Information Sciences 487 (2019) 1–17

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

External clustering validity index based on chi-squared statistical test

José María Luna-Romera*, María Martínez-Ballesteros, Jorge García-Gutiérrez, José C. Riquelme

Department of Computer Languages and Systems, ETSII, University of Seville, Spain

ARTICLE INFO

Article history:
Received 2 July 2018
Revised 15 February 2019
Accepted 17 February 2019
Available online 18 February 2019

Keywords:
Clustering analysis
External validity indices
Comparing clusters
Big data

ABSTRACT

Clustering is one of the most commonly used techniques in data mining. Its main goal is to group objects into clusters so that each group contains objects that are more similar to each other than to objects in other clusters. The evaluation of a clustering solution is a task carried out through the application of validity indices. These indices measure the quality of the solution and can be classified as either internal that calculate the quality of the solution through the data of the clusters, or as external indices that measure the quality by means of external information such as the class. Generally, indices from the literature determine their optimal result through graphical representation, whose results could be imprecisely interpreted. The aim of this paper is to present a new external validity index based on the chi-squared statistical test named Chi Index, which presents accurate results that require no further interpretation. Chi Index was analyzed using the clustering results of 3 clustering methods in 47 public datasets. Results indicate a better hit rate and a lower percentage of error against 15 external validity indices from the literature.

1. Introduction

Clustering is one of the many techniques in data mining. Its goal is to partition unlabelled data into clusters where instances within the same cluster are similar and instances grouped in other clusters are dissimilar to said clusters [1]. This technique has been applied in many fields, such as biological sciences [2], medicine [3], energy [4], chemical [5].

There are numerous clustering methods, and in general, each method produces a different clustering solution. In certain cases, the same method with different parameters could result in different solutions. The evaluation of the results is one of the most important issues in cluster analysis. Measuring the quality of a clustering solution is as important as the clustering method itself [6]. There exist clustering validity indices (CVI) that measure the quality of the solution, and these CVIs have commonly been used in the literature [7–13].

These measures could be classified into either internal or external CVIs. Internal CVIs are based on how the instances are distributed across the clusters by using the data by itself. When there is no external information, these kinds of indices present the only option available for the evaluation of the clustering solution because they depend on certain properties of the results, such as the compactness of the clusters or the separation between them. Compactness of clusters could be

* Corresponding author.
E-mail addresses: jmluna@us.es (J.M. Luna-Romera), mariamartinez@us.es (M. Martínez-Ballesteros), jorgarcia@us.es (J. García-Gutiérrez), riquelme@us.es (J.C. Riquelme).

https://doi.org/10.1016/j.ins.2019.02.046
0020-0255/

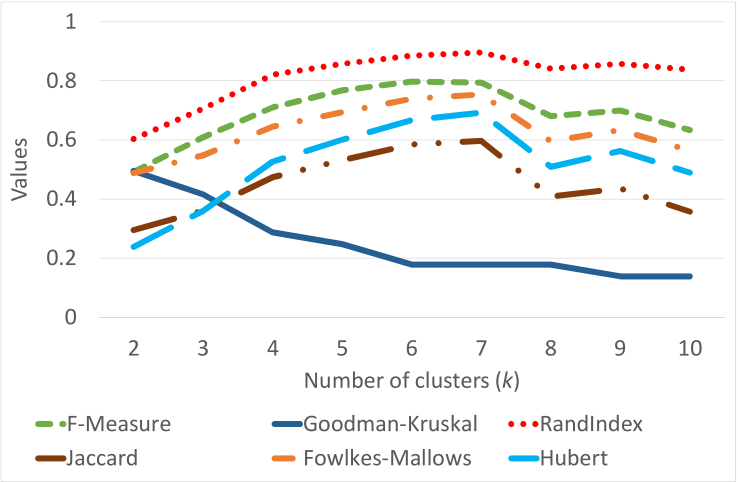


Fig. 1. Results of the CVIs from the literature for $k = 2$ to 10 number of clusters for zoo dataset whose optimal number of clusters is 7.

defined as the mean distance of separation between the instances within a cluster. Separation by itself is defined as the distance between the instances of different clusters. These indices seek a high level of compactness within each cluster and a considerable gap between clusters [14].

On the other hand, external indices use external information, such as class labels, to measure the quality of a clustering solution. These kinds of indices verify the quality of the clustering result by comparing it with the ground truth partition. In this case, the indices know in advance the optimal number of clusters for a dataset since ground truth holds this information [15]. This paper focuses on these external CVIs. Generally, CVIs from the literature determine their optimal result with a local minimum, a local maximum, or by following the elbow method [16–18], and the results could be imprecisely interpreted.

The purpose of this paper is to present an innovative external CVI based on the chi-squared statistical test, henceforth named Chi Index, which presents the results accurately without the need for interpretation. The effectiveness of the new index has been compared with 15 indices from the literature using 47 public datasets and 3 clustering methods from Spark MLlib [19] which made it possible to use this index in big data environments.

The remainder of this paper is organized as follows. Section 2 discusses the literature of external CVIs. In Section 3, the proposed new index is defined. Section 4.3 presents the experimental setup, the methodology followed and the results. The paper ends with the conclusions and suggested future work in Section 5.

2. External indices

An external index evaluates a clustering result C by comparing it against the ground truth partition G. A taxonomy of external indices could be established that depends on the criterion of how the clustering result and the ground truth partition are compared [20]. These indices can be classified into *set matching, pair-counting*, and *information theory*.

- *Set matching* is the category which assumes that the instance label of every cluster has corresponding instances in said cluster. Indices from the literature based on *set matching* include those known as *purity* [21], *F-measure* [22], *Criterion H* [23], *CSI* [24], *PSI* [20], and *Goodman–Kruskal* [25].
- The criterion known as *pair-counting* is based on the comparison between the number of instances with the same label and the cluster result. This category includes the *Rand index* [26], the *adjusted Rand index* [27], *Jaccard* [28], *Fowlkes–Mallows* [29], *Hubert Statistic* [30], and *Minkowski score* [31].
- Indices based on *information theory*, such as *entropy* [21], *variation of information* [32], and *mutual information* [33], have also been applied in the literature.

A list of the equations of these indices is given in Table 1. As mentioned above, the results that show these indices need to be interpreted since each index indicates the optimal number following the rules of the local maximum, the local minimum, or the “elbow method”. Figs. 1 and 2 illustrate two examples of the results for the CVIs from the literature for zoo and gesture datasets from the UCI repository whose optimal number of clusters is 7 and 5, respectively. In Fig. 1, it could be said that the CVIs follow a pattern, whereby the majority indicate point out the optimal number of clusters to be 7 with a local maximum, although Goodman–Kruskal indicates the optimal by following the elbow method. This figure shows that most of the CVIs also have a local maximum at 9, and this could be misleading in the cases when the optimal number of clusters remains unknown in advance. Fig. 2 corresponds to a dataset whose optimal number of clusters is 4; however, no index clearly shows the solution. The F-Measure, Jaccard, Fowlkes–Mallows, and Hubert indices, which indicate the optimal number with maximum values, all have a local maximum not only at 5 but also at 8. Furthermore, the

Table 1
Equations of external clustering validity indices from the literature equations.

Preliminaries	
Total elements in the dataset	n
Elements in cluster i in class j	n_{ij}
Total elements in cluster i	$n_{i.}$
Total elements in class j	$n_{.j}$
Rate of the cell ij	$p_{ij} = \frac{n_{ij}}{n}$
Rate of the row i	$p_{i.} = \frac{n_i}{n}$
Rate of the colum j	$p_j = \frac{n_j}{n}$
Set matching	
Purity [42]	$P = \sum_i p_i (max_j \frac{p_{ij}}{p_i})$
F-Measure [20]	$FM = \sum_j p_j max_i (2 \frac{\frac{p_{ij}}{p_i}\frac{p_{ij}}{p_j}}{\frac{p_{ij}}{p_i}+\frac{p_{ij}}{p_j}})$
Goodman-Kruskal [12]	$GK = \sum_i p_i (1 - max_j \frac{p_{ij}}{p_i})$
Criterion H [25]	$CH = 1 - \frac{1}{n} \max \sum_{i=1}^{k} n_{ij}$
CSI [10]	$CSI = \frac{\sum_{i=1}^{k} n_{ij} + \sum_{i=1}^{k'} n_{ji}}{2n}$
PSI [30]	$PSI = \begin{cases} \frac{S - E(S)}{\max(k,k') - E(S)} & S \geq E(S), \max(k,k') > 1 \\ 0 & S < E(S) \\ 1 & K = K' = 1 \end{cases}$
Pair-counting	
Rand index [29]	$RI = [\binom{n}{2} - \sum_i \binom{n_{i.}}{2} - \sum_j \binom{n_{.j}}{2} + 2\sum_{ij} \binom{n_{ij}}{2}] / \binom{n}{2}$
Adjusted rand index [36]	$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}] / \binom{n}{2}}{\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}] / 2 - [\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}] / \binom{n}{2}}$
Jaccard [33]	$J = \frac{\sum_{ij} \binom{n_{ij}}{2}}{\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2} - \sum_{ij} \binom{n_{ij}}{2}}$
Fowlkes and Mallows [9]	$FM = \frac{\sum_{ij} \binom{n_{ij}}{2}}{\sqrt{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}}$
Hubert Statistic [17]	$H = \frac{\binom{n}{2} \sum_{ij} \binom{n_{ij}}{2} - \sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\sqrt{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} [\binom{n}{2} - \sum_i \binom{n_{i.}}{2}] [\binom{n}{2} - \sum_j \binom{n_{.j}}{2}]}}$
Minkowski Score [3]	$MS = \frac{\sqrt{\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2} - 2\sum_{ij} \binom{n_{ij}}{2}}}{\sqrt{\sum_j \binom{n_{.j}}{2}}}$
Information Theory	
Entropy [42]	$E = -\sum_i p_i (\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i})$
Variation of Information [24]	$VI = -\sum_i p_i \log p_i - \sum_j p_j \log p_j - 2\sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$
Mutual Information [2]	$MI = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$

Table 2
Three different distribution examples with 3 classes (A, B, C) and 3 clusters (1, 2, 3).

(a) Contingency table where chi-squared is 0.				(b) Contingency table where chi-squared reaches its maximum value.				(c) Contingency table in which the distribution of the instances could be found on a real scenario.			
Cluster	A	B	C	Cluster	A	B	C	Cluster	A	B	C
1	2	2	2	1	6	0	0	1	3	3	0
2	1	1	1	2	0	0	3	2	0	3	0
3	3	3	3	3	0	9	0	3	0	0	9

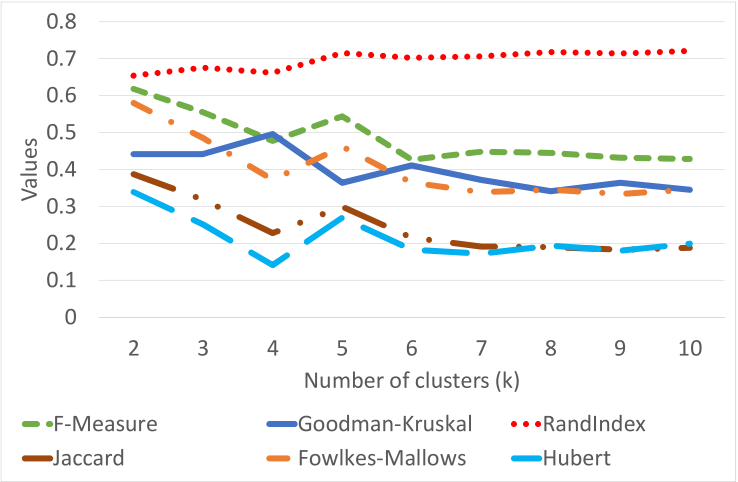


Fig. 2. Results of the CVIs from the literature for $k = 2$ to 10 number of clusters for knowledge dataset whose optimal number of clusters is 4.

Goodman–Kruskal index, which reaches its optimal number of clusters at the minimum value, has a low local minimum at 5 and at 8. Additionally, the Rand Index, following the elbow method, marks the optimal number at 5. In summary, CVIs indices can be misleading due to the interpretation of its results.

In recent years, several studies that propose new external indices for clustering validation have been published in the literature.

A new *pair-counting* index, which is based on an intuitive probabilistic approach, is employed to compare solutions that may have a certain degree of overlap in [34]. This index was tested using four artificial datasets with 6 classes and 4 real datasets from the UCI repository [35].

A new index was also presented in [36], but in this case, it is based on Max-Min distance between data points and prior information. This external index could be classified in the category of *set matching*. The performance of this index was compared with *set matching* and *pair-counting* indices using 6 artificial datasets and two real datasets also from the UCI repository.

The authors of the work presented in [37], proposed a new index based on an ensemble of supervised classifiers. We may classify this index as a *pair-counting* index. Fifty real datasets from the UCI repository were used for the experiments and the results were compared with several internal indices.

A new *pair-counting* index for analytical comparisons was presented in [20]. It applies a correction for chance and normalizes for each cluster separately. The experiments were carried out with artificial datasets with 3 classes and 6000 instances in each dataset. This new index obtained better results than other external CVIs such as purity, adjusted rand index, and mutual information.

In [10], other authors suggested a new *set-matching* index based on the conception of a degree of freedom that measures the decision interval between two classes. This index measures the quality of the clustering by comparing it with internal and *set matching* external indices. Fourteen real datasets were used to test the performance of the index.

Most of these clustering validation techniques are verified by comparing the clustering results with CVIs from the literature and by using synthetic datasets. This work strives to provide a reliable, and accurate CVI based on the chi-squared statistical test as the basis for clustering analysis.

3. Proposed external clustering validity index based on the chi-squared test

3.1. Chi-squared

The Pearson chi-squared statistical test is a method that determines whether there exists a significant difference between the expected values and the observed values in a distribution between two variables [38]. The following equation is applied to verify this correlation:

$$\chi^2 = \sum_{i}^{r} \sum_{j}^{c} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{1}$$

where r is the number of rows, c is the number of columns, n_{ij} is the observed value and E_i is the expected value. E_i is given by

$$E_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n} \tag{2}$$

Table 3
Contingency tables of Table 2c expressed in terms of relative frequencies.

(a) By relative frequencies per row.

#	A	B	C	
1	50%	50%	0%	100%
2	0%	100%	0%	100%
3	0%	0%	100%	100%

(b) By relative frequencies per column.

#	A	B	C
1	100%	50%	0%
2	0%	50%	0%
3	0%	0%	100%
	100%	100%	100%

where n is the total number of instances.

The χ^2 value is employed to determine the suitability of the value through the significant interval. In this way, χ^2 approaches to zero when the observed value resembles the expected value. Therefore, if the observed values are similar to the mean, χ^2 indicates that there is no dependence between the two variables that are being analysed.

3.2. Motivation

External validity clustering indices measure the quality of the clustering result by focusing on a ground truth. Our Chi Index may be considered a set-matching measure since it matches the clusters, and measures the similarity between the clustering and the ground truth, which is given by the maximum value that Chi Index could reach. In addition, the Chi Index is normalized in order to be influenced neither by the number of clusters nor the number of classes. The strategy of the Chi Index is, in general terms, to set the instances of the same class in separate clusters in such a way that the instances which belong to the same class are grouped together as much as possible. In addition, the Chi Index aims to define each cluster by a single class as far as possible. Therefore, the Chi Index looks for the clustering solution that, on the one hand, separates the classes into clusters, and, on the other hand, splits the clusters so that each one can be identified by a class.

The chi-squared test measures the difference between the expected frequencies and the observed frequencies in a distribution. The lower the chi-squared value, the more similar the expected values are to the observed values, that is, if the observed values of the distribution are closer to the mean, then the chi-squared value approaches zero.

Table 2 presents 3 contingency matrices for a distribution with 3 classes (A, B, C) and 3 clusters (1, 2, 3). The values in Table 2a are the same for all the clusters within the classes; in this case, the chi-squared value is 0. The Chi Index seeks exactly the opposite scenario, where the clusters are formed by only one class and where each class is only presented in one cluster, as illustrated in Table 2b. Table 2c presents a distribution where cluster 1 is formed of instances of classes A and B, cluster 2 is composed of instances of only class B, and cluster 3 is consisted of instances from class C.

In order to ensure that each class is only presented in one cluster and each cluster has only one class, the values of the contingency matrix have to be expressed in relative terms. To this end, the absolute frequency contingency table has to be transformed into 2 contingency matrices, one for the relative frequencies per row, and the other for the relative frequencies per column. Hence, in the first contingency matrix, the sum of the rows is 100%, and in the second contingency matrix, the sum of the columns is also 100%.

Taking Table 2c as an example, Table 3a and b are built transforming the absolute frequencies into relative frequencies. As mentioned before, the tables are expressed in relative terms to the total of rows and columns.

In this way, Table 3a indicates that cluster 1 is evenly split between classes A and B, cluster 2 is composed of instances from class 2, and cluster 3 has instances only from class 3. Alternatively, Table 3b shows that the instances from class A are only in the cluster 1, the instances from class B are evenly split between clusters 1 and 2, and the instances from class C are only in cluster 3.

In addition, the Chi Index has an accurate result that needs no interpretation. If we analysed the results for the Chi Index iterating over the number of clusters k, we would obtain two curves, one for each contingency matrix. In general, the clusters tend to become more specialized as the number of clusters increases, that is, there is a higher percentage of points of the same class in each cluster which will increase the chi-squared value for the matrix per row. On the other hand, when the records of each class are distributed across a greater number of clusters, then the value of the chi-square per column will tend to decrease. Our goal is to simultaneously maximize both values by encouraging their tendency to diverge. The first value where both series are cut off (or the distance between them is minimized as we cannot be sure wheter they will be crossed) sets the optimal number of clusters in our proposal. Henceforth, the Chi Index identifies the optimal solution as the minimum difference between the chi-squared values of the curves, thereby rendering it unnecessary to interpret the result thanks to its accuracy.

3.3. Chi Index toy example

Fig. 3 illustrates the spatial distribution of the instances of our toy example dataset with 24 instances and 3 classes. Each dot represents an instance and its colour defines the class to which it belongs.

Before applying a clustering method to this dataset, the number of clusters has to be previously determined. Fig. 4 shows the clustering solution from $k = 2$ to 4. It is difficult to determine which clustering solution is the best at a glance.

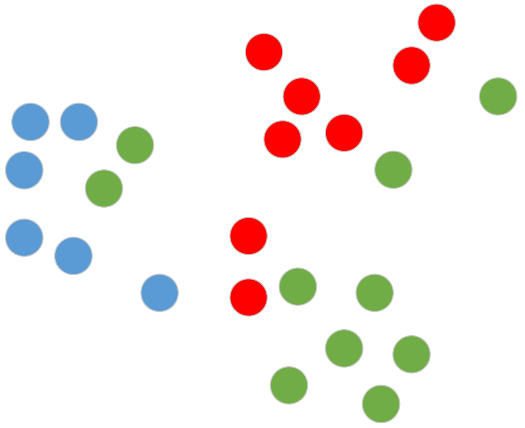
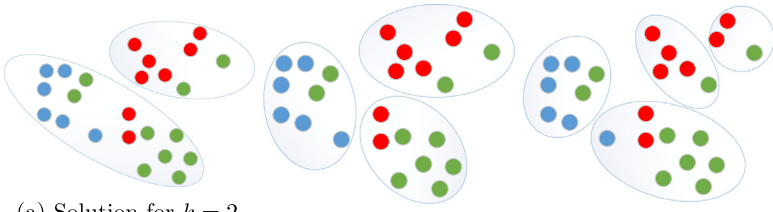


Fig. 3. Representation of the instance distribution of the toy example.



(a) Solution for $k = 2$. (b) Solution for $k = 3$. (c) Solution for $k = 4$.

Fig. 4. Clustering solution representation for $k = 2$ to 4.

To this end, an index that measures the quality of each clustering solution and selects the best one is required. The Chi Index measures the quality of the clustering based on the chi-squared test.

If we focus on the toy example, Fig. 4a represents the clustering solution for $k = 2$. This figure shows that cluster 1 has 2 instances from the red class, 8 green instances, and 6 blue instances, while cluster 2 has 6 red instances, 2 greens instances, and none from blue class. This information is shown in a contingency table in Table 4a, where the clusters are represented by rows, and the classes red, green, and blue are R, G, and B respectively. This table could be analysed in two ways: by rows, where we can conclude that cluster 1 is mainly composed of green instances, but it also has red and blue instances. However, cluster 2 is only composed of red and green instances.; by columns, where blue instances are only in cluster 1, red and green instances are distributed in both clusters.

This analysis is illustrated in Table 4d, where the relative frequency of the instances are expressed in relation to the total of rows (left-side) and columns (right-side).

A complete representation of each clustering solutions from $k = 2$ to 4 is presented in Table 4 with a pair of tables: the contingency table with the absolute frequency, and the contingency tables with relative values by rows and by columns.

Once we have the contingency tables with the relative values, we need to obtain the chi-square value of these tables for each iteration. In our toy example, the Chi Index has been calculated for the clustering solutions with $k = 2$ to 4. The goal is to maximize the values of the Chi Index in both tables and minimize the difference between them. Thus, the Chi Index result will ensure that the observed and expected values differ as much as possible, thereby keeping the solution with the highest percentage of classes in each cluster. Eqs. (3) and (4) detail how the chi square value by row and by column are calculated respectively for $k = 2$.

$$\chi^2_{row_{k=2}} = \frac{\left(13-\frac{88}{2}\right)^2}{\frac{88}{2}} + \frac{\left(50-\frac{75}{2}\right)^2}{\frac{75}{2}} + \frac{\left(37-\frac{37}{2}\right)^2}{\frac{37}{2}} + \frac{\left(75-\frac{88}{2}\right)^2}{\frac{88}{2}} + \frac{\left(25-\frac{75}{2}\right)^2}{\frac{75}{2}} + \frac{\left(0-\frac{37}{2}\right)^2}{\frac{37}{2}} = 89.01 \tag{3}$$

$$\chi^2_{column_{k=2}} = \frac{\left(25-\frac{205}{3}\right)^2}{\frac{205}{3}} + \frac{\left(80-\frac{205}{3}\right)^2}{\frac{205}{3}} + \frac{\left(100-\frac{205}{3}\right)^2}{\frac{205}{3}} + \frac{\left(75-\frac{95}{3}\right)^2}{\frac{95}{3}} + \frac{\left(20-\frac{95}{3}\right)^2}{\frac{95}{3}} + \frac{\left(0-\frac{95}{3}\right)^2}{\frac{95}{3}} = 139.40 \tag{4}$$

Table 5 shows the Chi Index results for our toy example. Chi Index reaches its maximum value at $k = 3$, therefore, we may conclude that the optimal number of clusters that achieved the best clustering solution with this class is with 3 clusters. It should be highlighted that the solution is reached by taking the maximum value of all the solutions because it is the one that achieve the largest value of chi values with both components, and also achieved the minimum difference between them.

Table 4
Toy example contingency tables in which clusters are represented by the rows, and the classes are represented by R (red), G (green), and B(blue). The tables on the left are the contingency tables in absolute values, while tables on the right belongs to the contingency tables with relative values taking as total the sum of the rows (left-side) and the sum of the columns (right-side).

#	R	G	B	
1	2	8	6	16
2	6	2	0	8
	8	10	6	24

(a) $k = 2$.

#	R	G	B	
1	0	2	6	8
2	6	2	0	8
3	2	6	0	8
	8	10	6	24

(b) $k = 3$.

#	R	G	B	
1	0	2	5	7
2	4	1	0	5
3	2	1	0	3
4	2	6	1	9
	8	10	6	24

(c) $k = 4$.

	By row			
#	R	G	B	
1	13%	50%	37%	100%
2	75%	25%	0%	100%
	88%	75%	37%	200%

	By column			
#	R	G	B	
1	25%	80%	100%	205%
2	75%	20%	0%	95%
	100%	100%	100%	300%

(d) Relative contingency tables for $k = 2$.

	By row			
#	R	G	B	
1	0%	25%	75%	100%
2	75%	25%	0%	100%
3	22%	67%	11%	100%
	97%	117%	86%	300%

	By column			
#	R	G	B	
1	0%	20%	100%	120%
2	75%	20%	0%	95%
3	25%	60%	0%	85%
	100%	100%	100%	300%

(e) Relative contingency tables for $k = 3$.

	By row			
#	R	G	B	
1	0%	29%	71%	100%
2	80%	20%	0%	100%
3	67%	33%	0%	100%
4	22%	67%	11%	100%
	169%	149%	82%	400%

	By column			
#	R	G	B	
1	0%	20%	83%	103%
2	50%	10%	0%	60%
3	25%	10%	0%	35%
4	25%	60%	17%	102%
	100%	100%	100%	300%

(f) Relative contingency tables for $k = 4$.

Table 5
Chi index solutions for $k = 2$ to 4.

k	χ^2_{row}	χ^2_{column}	$\chi^2_{row_{\max}}$	$\chi^2_{column_{\max}}$	$Chi\ Index(k)$
2	89.01	139.40	200	300	0.890
3	277.50	299.38	600	600	**0.925**
4	304.05	237.21	800	600	0.760

3.4. Chi Index definition

The Chi Index is defined as:

$$chi\,index(k) = row_{norm}(k) + col_{norm}(k) - |row_{norm}(k) - col_{norm}(k)| \tag{5}$$

where

$$row_{norm}(k) = \frac{\chi^2_{row}(k)}{\chi^2_{row_{max}}} \tag{6}$$

$$col_{norm}(k) = \frac{\chi^2_{column}(k)}{\chi^2_{column_{max}}} \tag{7}$$

$$\chi^2_{row}(k) = \sum_i^r \sum_j^c \frac{\left(\frac{n_{ij}}{n_{i\cdot}} - \frac{N_{\cdot j}}{r}\right)^2}{\frac{N_{\cdot j}}{r}} \tag{8}$$

$$\chi^2_{column}(k) = \sum_i^r \sum_j^c \frac{\left(\frac{n_{ij}}{n_{\cdot j}} - \frac{N_{i\cdot}}{c}\right)^2}{\frac{N_{i\cdot}}{c}} \tag{9}$$

$$N_{i\cdot} = \sum_j^c \frac{n_{ij}}{n_{\cdot j}} \tag{10}$$

$$N_{\cdot j} = \sum_i^r \frac{n_{ij}}{n_{i\cdot}} \tag{11}$$

and n_{ij} is the number of elements from the cluster i in the class j, $n_{i\cdot}$ is the total number of elements in cluster i, $n_{\cdot j}$ corresponds to the total number of elements in class j, and n is the total of elements in the dataset.

$$\chi^2_{row_{max}} = \begin{cases} 100 \cdot r \cdot (r-1) & r \leq c \\ 100 \cdot r \cdot (c-1) & r > c \end{cases} \tag{12}$$

$$\chi^2_{column_{max}} = \begin{cases} 100 \cdot c \cdot (r-1) & r \leq c \\ 100 \cdot c \cdot (c-1) & r > c \end{cases} \tag{13}$$

where r and c are the number of rows and columns respectively.

Chi index takes a value in [0, 2], where 0 is given by the worst clustering solution, and 2 is the best value that Chi Index can achieve. Hence, the optimal k is given by:

$$k^* = \underset{k}{\operatorname{argmax}}\, chi\, index(k) \tag{14}$$

4. Experimental results

This section describes the experimental study carried out with the aim of testing the proposed Chi Index over a variety of artificial datasets, and 47 public datasets in terms of certain benchmark evaluation criteria.

This section is composed of Section 4.1 that includes the experiments with the synthetic datasets. Section 4.2 defines the experimental design. Section 4.3 presents the results of the experiments carried out with the public datasets. Section 4.3.1 includes a statistical analysis to test the effectiveness of our proposed index for the public datasets. Finally, a discussion of the results is included in Section 4.3.2.

4.1. Chi Index validation

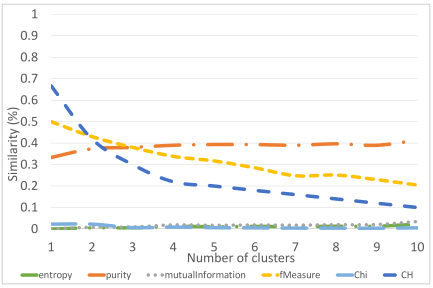
This section includes experimental results for artificial datasets to evaluate the behaviour of Chi Index on diverse clustering solutions based on the work published in [20]. In this case, clustering solutions are generated and compared with the ground truth (G). The results include the 15 CVIs from the state-of-art (Section 2) and our proposed Chi Index. Figs. 5–8 are composed of four subfigures:

- *Subfigure (a)* is a graphic representation of the generated clustering solutions (S1, S2, S3, ...) with G.
- *Subfigures (b,c,d)* are plots of the CVI results for each of the solutions. The y-axis represents the similarity in percentage, while the x-axis depends on a particular feature of each dataset. Detailed explanations are presented in their respective paragraphs.

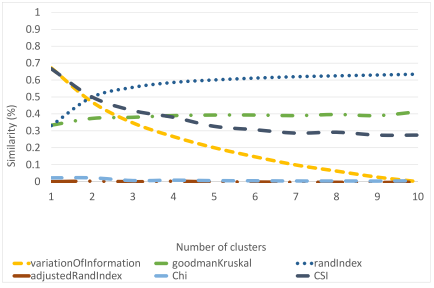
The similarity is defined as the affinity measured with the percentage of a clustering solution S_k compared with the ground truth G. It is expressed in relative terms to the best solution that could be found in the interval of the study. Its value lies in the range [0,1], whereby 0 indicates the worst result, and 1 indicates the solution that perfectly fits G.

Fig. 5 shows the results for clustering solutions with random partitions. The generated solutions go from 1 class up to 10. Fig. 5a shows the representation of G and the different clustering solutions from 1 class (S1) up to 5 classes (S5). In Figs. 5b–d, it is worth noting that the Chi Index, entropy, mutual information, adjusted rand index, Hubert, and PSI had its

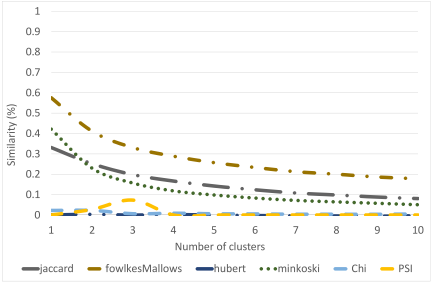
(a) Representation of the ground truth G and the solutions for $k = 1$ to 5 ($S1-S5$).



(b) Solutions for entropy, purity, mutual information, F-measure, CH, and Chi Index.



(c) Solutions for Variation of information, Goodman-Kruskal, Rand Index, Adjusted Rand Index, CSI, and Chi Index.



(d) Solutions for Jaccard, Fowlkes-Mallows, Huber, Minkowski, PSI, and Chi Index.

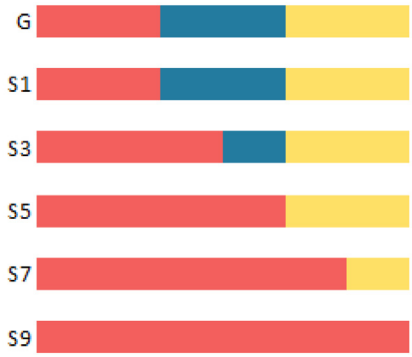
Fig. 5. Results for random generated clustering solution from $k = 1$ to 10 number of clusters.

values at zero. Mutual Information index (Fig. 5d) and the Rand Index (Fig. 5c), could imply that the optimal number is 3 because their curves converge. In addition, PSI had a higher value at 3, that may indicate that is the better solution, but it was with a value under 0.1.

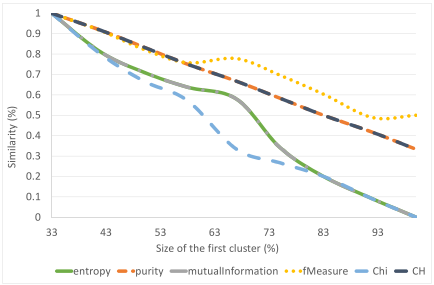
Fig. 6 shows the results for clustering solutions where the instances of the first cluster are increased in each dataset until completion. In Fig. 6a, S1 has the same distribution as G, and hence this is the best solution for all the indices. Figs. 6b–d show the distribution of the CVIs in these datasets. The x-axis represents the percentage of the instances of the first cluster, which ranges from 33% to 100%. It can be observed that all the indices presents a similar behaviour. Their best values are in the dataset that is equal to G and these values decrease until the last dataset whose all instances belong to cluster 1. We find that the Chi Index marks its optimal solution in $S1$ in a similar way than the rest of the indices, but Chi Index descends more linear than the rest of its competitors.

Fig. 7 shows the results for the solutions where the central cluster (in blue) is increased. Fig. 7a shows how the central cluster is increased on each solution where $S1$ is identical to G. The results are similar to the previous ones. Figs. 7b–d show that the indices behave similarly, since the best solution is $S1$, and these indices decrease until the central cluster fills the whole dataset. This result arises from the fact that our index is comparing the distribution of the points across the clusters and, when the dataset is composed of only 1 cluster, the index reaches the lowest value compared with the remaining solutions. We had a comparable situation for the indices of Mutual Information and Entropy (Fig. 7b), Variation of information (Fig. 7c), PSI, and Minkowski (Fig. 7d). It also should be highlighted that Chi Index reached similar results than PSI in this clustering solution.

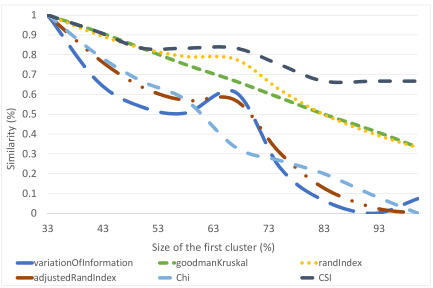
Fig. 8 displays the results of the indices for solutions where the number of incorrect instance labels regularly increases. As seen in Fig. 8a, $S1$ is also identical to G, and it can be observed that on each iteration some of the instances are incorrectly labelled and then this continues until all the instances are incorrectly labelled. Figs. 8b–d show that the Chi Index behaves in a similar way to the rest of the indices during the different datasets. The curves of the indices generally decreases from 1 until 0 in the dataset whose label are 100% incorrect labelled. As it can be seen, the Chi Index and PSI has a near linear since they begin in 1 and decrease to 0. In the case of the F-Measure, the purity, the CSI, and the CH, they start in 1 but they finish at 0.4. The rest of the indices also obtain a similarity of zero in the last dataset but do not describe a near linear behaviour.



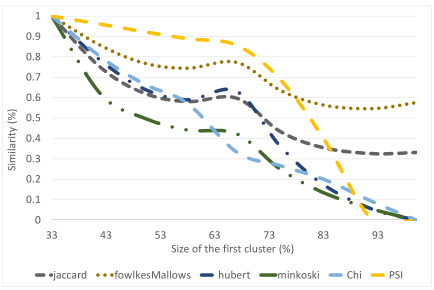
(a) Representation of the ground truth G and the solutions where cluster 1 (in red) is increased until the total of the distribution is reached.



(b) Solutions for entropy, purity, mutual information, F-measure, CH, and Chi Index.



(c) Solutions for Variation of information, Goodman-Kruskal, Rand Index, Adjusted Rand Index, CSI, and Chi Index.



(d) Solutions for Jaccard, Fowlkes-Mallows, Huber, Minkowski, PSI, and Chi Index.

Fig. 6. Results for generated clustering solution where the first cluster increases in each dataset until it fills the whole dataset.

4.2. Experimental design

To generate the clustering solutions, 3 clustering methods from Spark MLlib [19] were applied: k-means, bisecting k-means, and Gaussian mixture.

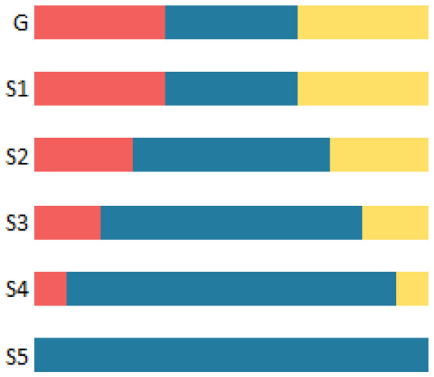
Each dataset, described in Section 4.2.1, was executed with each of these 3 clustering methods. In addition, these clustering methods require the number of clusters (k) into which the dataset is going to be partitioned. The k value was set in the range of $[D_k - 10, D_k + 10]$, where D_k is the correct number of clusters of each dataset and $k > 1$. The number of classes of the datasets was considered as the optimal number of clusters in the same way as carried out in [6,10,20,34,37,39]. With this configuration, we obtained a total of 2820 clustering solutions to test the CVIs. Each clustering solution was compared with the ground truth partition and was then evaluated by the 15 external CVIs described in Section 2. Our new proposed index was also applied in order to compare the results.

4.2.1. Datasets

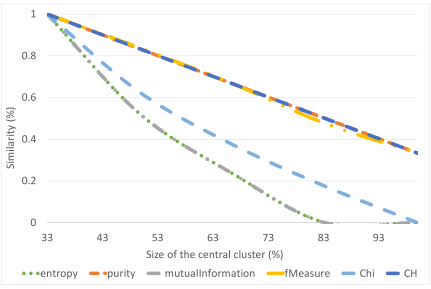
Table 6 presents the datasets used for the experiments and provides the following attributes for each dataset: name; number of classes to be used as the optimal number of clusters; number of features; and the number of instances. All these datasets were downloaded from the UCI machine-learning repository [35]. Note that due to the size of some of the datasets, such as *airlines, higgs, poker*, and *susy*, this could be considered big data. It should be borne in mind that all these datasets included the class information but were not involved in the clustering process. Class information was used in only the clustering analysis stage.

4.2.2. Validity index effectiveness

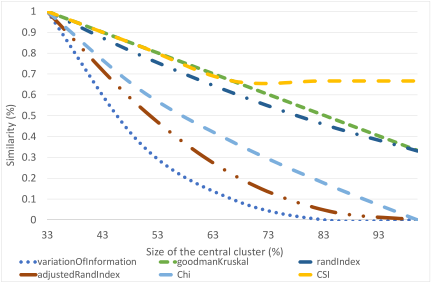
The effectiveness of a CVI measures its capacity to achieve the most coinciding matches while taking a benchmark from different clustering solutions into account. A clustering algorithm and different datasets with a ground truth solution are required in this process. The first step involces applying the clustering algorithm to the datasets and obtaining the multiple solutions. The second step evaluates the solutions with the CVIs. The third step compares the CVI results and selects the one with the highest score.



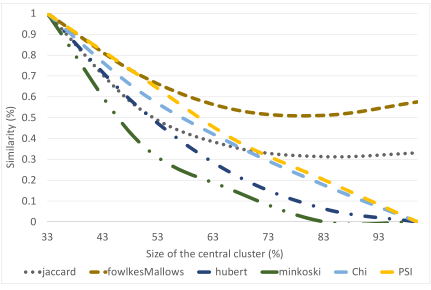
(a) Representation of the ground truth G and the solutions where the central cluster is increased in steps.



(b) Solutions for entropy, purity, mutual information, F-measure, CH, and Chi Index.



(c) Solutions for Variation of information, Goodman-Kruskal, Rand Index, Adjusted Rand Index, CSI, and Chi Index.



(d) Solutions for Jaccard, Fowlkes-Mallows, Huber, Minkowski, PSI, and Chi Index.

Fig. 7. Results for generated clustering solution where the central clusters are increasing step by step until the dataset is completed.

The effectiveness of a CVI depends on how often it takes the correct clustering result in accordance with the chosen criterion. Therefore, the effectiveness is given by counting how many times the index has hit the correct number of clusters. The benchmark employed to make the comparison between the indices includes the following values:

- Average number of hits: this value is given by the mean of the number of times that the index correctly predicted the optimal number of clusters per dataset.
- Average squared error: this is calculated as the average of the squared distances between the prediction of the index I_i and the correct number n_i:

$$Error = \frac{\sum_{i \in n} d(I_i, n_i)^2}{n} \tag{15}$$

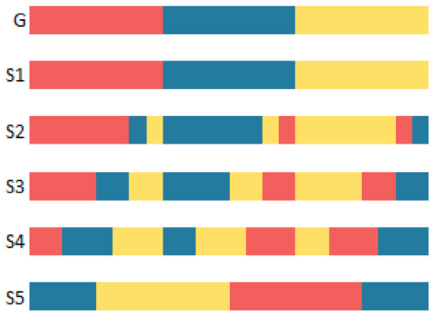
where n is the total number of datasets.

4.2.3. Statistical test

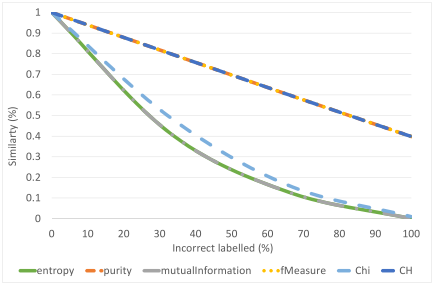
Finally, a statistical framework was applied to test the performance of the indices for the public datasets. The non-parametric Friedman test and Holm post-hoc procedure were chosen to statistically validate the significant differences in the mean ranks of the corresponding p-values reached. This statistical analysis was carried out using the open-source platform StatService [40].

The Friedman test is a non-parametric statistical test that evaluates the differences between more than two related sample means [41]. In our case, the related samples were the CVIs to be compared. The lower the p-value, the better the position in the ranking in the Friedman test.

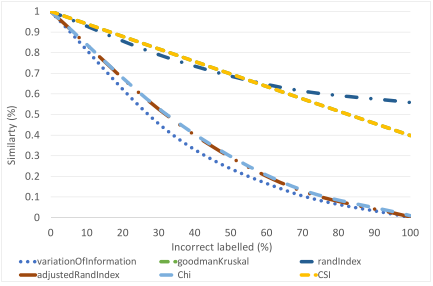
Average ranks for each index provide an objective comparison. The Friedman test could check whether the average ranks were significantly different from the mean rank expected under the null hypothesis. After checking that the measured average ranks are significantly different with an $\alpha = 0.05$, and provided that the Friedman test rejected the null hypothesis, then a post-hoc test could proceed to evaluate the relative performance of the studied CVIs against a control index (that



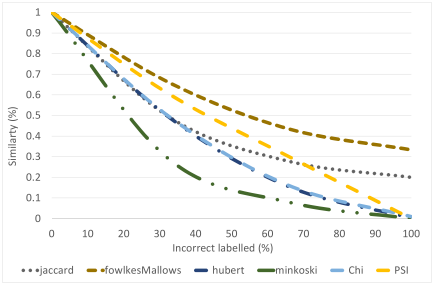
(a) Representation of the ground truth G and the clustering solutions where the number of incorrectly labelled objects are increasing.



(b) Solutions for entropy, purity, mutual information, F-measure, CH, and Chi Index.



(c) Solutions for Variation of information, Goodman-Kruskal, Rand Index, Adjusted Rand Index, CSI, and Chi Index.



(d) Solutions for Jaccard, Fowlkes-Mallows, Huber, Minkowski, PSI, and Chi Index.

Fig. 8. Results for generated clustering solution where the number of incorrectly labelled objects increase proportionally between the clusters.

with the best average rank) thereby avoiding any family-wise errors. This task will be carried out with the Holm step-down procedure by testing hypotheses sequentially ordered in terms of their significance [42].

4.3. Experimental results

This section presents the results obtained with the public datasets. Fig. 9a shows the average number of hits for each CVI in ascending order. It should be highlighted that the Chi Index achieves the highest rate of hits (58%) with a significant margin with its competitors. Indices from the literature had similar rates of hits, ranging from 43% in the case of the F-Measure to 36% for Mutual Information.

On the other hand, Fig. 9b presents the average squared error per index. It is worth noting that the Chi Index obtained the lowest percentage of error. This means that the Chi Index hits the optimal number of clusters most of the times and,when it is in error, it is still not far from the solution.

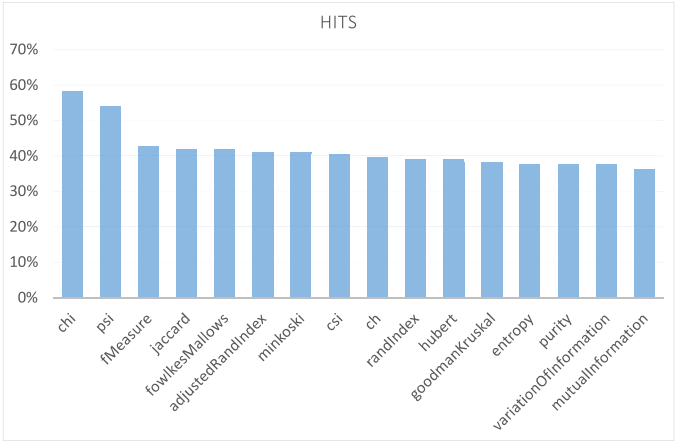
Fig. 10 presents the heatmaps of the distances to the optimal number of clusters of each CVI (rows) for each dataset (columns) represented by the numbers given in Table 6. In these figures, hits are highlighted in green and the farthest results from the solution are graded from white to red. Fig. 10a–c correspond to the results for the k-means, the bisecting k-means and Gaussian mixture methods, respectively.

As can be observed, the Chi Index had a higher rate of green cells than the rest of the CVIs. Although in certain datasets no CVI hit the correct number of clusters, in these cases, the Chi Index remained closer to the solution than its competitors.

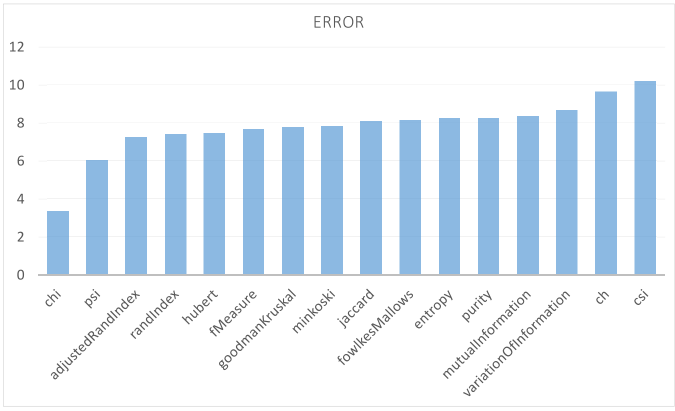
Fig. 11 illustrates the results of Chi Index for two datasets, *zoo* and *knowledge*, whose optimal number of clusters are 7 and 4, respectively. Fig. 11a shows how both curves are crossed at $k = 7$. Moreover, Fig. 11b presents the results for the dataset that has 4 clusters. As can be observed, the curves for the Chi Index by rows and by columns are cut off between $k = 4$ and $k = 5$. These results need no interpretation because the solution is given directly by the index.

4.3.1. Statistical analysis

The Friedman test rankings for every CVI are shown in Table 7a. The ranking was carried out with the results shown in Fig. 10. As previously indicated, the best result for a ranking was 1 and the worst was the last position. As the ranking shows, the Chi Index was in the first position with 6.415. The next index in the ranking was the PSI with a difference of more



(a) Average number of hits by CVI.



(b) Average squared error distance by CVI.

Fig. 9. Benchmark results for the public datasets.

than 1 point with respect to the Chi Index. From this index onwards there are only 0.5 points of difference, and hence, we may conclude that there is a dissimilarity between chi and the indices from the literature. The lowest value for the ranking was 6.415, and the rest ranged from 7.109 to 9.517. Such high values were presented because there were numerous ties in the results, and, in these cases, Friedman establishes the average of the sum of the ranking values of all the competitors. Therefore, for the dataset where all the indices hit the optimal number, Friedman set the ranking at 8.

The statistic for Friedman was 54.694, distributed according to a chi-squared distribution with 15 degrees of freedom. The p-value for Friedman was 0.000, which is lower than 0.05. Therefore, significant differences do exist and it rejected the null hypothesis that they all behaved in a similar way with a level of significance of $\alpha = 0.05$.

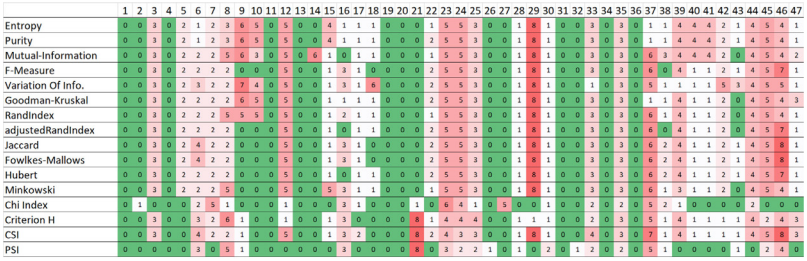
A post-hoc test has been performed in pairs to verify that our proposed Chi Index is significantly different from the other indices.

Table 7b shows the p-values, z-value and α_{Holm}, using the Chi Index as the control method since it obtained the best ranking. As can be observed, the null hypothesis is rejected for all the competitors' CVIs where the $p - value$ remains lower than the α_{Holm}. The null hypothesis was rejected by all the competitors but PSI, whose p-value (0.219) was higher than its α_{Holm} (0.050). Therefore, it may be concluded that the Chi Index generated the best results since it obtained the best average ranking, and that it was significantly different to all the competitors 'CVIs but PSI.

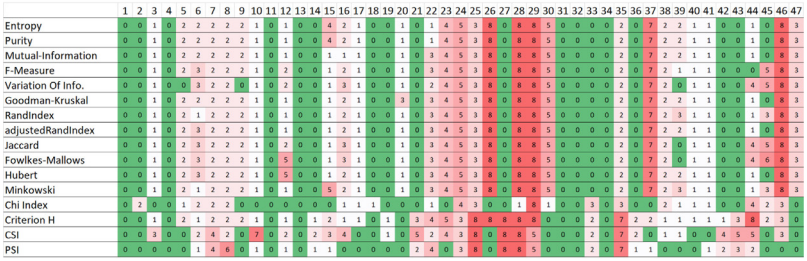
4.3.2. Discussion

The results of the experimental analysis for the public datasets from the UCI repository show that our proposed external index improves the rate of hits by almost 16% (Fig. 9a) with respect to the CVIs from the literature but just 2% from PSI. In addition, in the case of not being able to hit the correct number of clusters, our index obtained a rate of 3 points lower than the CVIs from the literature (Fig. 9b). Chi Index obtained similar rates of hits than PSI, but in case of error, its error is much lower. Our proposed index improves the results based on Friedman's test (Table 7a).

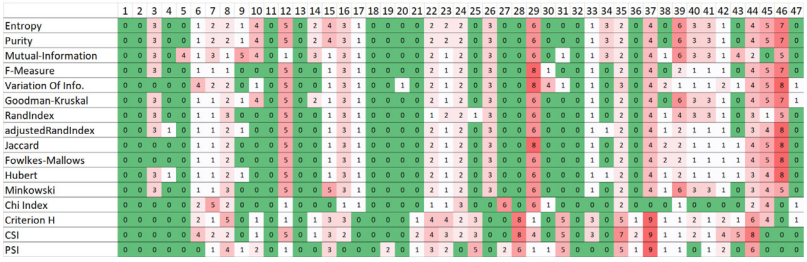
According to the heatmaps from Fig. 10, it can be stated that the Chi Index produced promising results since it hit the optimal number of clusters for most of the datasets and on the according when it failed, its error was not far from the



(a) Heatmap for k-means results.

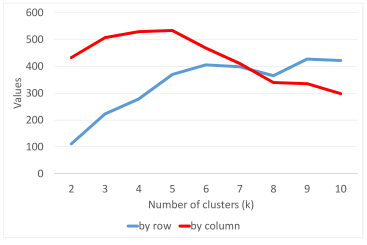


(b) Heatmap for bisecting k-means results.

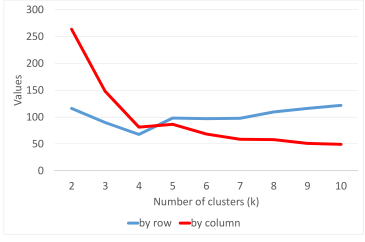


(c) Heatmap for Gaussian mixture results.

Fig. 10. Heatmaps of the distances to the optimal number of clusters of each CVI (rows) for each dataset (columns) represented by the number given in Table 6.



(a) Solution for zoo dataset whose optimal number of clusters is 7.



(b) Solution for knowledge dataset whose optimal number of clusters is 4.

Fig. 11. Representation of the Chi Index for $k = 2$ to 10 for two real datasets.

optimal. It is also interesting to note that there were several datasets in which none of the indices hit the optimal number of clusters. However, in numerous of datasets, it was only the Chi Index which hit the optimal number of clusters.

If we analyse the rate of hits and errors per clustering method, then the Chi Index obtained the best values. For k-means, the Chi Index and the PSI attained 60% hits, and 3.49 and 3.98 points of error respectively. The third in the ranking was the CH index with 49% hits and 5.68 points of error. Bisecting k-means results show that the Chi Index had the highest rate of hits with 64%, while the second mark was obtained by several indices with 49%. The Chi Index had 3.11 points of error and the next in the ranking was the Rand index with 4.32. Finally, the Gaussian mixture had similar results. The PSI index had 53% hits and 8.00 points of error, and in the second position was the Chi Index with 51% hits and 3.53 points of error. K-means and bisecting k-means obtained similar results while Gaussian mixture solutions obtained a lower rate of hits and a higher error.

Table 6
Dataset description.

#	Dataset	Classes	Features	Instances
1	airlines	2	7	539,383
2	bankmarketing	2	16	45,228
3	banknote	2	4	1372
4	biodeg	2	41	1055
5	breast cancer wisconsin	2	9	699
6	breast-tissue	6	9	106
7	car	4	6	1728
8	cloud	4	10	1024
9	column_2C	2	6	310
10	column_3C	3	6	310
11	diabetes	2	20	768
12	ecoli	8	7	336
13	electricity	2	8	45,312
14	faults	2	27	1941
15	forest type mapping	4	27	523
16	gesture phase dataset	5	32	9873
17	glass	6	9	214
18	haberman	2	3	306
19	higgs	2	28	11,000,000
20	iris	3	4	150
21	kddcup99	2	41	494,020
22	knowledge	4	5	403
23	leaf	36	14	340
24	letter	26	16	20,000
25	movement	15	90	360
26	optdigits	10	64	5620
27	ozone	2	72	2534
28	pendigits	10	16	10,992
29	poker	10	10	829,202
30	relax	2	13	182
31	satimage	7	36	6435
32	seeds	3	7	210
33	segment	7	19	2310
34	spambase	2	57	4601
35	spectrometer	4	100	531
36	susy	2	12	5,000,000
37	urban land cover	9	147	675
38	vehicle	4	18	846
39	vowel	11	10	990
40	waveform-1	3	21	5000
41	waveform-2	3	40	5000
42	wholesale	2	7	440
43	wine	3	13	178
44	wine quality red	6	11	1599
45	wine quality white	7	11	4898
46	yeast	10	8	1484
47	zoo	7	17	101

It is also interesting to note that the Chi Index illustrates the optimal clustering solution in an easy and concise way. Some of the solutions of indices in the literature need to be interpreted by following the elbow method or looking for a minimum or a maximum. The Chi Index points out the optimal solution in the intersection of the described curves.

5. Conclusions

In this paper, an innovative external CVI implemented in Spark has been proposed for its application in datasets regardless of their size. The proposed Chi Index is based on the chi-squared statistic test. In addition, we have shown the differences between our proposal and the indices from the literature.

The experimental study indicates that our external index is very competitive. Its effectiveness in public datasets with different sizes has been tested while varying the number of clusters, features, and the number of instances. The main achievements include the following:

- An external CVI based on the chi-squared statistic test is given.
- Our index allowed us to estimate the optimal number of clusters based on the class of the dataset.
- Chi-index results are clear to read and require no further interpretation.
- The proposed index is equipped to work with datasets that may be considered as Big Data.

Table 7
Statistical results.

(a) Sorted mean ranking for Friedman's test.

CVI	Ranking
Chi Index	6.415
PSI	7.109
CH	8.151
Adjusted Rand Index	8.383
F-Measure	8.415
Rand Index	8.489
Minkowski	8.545
Hubert	8.640
Goodman-Kruskal	8.753
Fowlkes-Mallows	8.781
Jaccard	8.799
Entropy	8.936
Mutual Information	8.954
Purity	8.982
Variation of Information	9.123
CSI	9.517

(b) Post-hoc analysis using Holm procedure and the Chi Index as the control index.

CVI	p	z	α_{Holm}
CSI	0.0000	5.490	0.0033
Variation of Information	0.0000	4.792	0.0036
Purity	0.0000	4.543	0.0038
Mutual Information	0.0000	4.493	0.0042
Entropy	0.0000	4.462	0.0045
Jaccard	0.0000	4.219	0.0050
Fowlkes–Mallows	0.0000	4.187	0.0056
Goodman–Kruskal	0.0000	4.137	0.0063
Hubert	0.0001	3.938	0.0071
Minkowski	0.0002	3.770	0.083
Rand Index	0.0002	3.670	0.0100
F-Measure	0.0004	3.539	0.0125
Adjusted Rand Index	0.0005	3.486	0.0167
CH	0.0021	3.486	0.0250
PSI	0.2197	1.227	0.0500

- The size of the dataset does not directly influence the effectiveness of the index.
- The software of this contribution can be found as a spark-package at http://spark-packages.org/package/josemarialuna/ExternalValidity.
- The source code of the Chi Index and the other indices from the literature can be found at https://github.com/josemarialuna/ExternalValidity.

We are currently applying this Chi Index in a clustering analysis with employment data and promising results have been attained. The Chi Index is also being applied on electrical data in collaboration with a Spanish electricity company. As future work, it would be interesting to extend the application of the index to include multi-label datasets.

Acknowledgment

This work has been supported by the Spanish Ministry of Economy and Competitiveness under projects TIN2014-55894-C2-R and TIN2017-88209-C2-2-R. J.M. Luna-Romera holds a FPI scholarship from the Spanish Ministry of Economy and Competitiveness.

References

[1] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (8) (2010) 651–666.
[2] M. Castro-Franco, M. Córdoba, M. Balzarini, J. Costa, A pedometric technique to delimitate soil-specific zones at field scale, Geoderma 322 (2018) 101–111.
[3] R. Davoodi, M. Moradi, Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier, J. Biomed. Inform. 79 (2018) 48–59.
[4] R. Pérez-Chacón, J.M. Luna-Romera, A. Troncoso, F. Martínez-Álvarez, J.C. Riquelme, Big data analytics for discovering electricity consumption patterns in smart cities, Energies 11 (3) (2018).
[5] B. Zhao, J. Wang, Unification of particle velocity distribution functions in gas-solid flow, Chem. Eng. Sci. 177 (2018) 333–339.
[6] J. Rojas-Thomas, M. Santos, M. Mora, New internal index for clustering validation based on graphs, Expert Syst. Appl. 86 (2017) 334–349.
[7] J. Handl, J. Knowles, D.B. Kell, Computational cluster validation in post-genomic data analysis, Bioinformatics 21 (15) (2005) 3201–3212.
[8] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2) (2001) 107–145.
[9] J. Wu, H. Xiong, J. Chen, Adapting the right measures for K-means clustering, in: Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD, ACM, New York, NY, USA, 2009, pp. 877–886.
[10] C. Liu, W. Wang, M. Konan, S. Wang, L. Huang, Y. Tang, X. Zhang, A new validity index of feature subset for evaluating the dimensionality reduction algorithms, Knowl. Based Syst. 121 (2017) 83–98.
[11] S. Jabbar, A.A. Minhas, A. Paul, S. Rho, Multilayer cluster designing algorithm for lifetime improvement of wireless sensor networks, J. Supercomput. 70 (1) (2014) 104–132.
[12] R. Tibshirani, G. Walther, Cluster validation by prediction strength, J. Comput. Graph. Stat. 14 (3) (2005) 511–528.
[13] A. Paul, A. Ahmad, M.M. Rathore, S. Jabbar, Smartbuddy: defining human behaviors using big data analytics in social internet of things, IEEE Wirel. Commun. 23 (5) (2016) 68–74.
[14] V. Berikov, I. Pestunov, Ensemble clustering based on weighted co-association matrices: error bound and convergence properties, Pattern Recognit. 63 (2017) 427–436.
[15] Y. Lei, J.C. Bezdek, S. Romano, N.X. Vinh, J. Chan, J. Bailey, Ground truth bias in external cluster validity indices, Pattern Recognit. 65 (2017) 58–70.
[16] E. López-Rubio, E.J. Palomo, F. Ortega-Zamorano, Unsupervised learning by cluster quality optimization, Inf. Sci. 436–437 (2018) 31–55.
[17] H. Yahyaoui, H.S. Own, Unsupervised clustering of service performance behaviors, Inf. Sci. 422 (2018) 558–571.
[18] Y. Zhang, J. Madziuk, C.H. Quek, B.W. Goh, Curvature-based method for determining the number of clusters, Inf. Sci. 415–416 (2017) 414–428.
[19] A. Spark, Clustering - Spark 2.2.0 Documentation, 2018. https://spark.apache.org/docs/2.2.0/ml-clustering.html, [Online; accessed 6-april-2018].
[20] M. Rezaei, P. Fränti, Set matching measures for external cluster validity, IEEE Trans. Knowl. Data Eng. 28 (8) (2016) 2173–2186.
[21] Y. Zhao, G. Karypis, Criterion functions for document clustering: experiments and analysis, Technical Report, University of Minnesota, Department of Computer Science, Minneapolis, 2001.
[22] B. Larsen, C. Aone, Fast and effective text mining using linear-time document clustering, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD, ACM, New York, NY, USA, 1999, pp. 16–22.

[23] M. Meilă, D. Heckerman, An experimental comparison of model-based clustering methods, Mach. Learn. 42 (1) (2001) 9–29.
[24] P. Fränti, M. Rezaei, Q. Zhao, Centroid index: cluster level similarity measure, Pattern Recognit. 47 (9) (2014) 3034–3045.
[25] L.A. Goodman, W.H. Kruskal, Measures of Association for Cross Classifications, Springer New York, New York, NY, 1971, pp. 2–34.
[26] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (336) (1971) 846–850.
[27] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: is a correction for chance necessary? in: Proceedings of the Twenty Sixth Annual International Conference on Machine Learning, in: ICML, ACM, New York, NY, USA, 2009, pp. 1073–1080.
[28] R. Sokal, P. Sneath, Principles of Numerical Taxonomy, Books in biology, W. H. Freeman, 1963.
[29] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, J. Am. Stat. Assoc. 78 (383) (1983) 553–569.
[30] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1) (1985) 193–218.
[31] A. Ben-Hur, I. Guyon, Detecting stable clusters using principal component analysis, in: M.J. Brownstein, A.B. Khodursky (Eds.), Functional Genomics: Methods and Protocols, Humana Press, Totowa, NJ, 2003, pp. 159–182.
[32] M. Meilă, Comparing clusterings by the variation of information, in: B. Schölkopf, M.K. Warmuth (Eds.), Learning Theory and Kernel Machines, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 173–187.
[33] A. Banerjee, I.S. Dhillon, J. Ghosh, S. Sra, Clustering on the unit hypersphere using von Mises-Fisher distributions, J. Mach. Learn. Res. 6 (2005) 1345–1382.
[34] D. Campo, G. Stegmayer, D. Milone, A new index for clustering validation with overlapped clusters, Expert Syst. Appl. 64 (2016) 549–556.
[35] D. Dheeru, E. Karra Taniskidou, UCI machine learning repository, 2017. http://archive.ics.uci.edu/ml/citation_policy.html.
[36] A.K. Alok, S. Saha, A. Ekbal, A min-max distance based external cluster validity index: MMI, in: Proceedings of the Twelfth International Conference on Hybrid Intelligent Systems (HIS), 2012, pp. 354–359.
[37] J. Rodríguez, M. Medina-Pérez, A. Gutierrez-Rodríguez, R. Monroy, H. Terashima-Marín, Cluster validation using an ensemble of supervised classifiers, Knowl. Based Syst. 145 (2018) 1–14.
[38] P.E. Greenwood, M.S. Nikulin, A guide to chi-squared testing, Wiley-Interscience, New York, NY, 1996.
[39] M.A. Wani, R. Riyaz, A new cluster validity index using maximum cluster spread based compactness measure, Int. J. Intell. Comput. Cybern. 9 (2) (2016) 179–204.
[40] J.A. Parejo, J. Garcia, A. Ruiz-Cortes, J.C. Riquelme, Statservice: Herramienta de análisis estadístico como soporte para la investigación con metaheurísticas, Actas del VIII Congreso Expañol sobre Metaheurísticas, Algoritmos Evolutivos y Bio-inspirados, 2012.
[41] S. Garcia, J. Luengo, F. Herrera, Data Preprocessing in Data Mining, Springer Publishing Company, Incorporated, 2014.
[42] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (2) (1979) 65–70.