# Cluster Validation using a Probabilistic Attributed Graph

Ana L. N. Fred
*Instituto Superior Técnico*
*Instituto de Telecomunicações*
*afred@lx.it.pt*

Anil K. Jain
*Michigan State University*
*jain@msu.edu*

## Abstract

*We propose a new cluster validity index. A data partition is described by a set of disjoint sub-graphs, each corresponding to the minimum spanning tree of a cluster, taking as edge weight the dissimilarity between linked objects. Based on the assumption that each cluster has a characteristic parametric distribution of dissimilarity increments, graph probabilities are estimated. The validity index is defined as the Minimum Description Length for both estimated model parameters and data partition, according to this probabilistic model. This new index can be used to evaluate various partitions of a given data set obtained by: (i) a single clustering algorithm, (ii) different clustering algorithms, or (iii) cluster ensemble methods. Experimental evaluation of the proposed index on synthetic and real data taken from the UCI repository confirms the usefulness of the method in selecting good clustering solutions.*

## 1. Introduction

Most clustering algorithms require a priori definition of some design parameters, such as the number of clusters, that intrinsically control the quality of obtained data partition. We propose a new cluster validity index for the selection of a partition among different clustering results, thus subsuming the need of a priori knowledge of the number of clusters. We build on the work of Fred et al. [2] who showed that dissimilarity increments, within a "natural" cluster, follow an exponential distribution. Assuming such a dissimilarity increments distribution (DID), we model each data partition by a set of disconnected probabilistic attributed subgraphs. The overall probability of the graph (representing the partition) is derived and a cluster validity index is proposed as the corresponding Minimum Description Length (MDL).

## 2. Cluster Validity Index

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of $n$ objects represented in a $d-$dimensional feature space, $x_i \in \mathcal{R}^d$, and $P = \{C_1, C_2, \ldots, C_{k_P}\}$ be a partition of $X$ into $k_P$ clusters.

### 2.1. Dissimilarity Increments

Let $triplet_{ijk} = (x_i, x_j, x_k)$ consist of three nearest neighbor patterns: $x_j$ is the nearest neighbor of $x_i$, and $x_k$ is the nearest neighbor of $x_j$ such that $x_k \neq x_i$. The dissimilarity increment associated with this triplet is defined as

$$incr(triplet_{ijk}) = |d(x_i, x_j) - d(x_j, x_k)|, \quad (1)$$

were $d(.,.)$ is a distance measure, herein taken as the Euclidean distance. The underlying hypothesis in the present work is that the statistic of dissimilarity increments computed between neighboring patterns within a "natural" cluster follow an exponential model [2]:

$$incr \sim f(incr; \theta) = \theta e^{-\theta incr},\ incr \geq 0, \quad (2)$$

where $\theta$ is the parameter of the DID. Figure 1(b) presents the histogram, and corresponding fitted distribution (according to a maximum likelihood estimate), of dissimilarity increments from one of the clusters in data set Half Rings (figure 1(a)). A data partition, $P$, with $k_P$ clusters has, therefore, an associated vector $\Theta_P = [\theta_1, \theta_2, \ldots, \theta_{k_P}]$ of parameters.

### 2.2. Cluster Graph Generative Model

Let $(x_1,\ x_2)$ be two nearest neighbor patterns, at a distance $d_{12} = d(x_1, x_2)$. According to the DID in (2), a pattern $x_3$ located in the neighborhood of $x_2$ will be at a distance $d(x_2, x_3) = d_{12} \pm incr_{123}$, where $incr_{123}$ is a realization of the random process $f(incr; \theta)$. We introduce a graph generative model for the data in a cluster, where nodes correspond to objects and edges have

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

weights given by the distance between linked objects, with an underlying dissimilarity increments distribution $f(incr;\theta)$. The pseudocode for the graph generative model is given in Algorithm 1.

**Algorithm 1** Graph generative model

**procedure** GRAPHGENMODEL($n_{edges}$, $d_{ini}$, $\theta$)
  Form an initial edge, $e_1$, with weight $d_{ini}$
  **for** $i = 1$ **to** $(n_{edges} - 1)$ **do**
    Randomly select an edge $e_j$ and one of its vertices, $v_j$;
    Let $d_j$ be the weight of $e_j$;
    Draw an increment value, $incr_i$, according to the exponential pdf $f(incr;\theta)$;
    Add a new edge $e_i$ to the graph, connected to $v_j$ and with weight given by: $d_i = d_j \pm incr_i$;
  **end for**
**end procedure**

## 2.3. Probabilistic Attributed Graph (PAG)

We model a cluster structure by a probabilistic attributed graph, where nodes represent objects, and edges have two attributes: (i) the dissimilarity between linked objects, and (ii) the probability of edge formation. Given the nearest neighbor relationships underlying the dissimilarity increments definition, we obtain the graph topology of a cluster as the minimum spanning tree (MST) computed from the completely connected graph with type (i) edge weights (dissimilarity between linked objects). For each edge in the MST, the second attribute, the probability of edge formation, is estimated as follows. According to the graph generating model in section 2.2, an edge $e_i$ (with weight, dissimilarity, $d_i$) is obtained from a connecting edge $e_j$ (with weight $d_j$) with probability $f(incr_i;\theta)$, where $incr_i = |d_i - d_j|$. Since an edge may have several connections with other edges in the MST, and assuming that all these are equally likely sources for the generation of $e_i$, this probability is estimated as:

$$\begin{aligned}\hat{f}(edge_i;\theta) &= \sum_{c\_edge_m} f(incr_m;\theta)P(c\_edge_m) \\ &= \sum_{c\_edge_m} \frac{\theta exp(-\theta incr_m)}{\|c\_edges\|}, \end{aligned} \quad (3)$$

where $c\_edge_m$ is an edge $e_m$ connected to $e_i$, $incr_m$ is the corresponding dissimilarity increment computed from $e_i$ and $e_m$, and $\|c\_edges\|$ is the number of edges directly connecting to $e_i$.

A partition $P$ with $k_P$ clusters is represented by a graph $G$ composed of $k_P$ disconnected subgraphs, $G = \{G_1, G_2, \ldots, G_{k_P}\}$, each $G_i$ corresponding to a graph representation of a cluster $C_i$ in $P$, as described earlier. The probability of a partition $P$ according to this model is then given by:

$$\begin{aligned}\hat{f}(P) &= \prod_{i=1}^{k_P} \hat{f}(G_i|P) \\ &= \prod_{i=1}^{k_P} \prod_{edge_j \in G_i} \hat{f}(edge_j;\theta_i). \end{aligned} \quad (4)$$

## 2.4. Cluster Validity Index

The proposed cluster validity index corresponds to the minimum description length of the graph-based representation of partition $P$, using the probabilistic model in section 2.3, hereafter referred as G-DID index:

$$\text{G-DID}(P) \equiv -log\hat{f}(P) + \frac{k_P}{2}log(n), \quad (5)$$

where the first and second terms represent, respectively, the length of partition description according to the DID hypothesis, and the cost of encoding our estimation of the DIDs. The overall procedure for computing this index is summarized in algorithm 2.

**Algorithm 2** Computing the G-DID Index

Let $P = \{C_1, C_2, \ldots, C_{k_P}\}$ be a data partition.
**for** $i = 1$ **to** $k_P$ **do**
  Estimate $\hat{\theta}_i$ associated with $C_i$;
  Obtain the MST for cluster $C_i$;
  Determine the PAG, $G_i$:
  - $G_i$ topology is given by the MST;
  - edge weights are given by the distance between linked objects in $G_i$;
  - estimate edges probabilities ($\hat{f}(edge_j;\hat{\theta}_i)$, $\forall edge_j \in G_i$) according to eq. (3);

**end for**
Determine partition probability $\hat{f}(P)$ using eq. (4).
Return G-DID index computed using eq. (5):

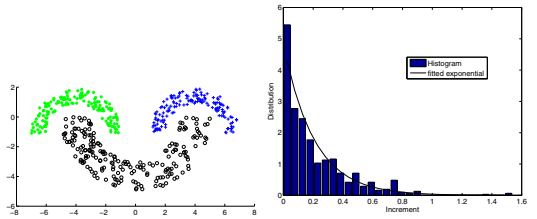
## 2.5. Cluster Validity Criterion

Using the proposed G-DID index, selection among a set of $N$ partitions $\mathcal{P} = \{P^1, P^2, \ldots, P^N\}$ follows a MDL criterion:

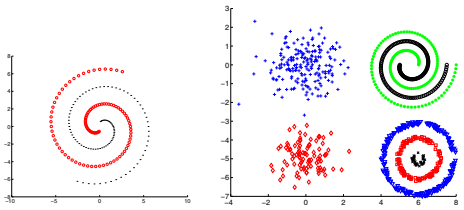
$$\text{Choose } P^i : i = \arg\min_j \{\text{G-DID}(P^j)\}. \quad (6)$$

## 3. Experimental Results and Discussion

The data sets used for evaluation of the proposed index consists of four 2-dimensional synthetic data sets (see figures 1 and 2) and 6 real data sets from the UCI Machine Learning Repository: *Iris*, Wisconsin *Breast-Cancer*, *Optical digits* (from a total of 3,823 samples, each with 64 features, we used a subset composed of the first 100 per digit, for a total of 1000 samples), *Pima Indians*, and *Log Yeast* and *Std Yeast* (consisting of the logarithm and the standardized version, respectively, of gene expression levels of 384 genes over two cell cycles of yeast cell data). For comparison purposes, we provide results obtained with Dunn's index [4].




(a) Half Rings (b) Histogram of increments for data in 1(a)



(c) Spiral (d) Mixed synthetic

**Figure 1. Synthetic data sets and illustrative cluster increments distribution.**

### 3.1. Selection of Algorithmic Parameter

In order to test the ability of the proposed index in selecting appropriate parameters for a specific clustering algorithm, we evaluate it based on the algorithm proposed in [2]. This algorithm has a design parameter, $\alpha$; according to [2], $\alpha = 3$ or 4 leads, in general, to good clustering solutions for well separated clusters, while smaller values of $\alpha$ are needed for data with touching clusters. In our experiments, we applied the algorithm with $\alpha = 1, 2, \ldots, 10$. The quality of each resulting partition $P$ was measured using the consistency index $CI(P, P^o)$, which is the percentage of agreement between $P$ and ground truth information $P^o$.

Figure 2(a) shows the evolution of the proposed G-DID index, normalized to the range [0,1], as a function of $\alpha$ for three data sets. As shown, the shape of these curves is data dependent. Selection of the $\alpha$ parameter based on the minimum G-DID value leads to the following choices: "d3" data set: $\alpha = [3 - 10]$; "Half Rings": $\alpha = [4 - 8]$; "Breast Cancer": $\alpha = 1$. These values of $\alpha$ correspond to the best clustering solutions for the first two data sets, and the second best solution for the Breast-Cancer data; the consistency index, CI, respectively for the three data sets are $CI = 1.0$, $CI = 1.0$, and $CI = 0.818$ (the best partition for breast-cancer had $CI = 0.833$). Figures 2(b) and 2(c) show clustering results for the 2-dimensional synthetic data set (d3) for different values of $\alpha$, where we can see the good quality of the choice made according to the proposed cluster validity index.

| Data Set | CI best | CI G-DID | $\alpha$ | CI Dunn | $\alpha$ |
|---|---|---|---|---|---|
| spiral | 1.00 | 1.00 | 2 | 1.00 | 2 |
| Half Rings | 1.00 | 1.00 | 4 | 0.70 | 2 |
| d3 | 1.00 | 1.00 | 3 | 1.00 | 3 |
| Mixed | 1.00 | 1.00 | 6 | 0.90 | 8 |
| iris | 0.67 | 0.67 | 2 | 0.67 | 2 |
| Breast-C | 0.83 | 0.82 | 1 | 0.83 | 2 |
| log yeast | 0.35 | 0.31 | 2 | 0.31 | 2 |
| std yeast | 0.53 | 0.53 | 1 | 0.41 | 2 |
| opti-digits | 0.41 | 0.41 | 1 | 0.30 | 6 |
| Pima | 0.65 | 0.65 | 2 | 0.65 | 2 |

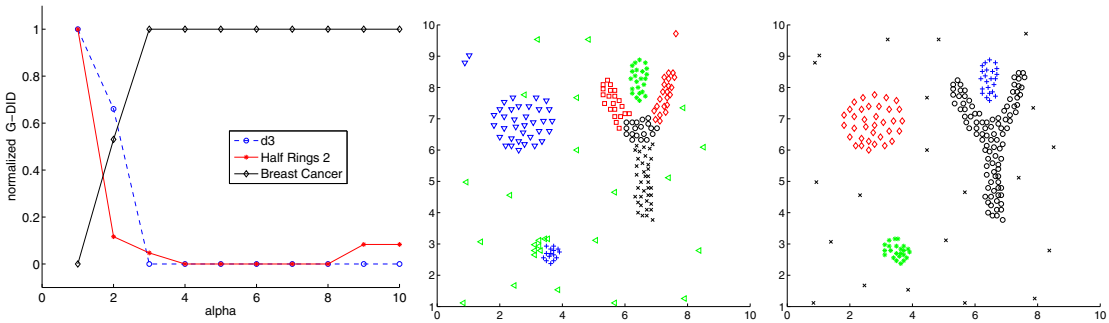
**Table 1. Consistency values ($CI$) for: the best partition (column 2); selected partition and $\alpha$ according to the proposed cluster validity index (columns 3 and 4); and selected partition and $\alpha$ according to Dunn's index (columns 5 and 6).**

Table 1 summarizes the results obtained for the ten data sets. As shown, the proposed criterion correctly selects the best partition (best value of parameter $\alpha$) for 8 out of the 10 data sets, and selects the second best partition for the remaining two data sets (Breast Cancer and Log yeast). Overall, these results outperform the partition selections based on Dunn's index.

### 3.2. Selection of Cluster Ensemble Strategy

Ensemble methods are a promising research area in clustering [3, 5, 1, 6]. Different cluster combination methods lead to distinct data partitioning, and the problem of cluster validity again needs to be addressed. Here we apply the proposed cluster validity index in the evaluation of clustering solutions produced by the Evidence Accumulation Clustering [3].

A single clustering ensemble is produced for each data set by applying the k-means algorithm with random $k$ values in the interval $k \in [10, 30]$, in a total of




(a) Normalized G-DID curves (b) Clustering of d3 data set: $\alpha = 2$ (c) Clustering of d3 data set: $\alpha = 3$

**Figure 2. Performance of proposed method in selecting the clustering parameter $\alpha$ in [2]. (a) Normalized G-DID index for 3 data sets. (b) and (c) Partition of d3 data set for $\alpha = 2$ and $\alpha = 3$.**

| Data Set | CI best | CI G-DID | Alg. | CI Dunn | Alg. |
|---|---|---|---|---|---|
| spiral | 1.00 | 1.00 | SL | 1.00 | SL |
| Half Rings | 1.00 | 1.00 | SL | 0.95 | AL |
| d3 | 0.91 | 0.91 | SL | 0.91 | SL |
| Mixed | 0.75 | 0.75 | SL | 0.75 | SL |
| iris | 0.67 | 0.67 | SL | 0.67 | SL |
| Breast-C | 0.69 | 0.69 | Ward | 0.65 | SL |
| log yeast | 0.37 | 0.37 | SL | 0.20 | CL |
| std yeast | 0.59 | 0.47 | Ward | 0.48 | SL |
| opti-digits | 0.81 | 0.20 | SL | 0.20 | SL |
| Pima | 0.65 | 0.65 | AL | 0.65 | AL |

**Table 2. EAC: Consistency values ($CI$) for the best partition (column 2), and selected partition and corresponding hierarchical algorithm according to the proposed index (columns 3 and 4) and Dunn's index (columns 5 and 6).**

100 runs, with random cluster center initialization. Different hierarchical algorithms are used for the extraction of the combined solution, namely: the Single link, complete link, average link, Wards, and centroid methods. The number of clusters in the combined solution is obtained by applying the lifetime criterion described in [3]. These five methods lead to differing combined solutions. Table 2 shows the best combination results and the selected combined partition with the corresponding hierarchical algorithm. As shown, the best solution is chosen in 8 out of the 10 data sets with the proposed method; Dunn's index only chooses the best in 5 out of ten cases.

## 4. Conclusions

We have addressed the problem of analyzing clustering solutions based on the formalism of probabilistic attributed graphs. Assuming that the dissimilarity values computed between neighboring patterns within a natural cluster follow an exponential distribution, we presented a graph generative model for the clusters. This formed the basis for the design of a new cluster validity index, that consists of the description length of the data partition, represented by a probabilistic attributed graph inferred from the data, conditioned on the given partition. Decision between clustering solutions based on the new index follows a MDL criterion.

We applied the proposed criterion in two distinct scenarios: the selection of a design parameter for a given clustering algorithm and the choice between combination results in a clustering ensemble approach. Results on several data sets, consisting of both synthetic and real data, reveal a good performance of the index in selecting a partition or design parameter.

## References

[1] H. Ayad and M. Kamel. Cluster-based cumulative ensembles. In N. Oza and R. Polikar, editors, *Proc. the 6th International Workshop on Multiple Classifier Systems*, pages 236–245. LCNS 3541, 2005.

[2] A. Fred and J. Leitão. A new cluster isolation criterion based on dissimilarity increments. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(8):944–958, 2003.

[3] A. L. Fred and A. K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, 2005.

[4] U. Maulik and S. Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, 2002.

[5] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2003.

[6] A. Topchy and A. K. Jain. Clustering ensembles: Models of consensus and weak partitionss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(27):1866–1881, 2005.