A Study on the Relationship between Internal and External Validity Indices Applied to Partitioning and Density-based Clustering Algorithms

Caroline Tomasini, Eduardo N. Borges, Karina Machado and Leonardo Emmendorfer
Centro de Ciências Computacionais, Universidade Federal do Rio Grande – FURG,
Av. Itália, km 8, 96203-900, Rio Grande, RS, Brazil

Keywords: Cluster Evaluation, Validity Index, Regression.

Abstract: Measuring the quality of data partitions is essential to the success of clustering applications. A lot of different validity indices have been proposed in the literature, but choosing the appropriate index for evaluating the results of a particular clustering algorithm remains a challenge. Clustering results can be evaluated using different indices based on external or internal criteria. An external criterion requires a partitioning of the data previously defined for comparison with the clustering results while an internal criterion evaluates clustering results considering only the data properties. In a previous work we proposed a method for selecting the most suitable cluster validity internal index applied on the results of partitioning clustering algorithms. In this paper we extend our previous work validating the method for density-based clustering algorithms. We have looked into the relationships between internal and external indices, relating them through linear regression and regression model trees. Each algorithm was run over synthetic datasets generated for this purpose, using different configurations. Experiments results point out that *Silhouette* and *Gamma* are the most suitable indices for evaluating both the datasets with compactness property and the datasets with multiple density.

1 INTRODUCTION

Clustering is an unsupervised data mining task based on the similarity between the instances (Tan et al., 2006). A cluster is a subset of instances that can be treated collectively as one group (Han et al., 2006). A clustering algorithm should maximize intragroup and minimize intergroup similarity. Nowadays, data clustering is widely used in several scientific or organizational applications such as complex data analysis, market research, image processing, test of hypotheses and profiles discovery (Xu and Wunsch, 2009).

Several clustering algorithms have been proposed in recent decades (Xu et al., 2005; Berkhin, 2006). *k-means* (Hartigan and Wong, 1979) is based on centroids and it requires that the user defines the *k* number of groups. Besides, *k-means* is not suitable for discovering clusters with nonconvex shapes or clusters of very different size. Other examples of partitioning methods are *k-medoids* and CLARANS. DBSCAN (Ester et al., 1996) is another well-known algorithm that grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise. This algorithm requires two user-defined parameters: a neighborhood specified by the radius ε and the minimum number of points $MinPoints$ in this neighborhood. Hierarchical methods work by grouping data objects into a tree of clusters (dendrogram) that shows how objects are grouped together step by step. DIANA and ROCK are examples of hierarchical algorithms (Han et al., 2006). There are also grid-based clustering algorithms which quantize the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. An overview of many other clustering algorithms can be found in the following surveys (Xu et al., 2005; Berkhin, 2006).

Validation of clustering results is essential to the success of clustering applications (Halkidi et al., 2001a). The quality of partitions generated by different algorithms can be evaluated using visual inspection and different indices based on external or internal criteria. Due to the high dimensionality and cardinality of the datasets, visual inspection becomes impracticable. An external criterion requires previously known data classes for comparison with the partitioning of the data resulted from the clustering algorithm (Xu and Wunsch, 2009). However, the vast majority of problems which require the use of grouping tech-

Tomasini, C., Borges, E., Machado, K. and Emmendorfer, L.
A Study on the Relationship between Internal and External Validity Indices Applied to Partitioning and Density-based Clustering Algorithms.
DOI: 10.5220/0006317000890098
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 1*, pages 89-98
ISBN: 978-989-758-247-9

niques do not have the data labelled *a priori*. Therefore, a common way to evaluate the clustering results is using indices based on internal criteria, which regard only the data properties, looking for partitioning with compact and well-separated clusters.

Several cluster validity indices have been proposed in the literature (Xu and Wunsch, 2009; Rand, 1971; Fowlkes and Mallows, 1983; Davies and Bouldin, 1979; Dunn, 1974; Rousseeuw, 1987; Baker and Hubert, 1975; Vendramin et al., 2010; Hubert and Levin, 1976). Each index focuses on a particular property of the partitions, and many of them are influenced by the impact of noise, density variation, or the presence of subclusters. It is not possible to point out a universally most reliable index (Liu et al., 2010; Vendramin et al., 2010). For this reason, selecting appropriate indices for evaluating the results of a particular clustering algorithm remains a challenge (Liu et al., 2010).

In order to help in this process of selecting the most suitable cluster validity internal indices we previously proposed a methodology based on the induction of regression models applied on the results of partitioning clustering algorithms (Tomasini et al., 2016). Now in this paper we have extended our previous work by adapting and validating the methodology for density-based algorithms, which were performed over synthetic datasets generated for this purpose, using different configurations. Clustering results were evaluated by different internal and external indices generating the input for regression models. Each external index is taken as the label attribute to be predicted, and internal indices are the regression attributes. Experiments results show the relationships between internal and external indices and point out that *Silhouette* and *Gamma* are the most suitable indices for evaluating both the datasets with compactness property using *k-means* and the datasets with multiple density using DBSCAN.

This paper is organized as follows. In section 2, we describe a set of cluster validity indices used in this work. In section 3, we present our methodology for choosing the most suitable cluster validity internal index. In section 4, we give details on the performed experiments and discuss the obtained results. Finally, in section 5, we draw our conclusions and point out some future work directions.

2 CLUSTERING VALIDATION

One of the most important issues in clustering analysis is the evaluation of results to find the partitioning that best fits the underlying data (Liu et al., 2010). This procedure is known under the terms clustering validation (Tan et al., 2006) or cluster validity (Halkidi et al., 2001a).

There are two criteria proposed for clustering validation and selection of an optimal clustering scheme (Berry and Linoff, 1996):

1. compactness – the members of each cluster should be as close to each other as possible;
2. separation – the clusters themselves should be widely spaced.

A common index of compactness is the variance (Han et al., 2006), which should be minimized. The separation can be measured by means of the distance between the clusters centers or between the nearest or most distant members. There are many different cluster validity indices that are very useful as quantitative measures for evaluating the quality of data partitions.

Although several indices had been proposed, each one focuses on a particular clustering property, and they may not deal with some aspects such as variation of density or noise. These properties or aspects turn each index able to outperform others in specific classes of problems. Based on the above arguments, choosing the appropriate validity index for evaluating the results of a particular clustering scheme remains a challenge.

Liu et al. (Liu et al., 2010) present a detailed study of eleven internal cluster validity indices investigating their validation from the impact of monotonicity, noise, density, subclusters and skewed distributions. The performed experiments show that most of indices have certain limitations in different application scenarios.

Vendramin et al. (Vendramin et al., 2010) proposed a statistical methodology for comparing cluster validity indices that is more robust than the traditional method used in the literature (Milligan and Cooper, 1985). The authors compute the Pearson correlation between internal and external indices in order to identify relationships between them. The experiments show that the larger the correlation value the higher the capability of an internal index to unsupervisedly mirror the behavior of the external index and properly distinguish between better and worse partitions.

As proposed by Vendramin et al. (Vendramin et al., 2010), our method described in section 3 discover the behavior and relationships between internal and external indices. These relationships are learned from linear regression models and regression model trees. We have performed a set of experiments presented in section 4. These experiments show that regression models can quantify the relationships between internal and external indices helping the user to choose the most suitable cluster validity internal index. The followings subsections describe a set of indices used in this work.

2.1 External Indices

The external indices are typically used to compare the cluster results with a previously known partitioning (Xu and Wunsch, 2009). This partitioning can reflect our intuition about the data structure, be suggested by a domain expert or be defined based on a matching between the clusters found and the labels already known.

Let P be a previously defined partition of the dataset X with n points. Let C be a clustering structure resulted from a clustering algorithm performed on X. The evaluation of C by an external index is achieved by comparing C to P. Considering a pair of data points $(x_i,x_j) \in \{X \times X\} | 1 \leq i \leq\ n, 1 \leq j \leq\ n$, one can compute the four different cases based on how x_i and x_j are placed in C and P (Xu and Wunsch, 2009), i.e., the frequency of pairs x_i and x_j which belong to:

a. the same group in C and the same category in P;

b. the same group in C, but different categories in P;

c. different groups in C, but the same category in P;

d. different groups in C and different categories in P.

In this work, we have applied the external indices *Jaccard* (Xu and Wunsch, 2009), *Rand* (Rand, 1971) and *Fowlkes-Mallows* (Fowlkes and Mallows, 1983). These indices are based on the frequency of instance pairs correct or incorrectly grouped according to one of the cases a, b, c and d. The following subsections specify each index.

2.1.1 Jaccard

Jaccard index J or *Jaccard* similarity coefficient (Xu and Wunsch, 2009) is a statistical measure used to compare the similarity and diversity between datasets. This index results in values that range in the close interval $[0,1]$. J returns a value closer to 0 when applied to different partitions and closer to 1 when computed on very similar partitions. The *Jaccard* index is defined by equation 1.

$$J = \frac{a}{a+b+c} \tag{1}$$

2.1.2 Rand

So as *Jaccard*, *Rand* index R (Rand, 1971) measures the similarity between two partitions P and C. This index also results in values in the range $[0,1]$, where 0 suggests that C and P are very different and 1 means highly similar partitions. *Rand* index is defined by equation 2.

$$R = \frac{a+d}{a+b+c} \tag{2}$$

2.1.3 Fowlkes-Mallows

The value of this index is directly related to the similarity between C and P, which means that higher the returned value, higher is the similarity between the partitions. *Fowlkes-mallows* index FM (Fowlkes and Mallows, 1983) is defined by equation 3.

$$FM = \sqrt{\frac{a}{a+b}\ \frac{a}{a+c}} \tag{3}$$

2.2 Internal Indices

In practice, external information such as class labels is often not available in many application scenarios. Therefore, in the situation that there is no external information available, internal validity indices are the only option for clustering validation (Liu et al., 2010). Typically, these indices are able to quantify the quality of clustering results using only frequencies and properties inherent to the dataset (Halkidi et al., 2001a), e.g., considering only the proximity matrix.

In this work, we have applied the internal indices *DBI* (Davies and Bouldin, 1979), *Dunn* (Dunn, 1974), *Gamma* (Baker and Hubert, 1975; Vendramin et al., 2010), *C-index* (Hubert and Levin, 1976; Vendramin et al., 2010) and *Silhouette* (Rousseeuw, 1987). Since we have applied the internal indices in different configurations of the clustering algorithms in order to better understand the behavior and the relationships among this clustering validation methods, we have used the internal indices as relative criteria. According to Xu et al. (Xu and Wunsch, 2009), relative criteria compare clustering results generated by different algorithms or the same algorithm but with different input parameters, i.e., they evaluate a clustering result comparing it to other clustering schemes (Halkidi et al., 2001a). The following subsections specify each index cited previously.

2.2.1 DBI

The Davies-Bouldin index *DBI* (Davies and Bouldin, 1979) is the ratio between the sum of the internal dispersion of clusters and the distance between them. Equation 4 defines *DBI* as

$$DBI = \frac{1}{n}\sum_{i=1}^{n} max_{i \neq j} \frac{\sigma_i + \sigma_j}{d(c_i,c_j)} \tag{4}$$

where n is the number of clusters, σ_i is the average distance of all points of the cluster i to its centroid c_i, σ_j is the average distance of all points in the cluster j to its centroid c_j and $d(c_i, c_j)$ is the distance between the centroids c_i and c_j.

It is clear for the above definition that *DBI* is the average similarity between each cluster and its most similar correspondent (Halkidi et al., 2001a). It is desirable for the clusters to have the minimum possible similarity to each other. Thus, smaller values of *DBI* correspond to compact clusters which centroids are distant from each other.

2.2.2 Dunn

The *Dunn* index D (Dunn, 1974) is calculated from the ratio between the shortest intergroup distance and longest intragroup distance. It returns values in the range $[0, \infty)$, where higher values attempts to identify compact and well separated clusters (Halkidi et al., 2001a).

Let $d(C_i, C_j)$ be the distance between two clusters C_i and C_j performed as the shortest distance between a pair of objects $x \in C_i$ and $y \in C_j$ (equation 5). Let $diam(C_i)$ be the diameter of C_i calculated as the maximum distance between two of its members (equation 6). *Dunn* index is formally defined by equation 7, where k is the number of clusters.

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} (d(x,y)) \tag{5}$$

$$diam(C_i) = \max_{x,y \in C_i} (d(x,y)) \tag{6}$$

$$D(k) = \min_{i=1,\ldots,k} \left(\min_{j+1,\ldots,k} \left(\frac{d(C_i, C_j)}{\max_{l=1,\ldots,K} (diam(C_l))} \right) \right) \tag{7}$$

The main disadvantage in relation to other indices is the high quadratic computational complexity. Furthermore, this criterion is very sensitive to noise.

2.2.3 Gamma

The *Gamma* index Γ (Baker and Hubert, 1975; Vendramin et al., 2010) computes the number of concordant pairs of objects S_+, which is the number of times the distance between a pair of objects from the same cluster is lower than the distance between a pair of objects from different clusters. This index also calculates the number of discordant pairs of objects S_-, which is the number of times the distance between a pair of objects from the same cluster is greater than the distance between a pair of objects from different clusters. Γ is defined by equation 8.

$$\Gamma = \frac{S_+ - S_-}{S_+ + S_-} \tag{8}$$

This index varies in the range $[-1, 1]$. Better partitions are expected to have higher values of S_+, lower values of S_- and, therefore, higher values of Γ.

2.2.4 C-Index

The C-index (Hubert and Levin, 1976; Vendramin et al., 2010) is defined by Equation 9

$$C = \frac{d_w - min(d_w)}{max(d_w) - min(d_w)} \tag{9}$$

where d_w is the sum of distances over all pairs of instances from the same cluster. Let j be the number of pairs of instances in the same cluster, $max(d_w)$ and $min(d_w)$ are the sum of the j largest and smallest distances, respectively, considering all pairs of distances. Thus, this index should be minimized and it varies within the range $[0, 1]$.

2.2.5 Silhouette

The *Silhouette* index S (Rousseeuw, 1987) defines the quality of clustering results based on the proximity among objects of a particular cluster and the neighborhood of these objects to the nearest cluster. Equation 10 computes this index for a single instance x, member of the cluster j, where $d(x, C_j)$ is the average dissimilarity between x and all the objects in j, h is the cluster nearest to x.

$s(x)$ varies within the range of $[-1, 1]$. The closer to 1 the better the object allocation. After computing s for all data objects in the cluster j, the average S_j is calculated (equation 11). Finally, *Silhouette* index is computed for the entire partition as defined by equation 12, where n_j is the number of objects in the cluster j and k is the number of clusters.

$$s(x) = \frac{d(x, C_h) - d(x, C_j)}{\max(d(x, C_h), d(x, C_j))} \tag{10}$$

$$S_j = \frac{\sum_{i=1}^{n_j} s(x_i)}{n_j} \tag{11}$$

$$S = \frac{\sum_{j=1}^{k} S_j}{k} \tag{12}$$

3 PROPOSED METHOD

This section presents our proposed method to relate internal and external cluster validity indices. It is split into five steps presented in Figure 1.

The 1st step sets the datasets to be used. The instances in this dataset must be labelled with predefined partitions that will be used in clustering validation external criteria. For this step we can select real

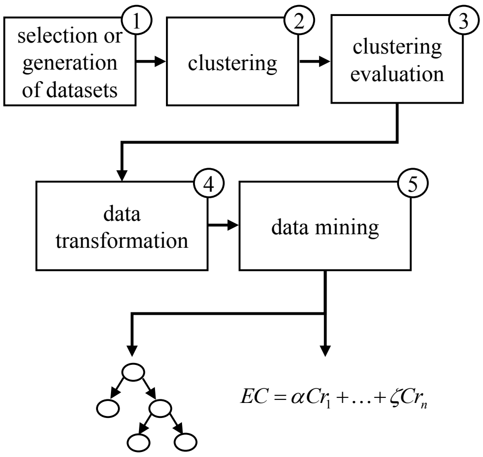
selection or generation of datasets
1
clustering
2
clustering evaluation
3
data transformation
4
data mining
5
$EC = \alpha Cr_1 + \ldots + \zeta Cr_n$

Figure 1: Proposed method to relate internal and external cluster validity indices.

datasets or generate synthetic ones. We notice that the use of synthetic datasets allows us the variation of properties such as number of instances, features and clusters, density, noise, and so on.

The 2nd step consists in selecting and applying a clustering algorithm on the datasets generated or selected in the previous step. This choice depends on the data properties. For instance, we recommend using a low computational complexity density-based algorithm if the datasets are very large having clusters of very different size with non convex shapes and multiple densities. The algorithm should be performed several times varying the input parameters looking for different partitions.

The 3rd step computes a set of validity internal and external indices for each clustering result, i.e., for each partition returned in previous step. These indices quantify the quality of clustering results.

In step 4, the index values previously computed are transformed to be used as input for data mining. Minimization criteria such as *DBI* and *C*-index are inverted since most indices are maximization criteria. Thus, for all transformed indices, compact and well-separated clusters will return higher values. A transformed index value t_{ip} is defined by equation 13, where v_{ip} is the index i value computed on the partition p and $\bar{v}_i$ is the average of index i values considering all partitions. In addition, all transformed values are normalized in the interval $[0,1]$.

$$t_{ip} = \begin{cases} 2\,\bar{v}_i - v_{ip}, & \text{if } i \text{ is a minimization criterion} \\ v_{ip}, & \text{otherwise} \end{cases} \tag{13}$$

Following our method, given a set of n partitions generated in 2nd step and their normalized internal index values computed in step 4, the performance of each index on the assessment of the whole set of partitions is evaluated with respect to a normalized external index in step 5. Thus the quality of clustering results are supervisedly quantify.

Table 1: An example of the training set with Jaccard (*J*) as target attribute.

	DBI	*S*	*D*	Γ	*C*	*J*
1	0.759	0.793	0.357	0.723	0.487	**0.410**
2	0.763	0.697	0.314	0.624	0.211	**0.410**
...						
n	0.871	0.753	0.247	0.675	0.401	**0.417**

Table 1 shows an example of training set where *Davies-Bouldin*, *Silhouette*, *Dunn*, *Gamma* and *C-index* were taken as predictive attributes and the external index *Jaccard* was the target attribute. These training sets are used as input of regression models to estimate the external index values based on the internal ones. The analysis of models helps to verify which internal index is most suitable for evaluating the datasets generated or selected in first step. A good internal index will be able to rank the partitions according to an ordering that is similar to that established by an external index, since it relies on supervised information about the data structure (known clusters).

4 EXPERIMENTAL EVALUATION

This section describes the experiments we conducted in order to empirically validate the method proposed in Section 3 and evaluate the quality of the regression models. Two hundred synthetic datasets were generated during the evaluation. These datasets contain multiple distributions of points in a two-dimensional space, which allows us to visually check the validity of the results (Halkidi et al., 2001b). The experiments were performed in a standard personal computer, using the statistical computing software R [1] in the steps 1 to 4. The last step was conducted using the data mining software Weka [2].

Experiments are divided into two different case studies (CS), using:

1. datasets with compactness property and a partitioning clustering algorithm;
2. datasets with multiple density and a density-based clustering algorithm.

The following metrics were used to evaluate the obtained models: correlation coefficient and root rela-

[1] http://www.r-project.org/

[2] http://www.cs.waikato.ac.nz/ml/weka/

tive squared error (Han et al., 2006). The experimental results point out that *Silhouette* and *Gamma* are the most suitable cluster validity internal indices for evaluating the scenarios specified in both case studies.

4.1 CS1: Datasets with Compactness Property and a Partitioning Clustering Algorithm

In the 1st step of the proposed method we have generated 100 distinct datasets, in which 150 instances was distributed in a two-dimensional space. We varied the number of real clusters n_c (classes) ranging from 2 to 5. The number of instances for each cluster was random always totaling 150. Half of the instances follows a Gaussian distribution, in which the radius r have been set from 1 to 10 and the standard deviation $sd = 0.33\ r$. The other half was randomly arranged between the largest and the smallest coordinates of the previously generated instances, simulating noise. Figure 2 (up) shows an example of dataset with three distinct classes identified by the color of the points.

After datasets generation, in 2nd step, we have applied the *k-Means* clustering algorithm setting up $k = n_c$. In 3rd step the clustering results were evaluated using all the cluster validity indices described in section 2. In step 4, for each partition, the cluster validity indices were transformed and normalized. Figure 2 (down) shows the partition returned when *3-means* was applied on the generated dataset (up).

In the step 5, we have trained a linear regression model for each cluster validity external index using the internal ones (Eq. 14-16). We have kept the Weka default parameters.

$$J = 0.4691\ S - 0.2766\ \Gamma - 0.2447\ C + 0.363 \quad (14)$$

$$R = -0.1746\ S + 0.2492\ \Gamma - 0.0543\ C + 0.508 \quad (15)$$

$$FM = 0.581\ S - 0.3891\ \Gamma - 0.2563\ C + 0.5333 \quad (16)$$

Observing the linear regression models, we notice that *DBI* and *Dunn* were irrelevant to estimate all external indices because they were not used in any regression. *Silhouette* had a positive impact on *Jaccard* and *Fowlkes-Mallows* and a negative on *Rand* index. The *Gamma* behavior was opposed because it positively affected only *Rand*. The *C-index* influence was always negative.

We have also learned regression model trees using the *M5* algorithm (Quinlan et al., 1992) to estimate the same three external indices. We have set the minimum number of instances to allow at a leaf node $M = 4$. Figure 3 shows the regression tree for the *Jaccard* index.

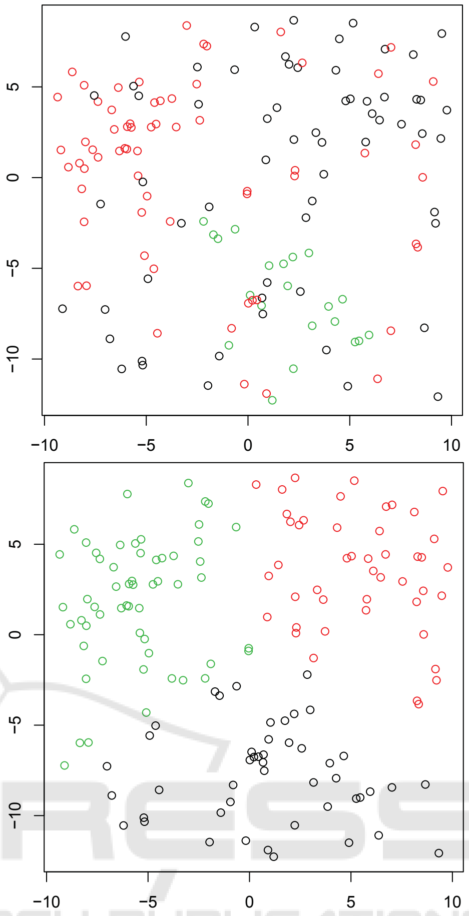

Figure 2: An example of generated dataset (up) and the partition returned by k-means (down).

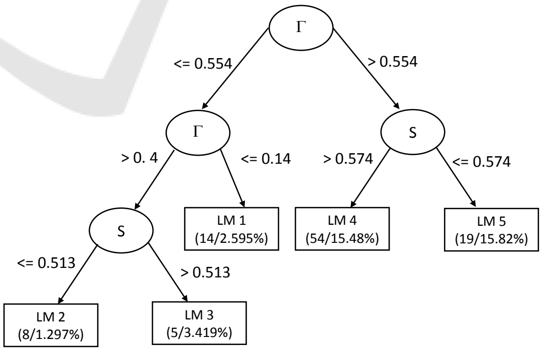

Figure 3: Model tree using the *Jaccard* index as target attribute.

Each leaf is a distinct linear regression model (LM) constructed using a subset of instances, which are selected using the rule defined by the path between the root and the leaf. These nodes present the number of instances and the prediction error of each LM. *Gamma* was the most discriminatory attribute (root) followed by *Silhouette*. Any other index helped to

Table 2: Regression evaluation using *k-means*.

Index	Regression	Correlation	RRSE
J	linear	97.82	20.78
	M5	98.75	15.83
R	linear	91.82	39.61
	M5	96.60	25.91
FM	linear	98.01	19.86
	M5	98.01	19.86

build the tree, but all LM have a minimal influence of *DBI* in addition to previously cited indices. For all LM, *Silhouette* had a positive impact in the Jaccard prediction while *Gamma* and *DBI* had a negative.

Regarding the use of the *DBI* as an internal node, the regression tree for the *Rand* index did not produce significant differences in the 10 LM (leaf nodes), which were composed by the same indices including *Silhouette* and *Gamma*. For the *Fowlkes-Mallows* index, *M5* produced a single node tree with the same LM presented in equation 16.

Table 2 compares the linear regressions with the regression model trees. For each learned model it shows two quality measures: correlation coefficient and root relative squared error (RRSE), both in percentages (Han et al., 2006). We notice that *M5* performed better than or equal to linear regression model.

The analysis of models allows us to verify that the most suitable cluster validity internal indices for evaluating the datasets using *k-means* were *Silhouette* and *Gamma*. These internal indices was strong related to all three external ones, i.e. they were able to evaluate the vast majority of partitions in a very similar way when compared to *Jaccard*, *Rand* and *Fowlkes-Mallows*, as shown by the high values of correlation and low error rates.

4.2 CS2: Datasets with Multiple Density and a Density-based Clustering Algorithm

In this second case study, we have applied the proposed method to multiple density datasets (Handl et al., 2005) and a density-based clustering algorithm.

In the 1st step we generated 100 distinct synthetic datasets with 150 instances in each one distributed in a two dimensional space. For these datasets we varied the number of classes as follows: 25 datasets with 2 classes, 25 datasets with 3 classes and so on. The 150 points of each dataset were generated using a bivariate Gaussian distribution (Goodman, 1963) defined by eq. 17, where μ_x is the mean of x, μ_y is the mean of y, σ_x is the standard deviation of x, σ_y is the standard deviation of y and ρ is the correlation.

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right) \quad (17)$$

We have decided to use the bivariate Gaussian distribution (Goodman, 1963) to generate datasets with different shapes. To achieve these shapes we varied the σ_x and σ_y randomly from 0.5 to 1.0 in each real cluster. This assures a random variation of spreading around x and y for each cluster. The variation of density was obtained considering a random number of points in each real cluster. We established that each cluster must have at least 2 points. Considering an example of dataset with two clusters, we have $Size_{C1} = random[2,148]$, $Size_{C2} = 150 - Size_{C1}$. Figure 4 (up) shows an example of a dataset with 4 real clusters. We can notice that these clusters have different density and shape.

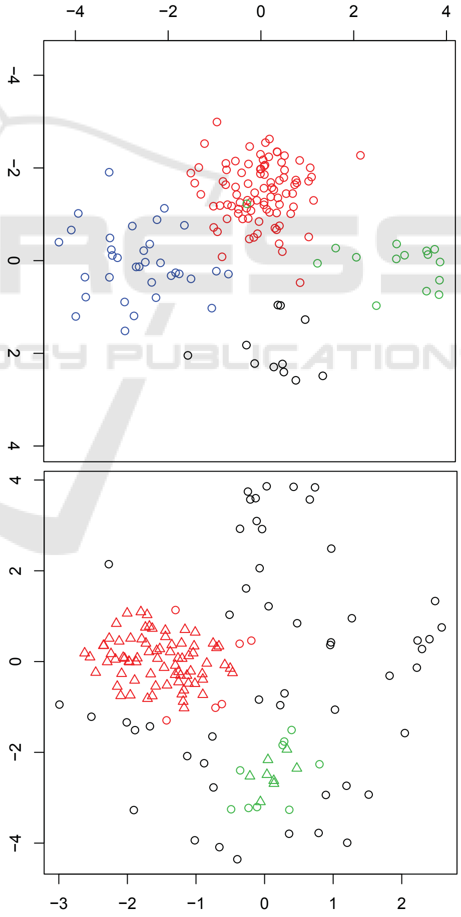

Figure 4: An example of generated dataset with 4 real clusters (up) and the partition obtained after perform DBSCAN (down) with 3 clusters.

The second step on our proposed methodology consists in applying the DBSCAN algorithm to each generated dataset. A classic problem on density-based clustering algorithms is to define the appropriate values for the radius (ε) and minimum number of points (*MinPoints*) to generate a good clustering result. Often this problem is solved by trial and error (Chaimontree et al., 2010). However since we have 100 datasets we decide to adopt the strategy proposed in (Zhou et al., 2012), which these parameters are adaptive and self-adjusting without manual intervention.

The approach to determine the parameters ε and *MinPoints* for each dataset is divided into four stages (Zhou et al., 2012):

i. Calculate the distance matrix for all data points and sort the values of this matrix an ascending order line by line. Compute the average of each column (ε_vector). So, $\varepsilon\_vector[i]$ is the average distance between a point and the i-th closer point.

ii. Calculate *MinPoints* using ε_vector. Firstly we verify how many points there are within ε considering the values of the distance matrix. For instance, on the first line of the distance matrix we verify how many distances are smaller than the $\varepsilon\_vector[1]$, we apply the same for second line and so on. These values compose the *MinPoints_vector*.

iii. Apply the *DBSCAN* algorithm using ε_vector and *MinPoints_vector* as parameters. For each pair $\varepsilon\_vector[i]$, *MinPoints_vector*$[i]$ we perform DBSCAN and store the number of resulted clusters, composing the parameter matrix as exemplified on Table 3.

iv. Using the parameter matrix we can find out when the number of groups stabilizes and locate the optimal values of ε and *MinPoints*.

Following our method in the 2nd step we apply the DBSCAN algorithm for the 100 generated datasets considering the parameters ε and *MinPoints* calculated for each dataset as previously explained. Then, in 3rd step, the clustering results are evaluated using all the cluster validity indices described in section 2.

Table 3: Example of parameter matrix generated to obtain the values of ε and *MinPoints* for a dataset.

n	ε	*MinPoints*	*k*	Stability
1	0.00	-1	150	-
2	0.23	-1	81	69
3	0.34	1	49	32
4	0.42	3	10	39
5	0.50	4	6	4
6	0.55	6	5	1
7	0.60	7	3	2
8	0.65	8	4	-1
9	0.69	**10**	**3**	1
10	0.73	11	3	0

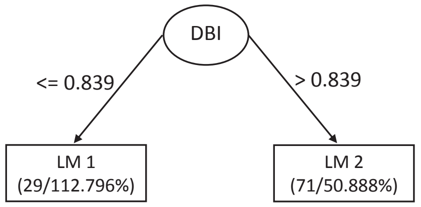

Figure 5: Model tree using *Jaccard* as target feature obtained on the second case study.

In step 4, the cluster validity indices were transformed and normalized to be used as input data for data mining. Figure 4 (down) shows the partition returned by DBSCAN applied on the original dataset (up) considering $\varepsilon = 0.493$ and $MinPoints = 6$.

In step 5, we apply the linear regression algorithm available on Weka for each cluster validity external index using the internal ones, considering the same parameter values used in case study 1 (Eq. 18-20).

$$J = 0.1927\,S + 0.1712\,\Gamma + 0.0268\,DBI + 0.353 \tag{18}$$

$$R = 0.5798\,\Gamma + 0.2868 \tag{19}$$

$$FM = 0.2585\,S + 0.6064 \tag{20}$$

We can notice that *Silhouette* has positive effect on *Jaccard* and *Fowlkes-Mallows* external indices. *Gamma* is also an important index for density-based algorithms since it has positive effect on *Jaccard* and *Rand*. The impact of *C-index* is insignificant since it does not appear on these linear regression models.

We have also applied the *M5* algorithm to obtain regression model trees which estimate *Jaccard*, *Rand* and *Fowlkes-Mallows*. For *Rand* and *Fowlkes-Mallows* indices we obtain single node trees with linear models equals to the linear regression equations 19 and 20 respectively. For *Jaccard* we obtained the model tree presented in figure 5. The LMs described on equations 21-22 were obtained by inducing linear regression models from the leaf nodes of this tree.

$$J_{LM1} = 0.0657\,S + 0.0584\,\Gamma + 0.0091\,DBI + 0.4464 \tag{21}$$

$$J_{LM2} = 0.0336\,S + 0.2484\,\Gamma + 0.0047\,DBI + 0.353 \tag{22}$$

From theses equations we can notice that *C-index* was insignificant again since does not appear on the LMs. DBI was important to split the instances in the root but had insignificant contribution on the LMs. Similar to linear regression *Gamma* index has positive contribution on the Jaccard value. Table 4 compares the results of linear regression and model trees. We can notice that *Rand* and *Fowlkes-Mallows* have equal results for the two algorithms. The difference is only for *Jaccard* results.

Table 4: Regression evaluation using DBSCAN.

Index	Regression	Correlation	RRSE
J	linear	65.48	75.58
	M5	67.59	73.73
R	linear	72.59	68.78
	M5	72.59	68.78
FM	linear	64.29	76.59
	M5	64.29	76.59

The analysis of models allows us to verify that the most suitable cluster validity internal indices for evaluating the generated datasets using DBSCAN were *Silhouette* and *Gamma*. The values of quality measures suggest a moderate correlation between them and all three external indices.

5 CONCLUSION

In this paper we have investigated the relationships between internal and external clustering validity indices learning a set of regression models. The analysis of these models allowed the inference of the most suitable internal index for each method of clustering algorithm. The experiments results point out that *Silhouette* and *Gamma* were the most suitable indices for evaluating the datasets with compactness propriety using *k-means* and the datasets with multiple density using DBSCAN.

Finally, our method can be seen as a template for a general strategy for selecting an internal validity index in which specific clustering or regression algorithms may be replaced by more effective or efficient ones in specific scenarios. As future work we highlight the performance of new experiments using different clustering algorithms and real datasets.

REFERENCES

Baker, F. B. and Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349):31–38.

Berkhin, P. (2006). A survey of clustering data mining techniques. In Kogan, J., Nicholas, C., and Teboulle, M., editors, *Grouping Multidimensional Data*, pages 25–71. Springer Berlin Heidelberg.

Berry, M. J. and Linoff, G. (1996). Data mining techniques for marketing, sales and customer support. john willey & sons. *Inc., 1997, 454 P.*

Chaimontree, S., Atkinson, K., and Coenen, F. (2010). Best clustering configuration metrics: Towards multiagent based clustering. In *Advanced Data Mining and Applications*, pages 48–59. Springer.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569.

Goodman, N. (1963). Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *Annals of mathematical statistics*, pages 152–177.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001a). On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3):107–145.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001b). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145.

Han, J., Kamber, M., and Pei, J. (2006). *Data mining, southeast asia edition: Concepts and techniques*. Morgan kaufmann.

Handl, J., Knowles, J., and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, pages 100–108.

Hubert, L. J. and Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological bulletin*, 83(6):1072.

Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). Understanding of internal clustering validation measures. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 911–916, Washington, DC, USA.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

Quinlan, J. R. et al. (1992). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. Singapore.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Tan, P.-N., Steinbach, M., Kumar, V., et al. (2006). *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston.

Tomasini, C., Emmendorfer, L., Borges, E. N., and Machado, K. (2016). A methodology for selecting the most suitable cluster validation internal indices. In *Proceedings of the 31st Annual ACM Symposium on*

Applied Computing, pages 901–903, New York, NY, USA. ACM.

Vendramin, L., Campello, R. J., and Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4):209–235.

Xu, R. and Wunsch, D. (2009). *Clustering.* IEEE Press, Piscataway, NJ, USA.

Xu, R., Wunsch, D., et al. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678.

Zhou, H., Wang, P., and Li, H. (2012). Research on adaptive parameters determination in dbscan algorithm. *Journal of Information & Computational Science*, 9(7):1967–1973.