Nejc ILC

Faculty of Computer and Information Science, University of Ljubljana

Modified Dunn's cluster validity index based on graph theory

Abstract. *Clustering methods serve as common tools for efficient data analysis in many fields of science. The essential, yet often neglected, step in the cluster analysis is validation of the clustering results. This paper presents a novel cluster validity index, which is the modification of the well-known Dunn's index. Our proposal is based on its generalization considering the shortest paths between data points in the Gabriel graph. The experiments show that the proposed index can be successfully applied in the validation of the partitions, even when they contain complex-shaped clusters.*

Streszczenie. *Klasteryzacja danych jest często wykorzystywanym narzędziem analizy w wielu dziedzinach nauki. Ważny, choć często zaniedbywany etap klasteryzacji to ocena wyników tego rodzaju analizy. W pracy tej zaprezentowano nowy indeks oceny klastrów, będący modyfikacją indeksu Dunna. Podejście proponowane w tej pracy jest uogólnieniem, bazującym na poszukiwaniu najkrótszej drogi pomiędzy punktami w grafie Gabriela. Przeprowadzone eksperymenty potwierdzają, że proponowany indeks może być stosowany do oceny podziałów zbiorów danych, nawet jeśli zawierają one klastry o skomplikowanych kształtach.* ***(Zmodyfikowany indeks oceny klastrów Dunna oparty na teorii grafów.)***

Keywords: cluster analysis, cluster validation, Dunn's index, Gabriel graph
Słowa kluczowe: analiza klastrów, ocena klastrów, indeks Dunna, graf Gabriela

Introduction

Clustering is a broadly-used data analysis discipline, fundamental in the field of machine learning, data mining, and pattern recognition [1]. The challenge for clustering methods is to partition input data into natural groups of objects, where a group or a cluster consists of similar objects and where the objects from different groups are as divergent as possible [2].

Unfortunately, no such technique for clustering exists that would cope effectively with any kind of the data, hence the essential step of a cluster analysis is cluster validation [3]. To assess the output of the clustering procedure, a wealth of cluster validity indices has been proposed so far [4]. Generally, they are classified into two groups: internal and external cluster validity indices. The internal indices evaluate the given partition of the data by measuring the compactness and the separation of the clusters on a basis of some objective criteria, without any information about how the true partition should look like. On the contrary, the external indices validate the clustering result with a reference to the partition that is known to be the ground truth. Usually, the ground truth is not known beforehand, so in the real world applications, one should use the internal indices to measure how well do the obtained partitions fit the input data. The majority of clustering algorithms require the expected number of clusters K in the data to be set in advance. Hence multiple runs of an algorithm are executed for different values of K and an internal index is employed afterwards to pick out the best partition.

One of the most used internal validity index was proposed by Dunn in 1973 [5] and since then some of its generalizations have been introduced, using different measures of the compactness and the separation between the clusters [6, 7]. In the paper, we propose a new Dunn-like validity index based on the shortest paths between the data points considering the Gabriel graph on the data. Our index can be seen as a modification of generalized Dunn's index, proposed by Pal and Biswas [6], yet improving its ability to correctly identify good-quality partitions when highly irregular or complex-shaped clusters are present in the data. We compare proposed modification with three other relevant indices while evaluating the partitions of six artificial and two real datasets that are obtained by the single-linkage algorithm [2] and the clustering with normalized cuts [8].

Let us first introduce the Gabriel graph and motivate its usage. Then we present more formally the generalization of the Dunn's index using the Gabriel graph and its proposed modification.

Gabriel graph

One of the most challenging problems for the internal indices is dealing with clusters of non-spherical shapes. Therefore, new approaches have emerged that are based on the graph theory concepts and are able to capture the real structure of the data more efficiently [6, 9]. Pal and Biswas used three types of graphs to impose a structure on the data, i.e., minimum spanning tree, relative neighbour graph, and Gabriel graph. Their results on various datasets show that the generalized Dunn's index based on the Gabriel graph [10] achieves the best performance. Furthermore, connectivity properties of the Gabriel graph prove to be beneficial in the terms of the cluster analysis as shown in [11]. These are the reasons why we adopted the Gabriel graph as the foundation of our research as well.

Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of D-dimensional data points, $\mathbf{x}_i \in \mathbb{R}^D$. A graph $G = (V, E)$ is an ordered pair, where $V = \{v_1, v_2, \ldots, v_N\}$ is a set of vertices and $E = \{e_1, e_2, \ldots, e_L\}$ is a set of edges between the vertices in V. For each data point $\mathbf{x}_i$ there is a vertex v_i that is its abstraction in the graph G, thus $|X| = |V| = N$. The proximity of the vertices $v_i, v_j \in V$ is defined as the Euclidean distance between the corresponding pair of data points, i.e., $d_E(\mathbf{x}_i, \mathbf{x}_j)$. Let edge $e_q = \{v_i, v_j\}$ link the vertex v_i with the vertex v_j and let G be undirected weighted graph – it means that the direction of an edge is neglected and that the edges are weighted by the Euclidean distance between the data points. Thus the weight of the particular edge $e_q = \{v_i, v_j\}$ is computed as $w(e_q) = d_E(\mathbf{x}_i, \mathbf{x}_j)$.

The Gabriel graph is a graph, in which there is an edge $e_q = \{v_i, v_j\}$, if

$$d_E^2(\mathbf{x}_i, \mathbf{x}_j) < d_E^2(\mathbf{x}_i, \mathbf{x}_k) + d_E^2(\mathbf{x}_k, \mathbf{x}_j)\ , \tag{1}$$

$\forall k : v_k \in V, k \neq i, k \neq j$. In other words, vertices v_i and v_j are connected, if there does not exist any other vertex v_k, such that its corresponding data point $\mathbf{x}_k$ would fall into the D-dimensional hypersphere with diameter $d_E(\mathbf{x}_i, \mathbf{x}_j)$ and its centre in $\mathbf{x}_i + (\mathbf{x}_j - \mathbf{x}_i)/2$. Fig. 1a) illustrates this notion with an example of three data points.

In order to compute the Gabriel graph with a greedy algorithm, we have to compute d_E, which has a single-pass time complexity of $O(D)$, and evaluate the condition in Eq. (1) for all the triplets of the data points. So, the overall time complexity of creating the Gabriel graph is $O(D \cdot N^3)$.

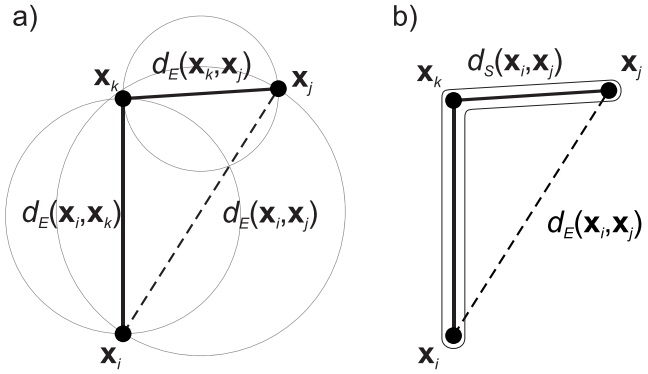

Fig. 1. a) A construction of the Gabriel graph on three data points. Solid line represents an edge between two vertices. b) The Euclidean distance d_E (dashed line) and the shortest distance d_S in the Gabriel graph (solid line highlighted with a contour) between the points $\mathbf{x}_i$ and $\mathbf{x}_j$

Dunn's index and its generalization

Suppose we have partitioned dataset X into K clusters C_i and have obtained the partition $\pi_K = \{C_1, C_2, \ldots C_K\}$, such that $X = \bigcup_{i=1}^{K} C_i$ and $C_i \cap C_j = \emptyset,\ i \neq j,\ C_i \neq \emptyset$. The Dunn's validity index [5] of the partition π_K is computed as

$$\mathrm{DN}(\pi_K) = \frac{\min_{1\leq i\leq K}\{\min_{1\leq j\leq K}\{\mathrm{dist}(C_i, C_j)\}\}}{\max_{1\leq k\leq K}\{\mathrm{diam}(C_k)\}}\ , \tag{2}$$

where

$$\mathrm{diam}(C_i) = \max_{\mathbf{x}_m,\mathbf{x}_n\in C_i}\{d_E(\mathbf{x}_m,\mathbf{x}_n)\} \text{ and} \tag{3}$$

$$\mathrm{dist}(C_i, C_j) = \min_{\substack{\mathbf{x}_m\in C_i,\ \mathbf{x}_n\in C_j \\ i\neq j}}\{d_E(\mathbf{x}_m,\mathbf{x}_n)\}\ . \tag{4}$$

High value of $\mathrm{DN}(\pi_K)$ indicates compact and well separated clusters in the partition π_K, thus we may compute partitions for different number of clusters K, or for any other parameter of the clustering algorithm, and consider the partition that maximizes the Dunn's index as the optimal solution.

To calculate the $\mathrm{DN}(\pi_K)$ we firstly have to compute the distances between all the data points. The between-point distances are then used to calculate the diameter of the cluster in Eq. (3) and the distance between clusters in Eq. (4). There are altogether $N(N-1)/2$ pairs of points, so the computation of the d_E distances requires $O(D\cdot N^2)$ time. The Eq. (2) itself requires $O(K^2)$ time, so the complexity of Dunn's index is $O(D\cdot N^2 + K^2)$ and, considering that in our experiments $K \leq \lceil\sqrt{N}\rceil$, it reduces to $O(D\cdot N^2)$.

However, Pal and Bezdek argue that such definitions of the inter-cluster diameter and the between-cluster distance are too sensitive to noisy data points and are also inconvenient for the validation of non-spherical clusters [7]. As the answer, some indices, enhanced via various types of graphs, were provided; one of these indices uses the concept of the Gabriel graph and in this paper we refer to it as the generalized Dunn's index [6]. The main idea about the generalization is to represent data points $x_i \in X$ with vertices $v_i \in V$ in the Gabriel graph and to use the redefinition of the cluster diameter and the between-cluster distance. Pal and Biswas defined the diameter of the cluster C_i in the following way

$$\mathrm{diam}_G(C_i) = \max_{1\leq q\leq |E_i|}\{e_q\}\ , e_q \in E_i\ , \tag{5}$$

where E_i denotes the set of edges in the Gabriel graph, such that every edge $e_q \in E_i$ connects a pair of vertices that both belong to the cluster C_i. Furthermore, they defined the distance between clusters C_i and C_j as the distance between the cluster centres

$$\mathrm{dist}_G(C_i, C_j) = d_E(\mu(C_i), \mu(C_j))\ , \tag{6}$$

where $\mu(C_i)$ denotes the mean value of all the data points in the cluster C_i. The generalized Dunn's index of the partition π_K is calculated in a very similar way as the original Dunn's index

$$\mathrm{DN}_G(\pi_K) = \frac{\min_{1\leq i\leq K}\{\min_{1\leq j\leq K}\{\mathrm{dist}_G(C_i, C_j)\}\}}{\max_{1\leq k\leq K}\{\mathrm{diam}_G(C_k)\}}\ , \tag{7}$$

where the higher value of DN_G indicates the better partition. To find the optimal partition π_K, we maximize $\mathrm{DN}_G(\pi_K)$ with respect to K. Authors demonstrated a good performance of the DN_G index for both the structural or chain-like clusters and the spherical clusters. They also argued that their proposed index is more resistant to the noise, although this hypothesis is not explicitly proven in the paper [6].

The DN_G index is more expensive to compute than the DN index due to the construction of the Gabriel graph, which takes $O(D\cdot N^3)$ time. Considering that there is at most $3N$ edges in the Gabriel graph [11], Eq. (5) can be computed in $O(N)$ time for all the clusters. The computation of the distances between all the pairs of clusters requires additional $O(D\cdot K^2)$ time, thus it all sums up to $O(D\cdot N^3 + D\cdot K^2 + N)$. Provided that in the experiments $K \leq \lceil\sqrt{N}\rceil$, it finally reduces to the complexity of $O(D\cdot N^3)$.

The proposed modified Dunn's index

Motivated by promising results of the generalized Dunn's index, we introduce its improvement as yet another modification of the way in which the diameter of cluster and the distance between clusters are computed.

The main idea of the proposed approach is to define the distance between a pair of data points in the terms of the shortest path between them in the Gabriel graph. In order to formalize the proposed distance, let us first introduce some essential definitions. Let $G = (V, E)$ be the Gabriel graph built on the dataset X. A path p between the vertices v_i and v_j is a sequence of vertices, such that there is an edge between each of the two successive vertices. A path with no repeated vertices is called a simple path and the length l_p of the simple path p equals the number of edges on p. With $e_r^p,\ r = 1, \ldots, l_p$ we denote an edge on a path p. Suppose there are P possible paths between a pair of vertices $v_i, v_j \in V$. We propose that the distance between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as the sum of edge weights along the shortest path between the vertices v_i and v_j

$$d_S(\mathbf{x}_i, \mathbf{x}_j) = \min_{1\leq p\leq P}\left\{\sum_{r=1}^{l_p} w(e_r^p)\right\}\ , \tag{8}$$

where $w(e_r^p)$ is the weight of the edge on the path p between the vertices v_i and v_j and it equals $d_E(\mathbf{x}_i, \mathbf{x}_j)$. The difference between the distances, i.e., d_E and d_S between the pair of data points $\mathbf{x}_i$ and $\mathbf{x}_j$ is depicted in Fig. 1b), where solid lines represent existing edges in the Gabriel graph. In this simple case, $d_S(\mathbf{x}_i, \mathbf{x}_j) = d_E(\mathbf{x}_i, \mathbf{x}_k) + d_E(\mathbf{x}_k, \mathbf{x}_j)$. To find

the shortest paths between all pairs of vertices in V, we employ a well-known Johnson's algorithm, which can be computed in $O(N^2 \log N + N^2)$ time assuming that $|E| = O(N)$ [12].

Now, the definitions of the cluster diameter in Eq. (3) and the between-clusters distance in Eq. (4) are altered slightly by a substitution of the Euclidean distance $d_E(\cdot,\cdot)$ with the distance $d_S(\cdot,\cdot)$

$$\text{(9)} \qquad \text{diam}_S(C_i) = \max_{\mathbf{x}_m,\mathbf{x}_n \in C_i} \{d_S(\mathbf{x}_m,\mathbf{x}_n)\} \text{ and}$$

$$\text{(10)} \quad \text{dist}_S(C_i,C_j) = \min_{\substack{\mathbf{x}_m \in C_i,\ \mathbf{x}_n \in C_j \\ i \neq j}} \{d_S(\mathbf{x}_m,\mathbf{x}_n)\} \ .$$

In the Gabriel graph, there always exists a path between any of two vertices (for proof, see [11]), so $d_S(\cdot,\cdot)$ is a non-negative real number for all the pairs of the data in X. We can now define the modified Dunn's index as

$$\text{(11)} \qquad \text{DN}_S(\pi_K) = \frac{\min_{1\leq i\leq K}\{\min_{1\leq j\leq K}\{\text{dist}_S(C_i,C_j)\}\}}{\max_{1\leq k\leq K}\{\text{diam}_S(C_k)\}} \ .$$

Its behaviour when validating the partition π_K is exactly the same as with the DN or the DN_G index – the larger the value of $\text{DN}_S(\pi_K)$, the better the partition π_K is considered to be. Thus, the maximum diameter of the clusters in the partition π_K should be minimized and the minimum distance between any of two clusters should be maximized in order to optimize the DN_S index.

When analysing the time complexity of the DN_S index, we consider three computational steps. Firstly, the Gabriel graph is built in $O(D \cdot N^3)$ time. Secondly, Johnson's algorithm is applied, which takes $O(N^2 \log N + N^2)$ time. Finally, Eq. (11) requires $O(K^2)$ time. Hence, the complexity of the DN_S index becomes $O(D \cdot N^3 + N^2 \log N + N^2 + K^2)$, which asymptotically equals $O(D \cdot N^3)$. The time complexity of the DN_S index is therefore higher than that of the DN index ($O(D \cdot N^2)$) and grows as fast as that of the DN_G index ($O(D \cdot N^3)$).

The reader should note that the proposed variant of the Dunn's index (DN_S) is also a generalization of the original Dunn's index (DN), but for the sake of clarity we refer to it as the modified Dunn's index to avoid a confusion with the DN_G index.

Experimental evaluation

We experimentally demonstrate the performance of the proposed index DN_S using the following protocol. We chose six artificial datasets, i.e., `wave`, `ring`, `moon`, `flag`, `spiral`, and `halfring`, and two real benchmark datasets, i.e., `iris` and `wine`. The first four datasets are available at [13], the datasets `spiral` and `halfring` are obtained from [14], whereas `iris` and `wine` are from the UCI repository [15]. Due to the fact that all the used datasets were synthetically generated or annotated by experts, we actually know the true partitions (cluster labels) and the expected number of clusters, denoted with K_T. See Table 1 for the detailed description of the datasets and Fig. 2 for the plots of the artificial ones. Duplicate data points were removed in the preprocess step.

To partition the datasets, two clustering algorithms were employed: the hierarchical single-linkage (SL) algorithm [2] and the normalized cuts (NC) algorithm [8]. Both algorithms require the expected number of clusters K to be given as an input parameter. The algorithms were executed for

Table 1. Datasets used in the experiments, containing N points in D dimensions. The number of clusters K_T is a man-given ground truth

dataset	N	D	K_T
`wave`	287	2	2
`ring`	800	2	2
`moon`	514	2	4
`flag`	640	2	3
`spiral`	200	2	2
`halfring`	400	2	2
`iris`	150	4	3
`wine`	178	13	3

$K = 2, \ldots, \lceil\sqrt{N}\rceil$, where N is the number of data points, as in [17]. Additionally, the algorithm NC requires parameter σ to be set in order to transform the distances between the data points to the similarities. In all experiments we used the following heuristic $\sigma = 0.05 \cdot \max_{\mathbf{x}_i,\mathbf{x}_j \in X} d_E(\mathbf{x}_i,\mathbf{x}_j)$, as recommended by the authors of the NC algorithm.

Let Π denote the set of the obtained partitions of a particular dataset using one of the clustering algorithms, $\Pi = \{\pi_K;\ K = 2, \ldots, \lceil\sqrt{N}\rceil\}$. Each partition was then validated by the proposed index DN_S and three other relevant validity indices: the Dunn's index (DN), the generalized Dunn's index (DN_G), and the index of connectedness (Conn) [9]. The latter evaluates the degree to which neighbouring data points have been placed in the same cluster. It is defined as

$$\text{(12)} \quad \text{Conn}(\pi_K) = \sum_{i=1}^{N}\left(\sum_{j=1}^{L} c_{i,\text{nn}_i(j)}\right) ,$$

$$\text{(13)} \qquad c_{i,\text{nn}_i(j)} = \begin{cases} \frac{1}{j} & \text{if } \nexists C_k : \mathbf{x}_i, \text{nn}_i(j) \in C_k \\ 0 & \text{otherwise} \end{cases} ,$$

where $\text{nn}_i(j)$ is the jth nearest neighbour of the data point x_i, and L is a parameter determining the number of neighbours that contribute to the index Conn. We followed the procedure in [16] and set the parameter L to a value of 5 for all the experiments. Minimum value of $\text{Conn}(\pi_K)$ indicates the optimal partition with respect to the number of clusters K.

Following the classical methodology for evaluating the cluster validity indices, every index makes a guess about the best partition considering the set of partitions Π on a given dataset. The best partition π_K^* predicted by the cluster validity index $\text{CVI}(\pi_K)$ is the partition, such that

$$\text{(14)} \qquad \pi_{K^*} = \arg\max_{\pi_K \in \Pi} \{\text{CVI}(\pi_K)\} \ ,$$

where K^* is the number of clusters in the best partition and function $\text{CVI}(\pi_K)$ represents an arbitrary index that should be maximized, i.e., DN, DN_G, and DN_S, whereas the Conn index should be minimized, so the following is the case

$$\text{(15)} \qquad \pi_{K^*} = \arg\min_{\pi_K \in \Pi} \{\text{Conn}(\pi_K)\} \ .$$

It is said that the index has made a correct guess, if the number of clusters K^* in the selected partition π_{K^*} equals the true number of clusters K_T, given as the ground truth. However, Gurrutxaga et al. [17] argue that this methodology makes an important and sometimes false assumption that the clustering algorithm works well. In other words, it is expected that the partition π_{K_T}, which contains the true num-

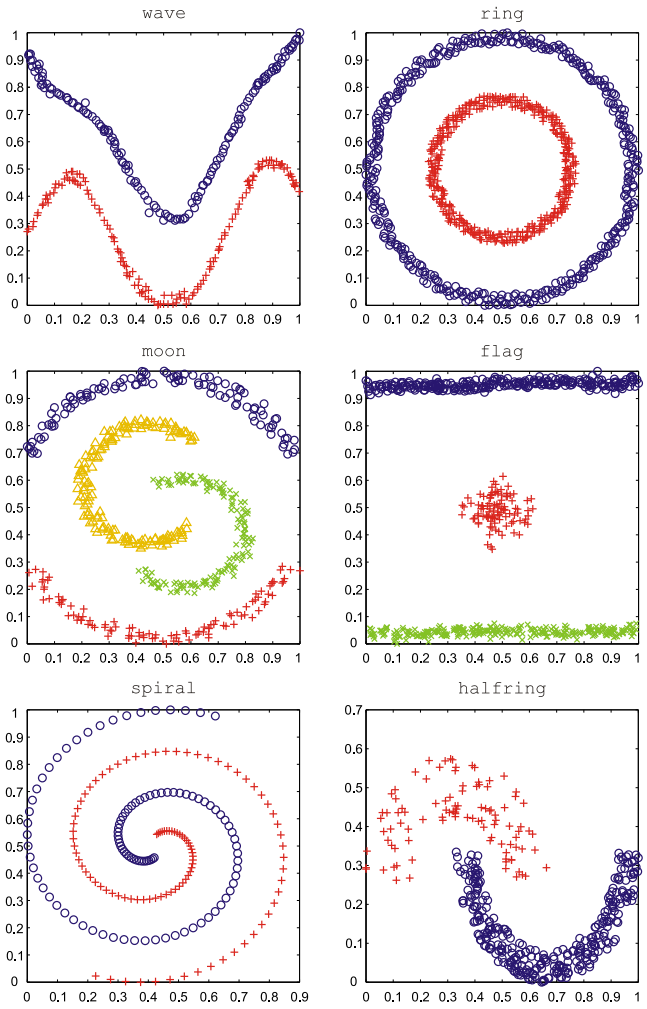

Fig. 2. The artificial datasets used in the experiments. The ground truth partitions are displayed using different shapes and colors

ber of clusters, fits the data better than any other partition in Π. As this is often not the case, especially it is not true for the datasets with complex-shaped clusters, an alternative methodology was proposed using external validity indices to measure the similarity between the partition π_K and the true partition denoted with π_T, which is assumed to be known. We followed this alternative methodology for the evaluation of the internal indices, so let us briefly formalize it.

Let $\mathrm{eCVI}(\pi_i, \pi_j)$ be an external validity index, called also a similarity measure, that returns high value, if the partitions π_i and π_j are similar and small value, if they are not similar. According to [17], the most similar partition π_{sim} with respect to the true partition π_T is defined as

$$\pi_{\mathrm{sim}} = \arg\max_{\pi_K \in \Pi} \left\{ \mathrm{eCVI}(\pi_K, \pi_T) \right\} . \tag{16}$$

We search among all the partitions made by a particular clustering algorithm to find the one, which is the most similar to the ground truth partition π_T. It follows that the internal index makes a successful guess, if it predicts π_{sim} as the best partition, i.e., if $\pi_{K^*} = \pi_{\mathrm{sim}}$. This approach is argued to be more general than the classical methodology and free of any assumption about the correctness of the used clustering algorithm, which is why we adopted it in the evaluation of the results.

In our experiments, we used three different external validity indices to determine the partition π_{sim}: Classification Accuracy (CA) [18], Normalized Mutual Information (NMI) [19], and Variation of Information (VI) [18]. The CA index is defined as the ratio of correctly clustered data points to all the data points. In order to calculate CA, the optimal covering, relating maximization of intersection between the partition π_K and true partition π_T is considered.

The NMI index is an external measure of cluster quality based on information theory and is computed as

$$\mathrm{NMI}(\pi_i, \pi_j) = \frac{\sum_{h=1}^{i} \sum_{l=1}^{j} N_{h,l} \log \left(\frac{N \cdot N_{h,l}}{N_h^i N_l^j} \right)}{\sqrt{\left(\sum_{h=1}^{i} N_h^i \log \frac{N_h^i}{N} \right) \left(\sum_{l=1}^{j} N_l^j \log \frac{N_l^j}{N} \right)}}, \tag{17}$$

where i is the number of clusters in the partition π_i and j is the number of clusters in the partition π_j. N is the number of all data points, N_h^i is the number of data points in the cluster $C_h \in \pi_i$, N_l^j is the number of data points in the cluster $C_l \in \pi_j$ and $N_{h,l}$ is the number of data points that are in the cluster $C_h \in \pi_i$ as well as in the cluster $C_l \in \pi_j$.

Finally, yet another information-theoretic index, the VI index, is defined as

$$\mathrm{VI}(\pi_i, \pi_j) = H(\pi_i) + H(\pi_j) - 2I(\pi_i; \pi_j) , \tag{18}$$

where the entropy of the partition π_i is defined as $H(\pi_i) = -\sum_{C_h \in \pi_i} p(C_h) \log p(C_h)$ and the mutual information between the partitions π_i, π_j is defined as $I(\pi_i; \pi_j) = \sum_{C_h \in \pi_i} \sum_{C_l \in \pi_j} p(C_h, C_l) \frac{\log p(C_h, C_l)}{p(C_h) p(C_l)}$. The probability $p(C_h)$ that a randomly chosen data point belongs to the cluster C_h is computed as $p(C_h) = N_h/N$, where N_h denotes the number of data points in the cluster C_h and N denotes the number of all data points in the dataset. Joint probability of two clusters $C_h \in \pi_i$ and $C_l \in \pi_j$ is defined as $p(C_h, C_l) = |C_h \cap C_l|/N$.

Results and discussion

Table 2 and Table 3 list the results of the comparison between four cluster validity indices, Conn, DN, DN_G, and DN_S, using three external validity indices, CA, NMI, and VI together with the information about the true clustering as the evaluation criteria. The partitions that were validated by the internal indices were created from eight datasets by the SL and the NC clustering algorithms. The upper-left part of the both tables shows the target number of clusters for each dataset; in the first column there is the true number of clusters K_T. It is followed by the number of clusters of the partition π_{sim} (see Eq. (16)) according to CA, NMI, and VI. The upper-right part of the tables contains the columns with the guessed number of clusters of the partitions π_{K^*} that are considered by the cluster validity indices to be the optimum (see Eq. (14) and Eq. (15)). The evaluation of the validity indices' performance, or better their correctness, can be assessed by counting the number of occasions the particular index made a successful guess. We call this number the score; for example, S_{VI} denotes the score of the particular internal validity index according to the external criterion VI. For each agreement between the target and the guessed number of clusters, we increase the score by one. Obviously, the higher the score, the better the performance of the index becomes.

Let us first take a look at Table 2 that presents the predicted number of clusters and the scores of the four compared indices when partitioning the datasets via the SL clustering algorithm. The first observation is that the target num-

Table 2. Results of the experiment on the partitions made by the SL clustering algorithm. The best scores are put in bold

dataset	target K				guessed K			
	K_T	K_{CA}	K_{NMI}	K_{VI}	Conn	DN	DN_G	DN_S
wave	2	2	2	2	2	2	3	2
ring	2	2	2	2	2	2	5	2
moon	4	4	4	4	2	2	5	3
flag	3	3	3	3	2	2	3	3
spiral	2	2	2	2	2	2	4	2
halfring	2	5	5	5	2	5	2	5
iris	3	6	2	2	2	2	2	2
wine	3	2	13	2	2	2	2	2
score				S_{VI}	5	6	3	**7**
				S_{NMI}	4	5	2	**6**
				S_{CA}	4	5	2	**6**
				S_{T}	**4**	3	2	**4**

Table 3. Results of the experiment on the partitions made by the NC clustering algorithm. The best scores are put in bold

dataset	target K				guessed K			
	K_T	K_{CA}	K_{NMI}	K_{VI}	Conn	DN	DN_G	DN_S
wave	2	2	2	2	2	2	2	2
ring	2	2	2	2	2	2	4	2
moon	4	4	4	4	2	2	4	4
flag	3	3	3	3	2	2	5	3
spiral	2	4	6	6	2	14	14	14
halfring	2	3	3	3	3	11	4	4
iris	3	3	2	2	2	2	2	2
wine	3	3	3	2	2	6	2	2
score				S_{VI}	5	3	4	**6**
				S_{NMI}	4	3	3	**5**
				S_{CA}	3	2	2	**4**
				S_{T}	3	2	2	**4**

ber of clusters determined by the external criteria differs from K_T for datasets halfring, iris, and wine – this is due to the inability of the SL algorithm to find the perfect partitions, as defined by the ground truth, of the mentioned datasets when partitioning the data into the K_T clusters. Indeed, these are the cases where the assumption of the clustering algorithm's correctness is not true and is better to compare the guessed number of clusters with the other values of the target K rather than with K_T. For instance, when the Conn and DN_G indices predicted two clusters in the halfring dataset, it is considered as a lucky guess, because all of the three external indices show that the most similar partition to the true partition is the one with five clusters. However, it is evident from the both tables that according to the S_{VI}, S_{NMI}, and S_{CA}, the proposed validity index DN_S achieved the best score. There is a tie between DN_S and the Conn index when the true partition is the considered criterion. Furthermore, the Conn and the DN index performed quite well, too, with one or two wrong predictions more than DN_S, depends on which external criterion we select as a reference. It is quite surprising that the DN_G index demonstrated the poorest results, although it was designed especially for the validation of complex-shaped, non-spherical clusters [6].

When evaluating the cluster validity indices one should be aware that different clustering algorithms produce different partitions for the same number of clusters, which is usually given as the input parameter. That is why we repeated the experiment, employing yet another method, the NC algorithm. The results are presented in Table 3 and we can say that in general they are quite similar to those in Table 2, especially when considering the scores in the lower part of the table. Again, the DN_S index performed better than the others. It is interesting to note that in the case of clustering the moon dataset with the NC algorithm the indices DN_G and DN_S were able to correctly identify the optimum partition with four clusters, while they were wrong in the case of the SL algorithm – DN_G and DN_S predicted five and three clusters respectively. This is due to the diversity within the partitions when they are made by different clustering algorithms. Furthermore, the Conn index has been lucky enough to guess that two is the optimum number of clusters in the dataset spiral even though the partition with two clusters demonstrated lower similarity when compared with the ground truth, than the partitions with four or six clusters. On the contrary, the Dunn-like indices DN, DN_G, and DN_S wrongly predicted that there are 14 clusters in the dataset spiral, which is a considerably large number. The same is evident also from the predictions of the DN index in the case of the dataset halfring. A more detailed inspection of the clustering results of the NC algorithm revealed that, in contrast to the SL algorithm, it tends to form compact clusters that are well balanced regarding the number of data points they contain. When increasing the number of clusters on the input of the NC algorithm, the cluster compactness grows as well. So the diameter of the cluster, as defined by Eq. (3), Eq. (5), and Eq. (9), decreases, which causes the value of DN, DN_G, and DN_S to increase. Therefore, it would be beneficial to consider a way to penalize solutions with high number of clusters in such cases, which remains a subject of further improvements.

Conclusion

In this paper we proposed a modification of the Dunn's cluster validity index that is based on the usage of the Gabriel graph to represent connections between the data points. Our contribution relates to the novel definition of the distance between a pair of data points, which is calculated as the sum of edge weights on the shortest path between the pair of points, considering connections in the Gabriel graph. We have experimentally demonstrated that the novel validity index DN_S performs similar or better than the other three relevant indices when identifying the best partitions of some artificial and real datasets. We conclude that the proposed index can be successfully employed in the cluster validation process, also when the input data consists of complex-shaped clusters. According to the results of the experiments, we consider the presented work as an efficient improvement over the Dunn's index and its generalization.

In the future, we plan to conduct additional experiments on high-dimensional biological data to assess the performance of the novel validity index in real-life situations. Furthermore, we will focus on the integration of the proposed index into the cluster-ensemble framework introduced by Vega-Pons et al. [16], where the partitions in the ensemble are weighted by the internal validity indices before they are merged into the consensus partition.

BIBLIOGRAPHY

[1] Bishop C.M., Pattern Recognition and Machine Learning, Springer, 2006

[2] Xu R., Wunsch II D., Survey of clustering algorithms, *IEEE Transactions on Neural Networks*, 16 (2005), 645–678

[3] Handl J.,Knowles J., Kell D.B., Computational cluster validation in post-genomic data analysis, *Bioinformatics*, 21 (2005), 3201–3212

[4] Halkidi M., Batistakis Y., Vazirgiannis M., On Clustering Vali-

dation Techniques, *Journal of Intelligent Information Systems*, 17 (2001), 107–145
[5] Dunn J.C., A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Journal of Cybernetics*, 3 (1973), 32–57
[6] Pal N.R., Biswas J., Cluster validation using graph theoretic concepts, *Pattern Recognition*, 30 (1997), 847–857
[7] Bezdek J., Pal N., Some new indexes of cluster validity, *IEEE Transactions on Systems, Man, and Cybernetics*, 28 (1998), 301–315
[8] Shi J., Malik J., Normalized Cuts and Image Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (2000), 888–905
[9] Handl J., Knowles J., Exploiting the Trade-off –The Benefits of Multiple Objectives in Data Clustering. In: Coello Coello C., Hernández Aguirre A., Zitzler E., editors, *Evolutionary Multi-Criterion Optimization*, vol. 3410 of LNCS, Springer Berlin / Heidelberg, 2005, 547–560
[10] Gabriel K.R., Sokal R.R., A New Statistical Approach to Geographic Variation Analysis, *Systematic Zoology*, 18 (1969), No. 3, 259–278
[11] Matula D.W., Sokal R.R., Properties of Gabriel Graphs Relevant to Geographic Variation Research and the Clustering of Points in the Plane, *Geographical Analysis*, 12 (1980), 205–222
[12] Cormen T.H., Leiserson C.E., Rivest R.L., Stein C., Introduction to algorithms, MIT Press, 2001
[13] Ilc N., Artificial datasets for clustering, URL: `http://laspp.fri.uni-lj.si/nejci/data/`, 2011
[14] Kuncheva L.I., Clustering data, URL: `http://pages.bangor.ac.uk/~mas00a/activities/artificial_data.htm`, 2011
[15] Frank A., Asuncion A., UCI Machine Learning Repository, URL: `http://archive.ics.uci.edu/ml/`, 2010
[16] Vega-Pons S., Correa-Morris J., Ruiz-Shulcloper J., Weighted partition consensus via kernels, *Pattern Recognition*, 43 (2010), 2712–2724
[17] Gurrutxaga I., Muguerza J., Arbelaitz O., Pérez J.M., Martín J.I., Towards a standard methodology to evaluate internal cluster validity indices, *Pattern Recognition Letters*, 32 (2011), 505–515
[18] Meila M., Comparing clusterings – an information based distance, *Journal of Multivariate Analysis*, 98 (2007), 873–895
[19] Strehl A., Ghosh J., Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions, *Journal of Machine Learning Research*, 3 (2002), 583–617

Author: *Dipl.Ing. Nejc Ilc, Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia, E-mail: nejc.ilc@fri.uni-lj.si*