Determining the number of clusters in CROKI2 algorithm

Malika Charrad RIADI and CEDRIC research Laboratories
National School of Computer Sciences
malika.charrad@riadi.rnu.tn
Mohamed Ben Ahmed RIADI research Laboratory
National School of Computer Sciences
mohamed.benahmed@riadi.rnu.tn
Yves Lechevallier
INRIA-Rocquencourt 78153 Lechesney cedex
yves.lechevallier@inria.fr
Gilbert Saporta
CEDRIC research Laboratory
Conservatoire National des Arts et Metiers
gilbert.saporta@cnam.fr

***Abstract*—One of the major problems in clustering is the need of specifying the optimal number of clusters in some clustering algorithms. Some block clustering algorithms suffer from the same limitation that the number of clusters needs to be specified by a human user. This problem has been subject of wide research. Numerous indices were proposed in order to find reasonable number of clusters. In this paper, we aim to extend the use of these indices to block clustering algorithms. Therefore, an examination of some indices for determining the number of clusters in CROKI2 algorithm is conducted on synthetic data sets. The purpose of the paper is to test the performance and ability of some indices to detect the proper number of clusters on rows and columns and to compare our new index with some other indexes..**

I. Introduction

Simultaneous clustering, usually designated by biclustering, co-clustering or block clustering, is an important technique in two way data analysis. The term was first introduced by Mirkin [16] (recently by Cheng and Church in gene expression analysis), although the technique was originally introduced much earlier by J.Hartigan [13]. The goal of simultaneous clustering is to find sub-matrices, which are subgroups of rows and subgroups of columns that exhibit a high correlation. A number of algorithms that perform simultaneous clustering on rows and columns of a matrix have been proposed to date. They have practical importance in a wide variety of applications such as biology, data analysis, text mining and web mining. A wide range of different articles were published dealing with different kinds of algorithms and methods of simultaneous clustering. Comparisons of several biclustering algorithms can be found, e.g., in [1], [2], [7] or [10]. One of the major problems of simultaneous clustering algorithms, similarly to the simple clustering algorithms, is that the number of clusters must be supplied as a parameter. To overcome this problem, numerous strategies have been proposed for finding the right number of clusters. However, these strategies can only be applied with one way clustering algorithms and there is a lack of approaches to find the best number of clusters in block clustering algorithms. In this paper, we are interested by the problem of specifying the number of clusters on rows and columns in CROKI2 algorithm proposed in [5]. This paper is organized as follows. In the next section, we present the simultaneous clustering problem. Then in section 4 we present CROKI2 algorithm. In section 5 and section 6, we present a review of approaches based on relative criteria for cluster validity and some clustering validity indices proposed in the literature for evaluating the clustering results. Moreover, an experimental study based on some of these validity indices is presented in section 7 using synthetic data sets.

II. Simultaneous clustering problem

Given the data matrix A, with set of rows $X = (X_1, ..., X_n)$ and set of columns $Y = (Y_1, ..., Y_n)$, a_{ij}, $1 \leq i \leq n$ and $1 \leq j \leq n$ is the value in the data matrix A corresponding to row i and column j. Simultaneous clustering algorithms aim to identify a set of biclusters $B_k(I_k, J_k)$, where I_k is a subset of the rows X and J_k is a subset of the columns Y. I_k rows exhibit similar behavior across J_k columns, or vice versa and every bicluster B_k satisfies some criteria of homogeneity.

	Y_1	...	Y_j	...	Y_m
X_1	a_{11}	...	a_{1j}	...	a_{1m}
...	...	...	...	...	...
X_i	a_{i1}	...	a_{ij}	...	a_{im}
...	...	...	...	...	...
X_n	a_{n1}	...	a_{nj}	...	a_{nm}

TABLE I
DATA MATRIX

III. CROKI2 ALGORITHM

CROKI2 algorithm is applied to contingency tables to identify a row partition P and a column partition Q that maximises Khi2 value of the new matrix obtained by grouping rows and columns. CROKI2 consists in applying K-means algorithm on rows and on columns alternatively to construct a series of couples of partitions (P^n, Q^n) that optimizes Khi2 value of the new data matrix. Given a contingency table $A(X, Y)$, with set of rows X and set of columns Y, the aim of CROKI2 algorithm is to find a row partition $P = (P_1, ..., P_K)$ composed of K clusters and a column partition $Q = (Q_1, ..., Q_L)$ composed of L clusters that maximizes Khi2 value of the new contingency table (P,Q) obtained by regrouping rows and columns in respectively K and L clusters. The new contingency table $T_1(P, Q)$ is defined by this expression:

$$T_1(k, l) = \sum_{i \in P_k} \sum_{j \in Q_l} a_{ij},\ k \in [1, ..., K]$$

and $l \in [1, ..., L]$.
Marginal frequencies in table T1 are :

$f_{kl} = \sum_{i \in P_k} \sum_{j \in Q_l} f_{ij}$
$f_{k.} = \sum_{i \in P_k} f_{i.}$
$f_{.l} = \sum_{j \in Q_l} f_{.j}$

Inputs of CROKI2 algorithm are: contingency table, number of clusters on rows and columns and number of runs.

IV. CLUSTER VALIDATION IN CLUSTERING ALGORITHMS

While clustering algorithms are unsupervised learning processes, users are usually required to set some parameters for these algorithms. These parameters vary from one algorithm to another, but most clustering algorithms require a parameter that either directly or indirectly specifies the number of clusters. This parameter is typically either k, the number of clusters to return, or some other parameter that indirectly controls the number of clusters to return, such as an error threshold. Moreover, even if user has sufficient domain knowledge to know what a good clustering "looks" like, the result of clustering needs to be validated in most applications. The procedure for evaluating the results of a clustering algorithm is known under the term cluster validity. In general terms, there are three approaches to investigate cluster validity [9]. The first one is based on the choice of an external criterion. This implies that the results of a clustering algorithm are evaluated based on a pre-specified structure, which is imposed on a data set and reflects user intuition about the clustering structure of the data set. In other words, the results of classification of input data are compared with the results of classification of data not participating in the basic classification. The second approach is based on the choice of an internal criterion. In this case, only input data is used for the evaluation of classification quality. The internal criteria are based on some metrics which are based on data set and the clustering schema. The main disadvantage of these two methods is their computational complexity. Moreover, the indices related to these approaches aim at measuring the degree to which a data set confirms an a priori specified scheme. The third approach of clustering validity is based on the choice of a relative criterion. Here the basic idea is the comparison of the different clustering methods. One or more clustering algorithms are executed multiple times with different input parameters on the same data set. The aim of the relative criterion is to choose the best clustering schema from the different results. The basis of the comparison is the validity index. Several validity indices have been developed and introduced for each of the above approaches ([8], [9] and [11]). In this paper, we focus only on indices proposed for the third approach.

V. VALIDITY INDICES

In this section some validity indices are introduced. These indices are used for measuring the quality of a clustering result comparing to other ones which were created by other clustering algorithms, or by the same algorithms but using different parameter values. These indices are usually suitable for measuring crisp clustering. Crisp clustering means having non overlapping partitions.

A. Dunn's Validity Index

This index [6] is based on the idea of identifying the cluster sets that are compact and well separated. For any partition of clusters, where C_i represent the cluster i of such partition, the Dunn's validation index, D, could be calculated with the following formula:

$$D = min_{1 \le i \prec j \le K} \frac{d(C_i, C_j)}{max_{1 \le k \le K} d'(C_k)}$$

where K is the number of clusters, $d(C_i, C_j)$ is the distance between clusters C_i and C_j (intercluster distance) and $d'(C_k)$ is the intracluster distance of cluster C_k.In the case of contingency tables, the distance used is Chi-2 distance. The main goal of the measure is to maximise the intercluster distances and minimise the intracluster distances. Therefore, the number of clusters that maximises D is taken as the optimal number of clusters.

B. Davies-Bouldin Validity Index

This index [4] is a function of the ratio of the sum of within-cluster scatter to between-cluster separation.

$$DB = \frac{1}{K} \sum_k max_{k \neq k'} \frac{S_n(c_k) + S_n(c_{k'})}{S(c_k, c_{k'})}$$

where K is the number of clusters, S_n is the average distance of all objects from the cluster C_k to their cluster centre c_k, $S(c_k, c_{k'})$ distance between clusters centres c_k and $c_{k'}$. In the case of contingency tables, the distance used is Chi-2 distance. Hence, the ratio is small if the clusters are compact and far from each other. Consequently, Davies-Bouldin index will have a small value for a good clustering.

C. Xie- Beni Index

It was defined by Xie and Beni [11] to measure the compactness and separation of clusters. Xie-Beni index validity is the combination of two functions. The first calculates the compactness of data in the same cluster and the second computes the separateness of data in different clusters. Let

XB represent the overall validity index, π be the compactness and s be the separation of the c-partition of the data set. The Xie-Beni validity can now be expressed as:

$$XB = \frac{\pi}{s}$$

where

$$\pi = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n}\mu_{ik}^2 d^2(x_i,c_k)}{n}$$

and

$$s = (d_{min})^2$$

K is the number of clusters and d_{min} is the minimum chi-2 distance between cluster centres c_k and $c_{k'}$, given by

$$d_{min} = min_{kk'}d(c_k, c_{k'}).$$

Smaller values of π indicate that the clusters are more compact and larger values of s indicate that the clusters are well separated. Thus, a smaller index reflects that the clusters have greater separation from each other and are more compact. s was originally proposed to identify separation for fuzzy c-partitions, and μ_{ik} is the membership degree criterion of the object x_i in the cluster C_k, we propose:

$$\mu_{ik} = 1 \text{ if } x_i \in C_k$$
$$\mu_{ik} = 0 \text{ if } x_i \notin C_k$$

D. CS Index

CS Index proposed in [3] is a combination between clusters diameters and minimum distance between clusters centers. It is computed as follows:

$$CS = \frac{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{x_i \in C_k} max_{x_j \in C_k} d(x_i,x_j)}{\frac{1}{K}\sum_{k=1}^{K} min_{k,k\neq k'} d(c_k,c_{k'})}$$

where d is a chi-2 distance function, K is the number of clusters and $|C_k|$ is the number of elements in cluster C_k. CS index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. It deals with clusters with different densities and/or sizes and has a small value for a good clustering.

E. Chi2 penalized index

The Chi2 value is another index of good clustering. In fact, CROKI2 tries to find partitions on rows and columns that maximize the Chi2 value of the new matrix generated by new partitions. Therefore, Chi2 value increases when the number of clusters increases. Its maximum value is obtained when each row and each column is affected to a cluster. That is why this index needs to be penalized by KxL, where K is the number of clusters on rows and L is the number of clusters on columns. Chi-2 penalized index could be calculated with the following formula:

$$\text{Chi2P} = \frac{Chi2(P,Q)}{\sqrt{K \times L}}$$

VI. Experimental results

CROKI2 algorithm uses k-means to cluster rows and columns. Therefore, the number of clusters needs to be specified by user. Once CROKI2 is applied to data set, we use all indices presented above to validate clustering alternatively on rows and columns. For our study, we used 3 synthetic two-dimensional data sets further referred to as DataSet1, DataSet2 and DataSet3. DataSet1 is composed of 6000 rows and 4000 columns generated around 5 clusters on rows and 4 clusters in columns. DataSet2 is composed of 4000 rows and 4000 columns generated around 4 clusters on rows and 4 clusters in columns. DataSet3 is also composed of 5000 rows and 4000 columns generated around 6 clusters on rows and 4 clusters in columns (See table II) .

1st	2nd	3rd
6000x4000	5000x4000	5000x4000
(5,4)	(4,4)	(6,4)

TABLE II
DataSets table

Tables below summarize the results of the validity indices (DB, Dunn, CS, Xie-Beni, Chi2 penalized), for different clustering schemes of the above-mentioned data sets as resulting from the simultaneous clustering using CROKI2 with its input value (number of clusters on rows and columns), ranging between 2 and 8. Indices Chi2 penalized, Dunn and CS propose the partitioning of rows of DataSet1 into five clusters and its columns into four clusters which is the correct number of clusters fitting the data set while DB and Xie-Beni indexes select five clusters on rows and five clusters on columns as the best partitioning (See table III).

DataSet 1	1st	2nd	3rd
DB	(5,5)	(5,6)	(5,4)(6,5)
Dunn	(5,4)	(6,4)	(5,5)(6,5)
CS	(5,4)(5,6)	(6,4)(6,6)	(5,5)
Xie-Beni	(5,5)	(5,4)(6,5)	(6,6)
$\frac{Chi2}{\sqrt{K \times L}}$	(5,4)	(6,4)	(5,6)

TABLE III
DataSet 1 biclustering

DataSet 2	1st	2nd	3rd
DB	(4,4) (3,3)	(3,5)	(5,5)(5,4)
Dunn	(4,4) (3,3)	(4,3)	(4,5)
CS	(4,4)	(3,3)(5,4)	(4,5)
Xie-Beni	(4,4)(3,3)	(3,4)(5,4)	(3,5)(4,5)
$\frac{Chi2}{\sqrt{K \times L}}$	(4,4)	(4,5)	(5,4)

TABLE IV
DataSet 2 biclustering

Also, all indices propose four clusters on rows and columns as the best partitioning for DataSet2 (See table IV). In the case of DataSet3, DB and Dunn find the correct number of clusters on rows and columns (i.e. six clusters on rows and 4 clusters

on columns) on the contrary to Chi2 penalized index, CS and Xie-Beni, which propose four clusters on rows and columns as the best partitioning (See table V).

DataSet 3	1st	2nd	3rd
DB	**(6,4)**	(4,4)	(6,5)(5,4)
Dunn	**(6,4)**(4,4)	(4,3)	(4,5)
CS	(4,4)	(5,4)(4,5)	**(6,4)**
Xie-Beni	(4,4)	**(6,4)**	(6,5)(4,6)
$\frac{Chi2}{\sqrt{K \times L}}$	(4,4)	(5,4)	(4,5)

TABLE V
DATASET 3 BICLUSTERING

VII. CONCLUSION

In this paper, we proposed to extend the use of some indices used initially for classic clustering to biclustering algorithms, especially CROKI2 algorithm for contingency tables. Experimental results show that these indices are able to find correct number of clusters when applied with biclustering algorithms or to give an indication of a partitioning that best fits a data set. The Chi2 penalized is also able in some cases to find the best partitioning but it can only be used with CROKI2 algorithm.

REFERENCES

[1] A. Prelic, S. Bleuler, P. Zimmermann, A.Wille, P. Bhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics, 9(22):11221129, 2006.
[2] A. Tanay, R. Sharan, , and R. Shamir. Biclustering algorithms: A survey. In Handbook of Computational Molecular Biology, 2004.
[3] C.-H.Chou, M.-C. Su, and E. Lai. A new cluster validity measure for clusters with different densities. in: IASTED International Conference on Intelligent Systems and Control, pages 276281, 2003.
[4] D. Davies and D. Bouldin. A cluster separation measure. IEEE Trans. Pattern Anal. ,Machine Intell., 4(1):224227, 1979.
[5] G. Govaert. Classification croise. These de doctorat, Universit Pierre et Marie Curie Paris VI.
[6] J. Dunn. Well separated clusters and optimal fuzzy partitions. J.Cybern, (4):95104, 1974.
[7] M. Charrad, Y. Lechevallier, G. Saporta, and M. B. Ahmed. Le bi-partitionnement : Etat de lart sur les approches et les algorithmes. EcolIA08, 2008.
[8] M. Halkidi, M. Vazirgiannis, and I. Batistakis. Quality scheme assessment in the clustering. Process. In Proceedings of PKDD, 2000.
[9] S. Theodoridis and K. Koutroubas. Pattern recognition. Academic Press, 1999.
[10] S. Madeira and A. Oliveira. Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics, (1), 2004.
[11] X. L. Xie and G. Beni. A validity measure for fuzzy clustering. IEEE Trans. Pattern Analysis and Machine Intelligence, 13, 1991.
[12] P. R. Bushel, R. D. Wolfinger, and G. Gibson. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. BMC Systems Biology, 2006.
[13] J. Hartigan. Direct splitting. Chapter 14 in Clustering Algorithms John Wiley and Sons New York, pages 251277, 1975.
[14] E. Saka and O. Nasraoui. Simultaneous clustering and visualization of web usage data using swarm-based intelligence. Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence, 01, 2008.
[15] G. Sanguinetti, J. Laidler, and N. D. Lawrence. Automatic determination of the number of clusters using spectral algorithms. Machine Learning for Signal Processing Proc. 15th IEEE Sig. Proc. Soc. Workshop, pages 5560, 2005.
[16] B. Mirkin. Mathematical classification and clustering. 1996.