SD+DS+DD)} \quad (1)$$

$$R = \frac{(SS+DD)}{M} \quad (2)$$

According to [4] the Rand index lies between 0 and 1. When the two partitions agree perfectly, the Rand index is 1. The disadvantage of the Rand index is that the expected value of the Rand index of two random partitions does not take a constant value (say zero) [5].

- **Adjusted Rand Index:** According to the Web page Wikipedia (2011), the Adjusted Rand index proposed by [11] is the corrected for chance version of the Rand index (see Equation (3)).

$$(ARI) = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}\right] / \binom{M}{2}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] - \frac{\left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}\right]}{\binom{M}{2}}} \quad (3)$$

The Rand index is bounded by 1, and the value 0 is taken when the index equals its expected value [5].

- **Jaccard Index:** Another useful approach to measure the overlapping between partitions is the Jaccard index [1][5]. The Jaccard index may be expressed such as Equation (4):

$$Jaccard\ index\ (J) = \frac{SS}{(SS+SD+DS)} \quad (4)$$

Same as the Rand index, each given attribute of two objects, lies between 0 and 1.

- **Folkes and Mallows Index:** [3] develop another method for comparing partitions. On [11] they give a brief descriptions how this work. Defined such as Equation (5).

$$FM = \sqrt{\frac{SS}{(SS+SD)} \cdot \frac{SS}{(SS+DS)}} \quad (5)$$

Metrics Based on Entropy

The Entropy index (Equation (6)) calculates the sum of the entropies of each cluster n_j. Therefore, if the result consists of objects with only a single label, the entropy is 0. This means that a perfect clustering solution is when the entropy is 0. However, if the clusters that contain documents from single class labels become more diverse, the entropy value will grow more redundancy. The lower the entropy value, the better the clustering is.

$$Entropy\ Index = \sum_{j=1}^{k} \frac{n_i}{N} E_j \quad (6)$$

Evaluation by Set Matching

The Purity index (Equation (7)) is very similar to the entropy index. In order to calculate the purity of the cluster we have to calculate the purity P_j in each cluster. Then to calculate the overall purity index we use the weighted sum of the individual cluster purities.

$$Purity\ index = \sum_{j=1}^{k} \frac{n_j}{N} P_j \quad (7)$$

RESULTS

In this section it is shown an experimental testing using a flat clustering result example. As shown in Table 2, computing validation indexes can take a long time to complete using only one computer to analyze millions of objects. Parallel or distributed computing takes advantage of these validation indexes, by arranging them to work together on the same problem, therefore reducing the time needed to obtain the evaluation solution.

Almost all indexes that are based on the prior knowledge about the data (external criteria) can be described using the so-called confusion matrix, or association matrix or contingency table for the evaluation on counting pairs (see Table 2). The confusion matrix [11] is a K×K′ matrix, whose n_{ij} elements is the number of points that overlaps the pairs of points in the given set of n objects $D = \{O_1, \dots, O_n\}$. Let $n_{i.}$ and $n_{.j}$ be the number of elements that are in whole data set using the next Equation (8) (see Figure 3).

$$n_{ij} = \left|C_K \cap P_{K'}\right| \tag{8}$$

We used a data-set (see Figure 5). This data-set were first evaluated in sequential and then in parallel processing.

PID	P	C
0	1	3
1	4	1
2	1	0
3	0	3
4	0	1
5	0	2
6	4	1
7	1	2
8	4	3
9	4	0

PID	P	C
10	2	1
11	0	3
12	2	0
13	4	0
14	2	1
15	0	2
16	0	0
17	1	2
18	3	2
19	4	0

PID	P	C
20	4	0
21	4	2
22	0	1
23	3	0
24	1	1
25	0	1
26	1	2
27	4	3
28	4	3
29	3	2

PID	P	C
30	1	0
31	0	1
32	4	3
33	1	1
34	4	2
35	4	0
36	4	1
37	2	1
38	1	3
39	3	0

Figure 5
Data Set

Sequential

First, by using the given set of n objects $D\{O_1, \dots, O_n\}$, supposing that $C = \{C_1, \dots, C_k\}$ and $P = \{P_1, \dots, P_{k'}\}$ represent the two partition from the objects in D, we can calculate the confusion matrix shown in Figure 3. The confusion matrix will be $5{\times}5'$, whose n_{ij} elements are represented in Figure 6.

Partition/Cluster	P_0	P_1	P_2	P_3	P_4	Sum
C_0	1	2	1	2	5	11
C_1	4	2	3	0	3	12
C_2	2	3	0	2	2	9
C_3	2	2	0	0	4	8
C_4	0	0	0	0	0	0
Sum	9	9	4	4	14	n = 40

Figure 6
Confusion Matrix from the Data Set

Then, thru computing the confusion matrix, now we can use the counting pairs matrix (see Figure 4). As presented (Equation (9)) [5], [11] that SS + DD can be simplified to a linear transformation of:

$$E\left[\sum_{i,j}\binom{n_{ij}}{2}\right] = \frac{\left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right]}{\binom{n}{2}} = \sum_{i,j}\frac{n_{ij}(n_{ij}-1)}{2} \tag{9}$$

With simple algebra, the counting pairs matrix can be calculated (see Figure 7):

	P	P'
C	$\sum_{i,j}\binom{n_{ij}}{2} = 39$	$\sum_i \binom{n_{.j}}{2} - \sum_{i,j}\binom{n_{ij}}{2} = 136$
C'	$\sum_i \binom{n_{i.}}{2} - \sum_{i,j}\binom{n_{ij}}{2} = 146$	$\binom{n}{2} = 459$

Figure 7
Counting Pairs Matrix Results

The maximum numbers of all pairs in our data set is $M = 39 + 136 + 146 + 459 = 780$ and the total number of points $n = 40$. Now we can directly compute the external validity indexes defined in Table 1 to measure the degree of similarity between P and C. Using the direct values from the counting pairs matrix we can calculate the Rand, Adjusted Rand, Jaccard, and Folekes & Mallows indexes (see Figure 8).

Validation Indexes		Results x	Similarity $0 \le x \le 1$	
			Weak	Strong
1	R	0.6385		√
2	ARI	0.018098	√	
3	J	0.1215	√	
4	FM	0.2168	√	

Figure 8
Validation Indexes Results Using the Figure 7

Table 2
External Indexes Constraints

	Validation Indexes	Notation	Clustering Method	O(*)
1	Rand	$\frac{(SS+DD)}{(SS+SD+DS+DD}$	Flat	$O(n)$
2	Adjusted Rand	$\frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}]/\binom{M}{2}}{\frac{1}{2}\left[[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}] - \frac{[\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}]}{\binom{M}{2}}\right.}$	Flat	$O(n)$
3	Jaccard	$\frac{SS}{(SS+SD+DS)}$	Flat	$O(n)$
4	Folkes and Mallows	$\sqrt{\frac{SS}{(SS+SD)}\cdot\frac{SS}{(SS+DS)}}$	Flat	$O(n)$
5	Entropy	$\sum_{j=1}^{k}\frac{n_{i.}}{n}E_j;\ E_j = \sum_i p_{ij}\,log_2(p_{ij})$	Flat	$O(n)$
6	Purity	$\sum_{j=1}^{k}\frac{n_{i.}}{n}P_j;\ P_j = \frac{1}{n_{i.}}max_i(n_{ij})$	Flat	$O(n)$

For the other two validity indexes presented in the Table 2, Entropy and Purity, we will use the confusion matrix (see Figure 9).

Cluster	P_0	P_1	P_2	P_3	P_4	Entropy	Purity
C_0	1	2	1	2	5	2.04037	0.4545
C_1	4	2	3	0	3	1.95915	0.3333
C_2	2	3	0	2	2	1.97494	0.3333
C_3	2	2	0	0	4	1.5	0.5
C_4	0	0	0	0	0	0	0
			Total			**1.95915**	**0.4**

Figure 9
Entropy and Purity Calculation

As we can see, with respect to the data set, the validity evaluation proves that we fail on having a strong fit in the data set. In other words, the results show that our clustering C posses a weak similarity with P.

Parallel

Task parallelism is a form of computation in which many task are divided to compute a calculation simultaneously distributed, operating with the principle that large scale of data-sets could be computed in small partitions, resulting with the same expected sequential result.

The time complexity of large-scale data sets generally increases, and so with the cluster validation process that evaluates them. The time complexity of sequential or parallel is not discussed in our research.

First, let's try to see if the counting pair's matrix is possible in parallel, which is the key element of calculating most of the external indexes. Lets Imagine that we have the data-set of n objects $D = \{O_1, \dots, O_n\}$ that we discussed earlier, supposing that $C = \{C_1, \dots, C_k\}$ and $P = \{P_1, \dots, P_{k'}\}$ represent the two partition from the objects in D. Then, imagine that the machine achieves to calculate the confusion matrix and then starts the partition of the matrix data set to the salve "tasks" (see Figure 10 and Equation (10)).

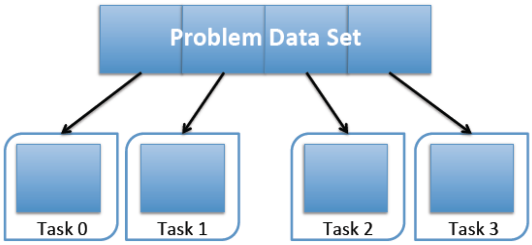

Figure 10
Parallelization of the Data Set

$$\sum_{i=1}^{3} \boldsymbol{S_i} = \boldsymbol{S_0} + \boldsymbol{S_1} + \cdots + \boldsymbol{S_n} \quad (10)$$

Now, let's expand the Equation (11) and we obtain:

$$\frac{n_{ij}(n_{ij}-1)}{2} = \frac{n_{ij_0}(n_{ij_0}-1)}{2} + \frac{n_{ij_1}(n_{ij_1}-1)}{2} \quad (11)$$

Thru the least simplification expression we can see that by the completing the square, a method derivate by the quadratic formula, is not currently possible the calculation of the counting pairs matrix in parallel (see Equation (12)).

$$n_{ij}{}^2 = \left(n_{ij_0} + n_{ij_1}\right)^2 \quad (12)$$

Instead of using the counting pairs matrix, lets use the confusion matrix. By inspection we can see that this method in parallel is easily done. Since, the confusion matrix is a total summation of the clusters, satisfy the early equation discussed.

First, the machine will create a partition of the data set. Then, is given to the slaves to calculate the small confusion matrix partition. Finally, the slave return the results to the machine to create the total confusion matrix and continue with the sequential calculation earlier explain.

CONCLUSION

The problem of determining the strong fit of large-scale data sets is an important issue for clustering processes and also challenging one. In this work, a method to improve the complexity of the evaluation of large-scale data-sets have been presented, based on external criterion inspired in an information theoretic approach to assess the validity of the clustering solutions in parallel.

The experiment carried out on a synthetic data set, using six external indexes; show that, when we create this method in parallel by theoretic the complexity drops, meaning that the validity evaluation will be faster.

As it has also been defined, another issue beyond the scope of this work is the problem of evaluating large-scale hierarchical clustering results. For future work, two possible approaches for validity methods were proposed to define a pair counting strategy that takes into account how close in the dendogram the points are, and to compute several measure for different flat results by 'cutting' the dendogram at various levels, and report their average.

Since, the hierarchical clustering is a structure exploration of the data that encompasses great further contributions to data-mining, information retrieval, among many others; still a need for developing quality measures that assess the quality of the hierarchical clustering.

References

[1] Amigo, E., et al., "A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints", *Springer Standard Collection,* 12(4), 2008, 461 – 486.

[2] Meila, M., "Comparing Clustering", *Association for Computing Machinery,* 2005, 577 – 584.

[3] Fowlkes, E., et al., "A Method for Comparing Two Hierarchical Clustering's", *American Statistical Association,* 78(383), 1983, 553 – 569.

[4] Rand, W., "Objective Criteria for the Evaluation of Clustering Methods", *Journal of the American Statistical Association,* 66, 1971, 846 – 850.

[5] Yeung, K., et al., "Details of the Adjusted Rand Index and Clustering Algorithms Supplement to the Paper – An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data", *Bioinformatics,* 17(9), 2001, 763.] Yeung, K., et al., "Details of the Adjusted Rand Index and Clustering Algorithms Supplement to the Paper – An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data", *Bioinformatics,* 17(9), 2001, 763.

[6] Halkidi, M., et al., "Cluster Validity Methods: Part 1", *ACM Special Interest Group on Management of Data Record,* 3(3), 2002, 42 – 45.

[7] Halkidi, M., et al., "Clustering Validity Checking Methods: Part 2", *ACM Special Interest Group on Management of Data,* 31(3), 2002, 19 – 27.

[8] MacQueen, J., "Some Methods for Classification and Analysis of Multivariate Observations", *Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability,* 5, 1967, 281 – 297.

[9] Heeringa, W., et al., "Validating Dialect Comparison Methods", Classification, Automation, and New Media, 2000, 445 -460.

[10] Pearson, K., "Mathematical Contributions to the Theory of Evolution", *Drapers' Company Research Memoris,* 1904, 13 – 17.

[11] Hubert, L., et al., "Comparing Partitions", *Journal of Classification,* 2(1), 1985, 193 – 218.

[12] Davies, D., L., et al., "A Cluster Separation Measure", *Pattern Analysis and Machine Intelligence, IEEE Transactions,* 1(2), 1979, 224 – 227.

[13] Dunn, J., C., "Well Separated Clusters and Optimal Fuzzy Partitions", *Journal of Cybernetica,* 4, 1974, 95 – 104.

[14] Fraley, C., et al., "How Many Clusters? Which Clustering Method? – Answers Via Model-Based Cluster Analysis", *The Computer Journal,* 41(8), 1998, 578 – 588.

[15] Grimmer, J., et al., "Quantitative Discovery from Qualitative Information: A General-Purpose Document Clustering Methodology", *PNAS,* 108(7), 2011, 2643 – 2650.

[16] Jain, A., K., et al., "Data Clustering: A Review", *ACM Computing Surveys (CSUR),* 31(3), 1999.

[17] Raftery, A., "A Note on Bayes Factor for Long-Linear Contingency Table Models with Vague Prior Information", *Journal of the Royal Statistical Society,* 48(2), 1986, 249 – 250.

[18] Zhaou, Y., "Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering", *Machine Learning,* 55(3), 2004, 311 – 331.