Twelfth Mexican International Conference on

Artificial Intelligence

Special Session • Revised Papers
24–30 November 2013
Mexico City, Mexico

MICAI
2013

Edited by
Félix Castro
Alexander Gelbukh
and Miguel González Mendoza

The papers in this book comprise the proceedings of the meeting mentioned on the cover and title page. They reflect the authors' opinions and, in the interests of timely dissemination, are published as presented and without change. Their inclusion in this publication does not necessarily constitute endorsement by the editors, the IEEE Computer Society, or the Institute of Electrical and Electronics Engineers, Inc.

IEEE Computer Society Order Number P2605
ISBN 978-1-4799-2605-3
BMS Part Number CFP1334B-PRT
Library of Congress Number 2013953790

Additional copies may be ordered from:

IEEE Computer Society
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: + 1 800 272 6657
Fax: + 1 714 821 4641
http://computer.org/cspress
csbooks@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: + 1 732 981 0060
Fax: + 1 732 981 9667
http://shop.ieee.org/store/
customer-service@ieee.org

IEEE Computer Society
Asia/Pacific Office
Watanabe Bldg., 1-4-2
Minami-Aoyama
Minato-ku, Tokyo 107-0062
JAPAN
Tel: + 81 3 3408 3118
Fax: + 81 3 3408 3553
tokyo.ofc@computer.org

Individual paper REPRINTS may be ordered at: <reprints@computer.org>

Editorial production by Juan E. Guerrero
Cover art production by Zoey Vegas
Printed in the United States of America by Applied Digital Imaging

CPS
Conference Publishing Services

IEEE Computer Society
***Conference Publishing Services* (CPS)**
http://www.computer.org/cps

2013 12th Mexican International Conference on Artificial Intelligence

MICAI 2013

Table of Contents

Logic, Knowledge-Based Systems and Multi-agent Systems

Robotics, Planning and Scheduling

Evolutionary and Nature-Inspired Metaheuristic Algorithms

Neural Networks and Hybrid Intelligent Systems

Machine Learning and Pattern Recognition

Data Mining

Applicability of cluster validation indexes for large data sets

M. Santibáñez, R. M. Valdovinos, A. Trueba
Universidad Autónoma del Estado de México
UAEM
Facultad de Ingeniería, Cerro de Coatepec s/n Toluca, México
monicass_isc@hotmail.com, li_rmvr@hotmail.com,atruebae@hotmail.com

E. Rendón
Instituto Tecnológico de Toluca
ITT
Av. Tecnológico s/n Fracc. La Virgen, Metepec, México
erendon@ittoluca.edu.mx

R. Alejo, E. López
Tecnológico de Estudios Superiores de Jocotitlan
TESJo
Carretera a Toluca Atlacomulco Km. 44.8 s/n Jocotitlan, México
ralejoll@hotmail.com

***Abstract*—Over time, it has been found there is valuable information within the data sets generated into different areas. These large data sets required to be processed with any data mining technique to get the hidden knowledge inside them. Due to nowadays many of data sets are integrated with a big number of instances and they do not have any information that can describe them, is necessary to use data mining methods such as clustering so it can permit to lump together the data according to its characteristics. Although there are algorithms that have good results with small or medium size data sets, they can provide poor results when they work with large data sets. Due to above mentioned in this paper we propose to use different cluster validation methods to determine clustering quality, as its analysis, so at the same time to determine in an empiric way the more reliable rates for working with large data sets.**

***Keywords*— data mining, clustering, cluster validation, validation indexes**

I. INTRODUCTION

Nowadays, the data, collected from different areas, represents one of the biggest and more interesting information sources, so their processing deliver hidden knowledge to professionals as patterns form, this allows to use it in crucial processes like medicine, financial operations and purchase trends analysis [1].

Data mining and its tools are in charge of processing data for obtaining knowledge that provides a guideline for decision making. These tools are algorithms capable of finding data patterns, either through a training process by which it gets learn to recognize common features among data as from a training sample (offline learning) or by a continuous processing, which data are unknown *a priori* (online learning) [2].

Focusing on the online learning is relevant to try with large data sets generated in real time, it might be possible makes a real time of updating [3], because it depends on the constantly entering patterns to the system and this, according to [4], reduce significantly the processing time.

There are different tasks in data mining that can be performed; one of them is clustering, it refers to data processing that does not have any type of information. For this task is implemented algorithms that can group data according to their similitude, either by determining a radius distance or through the density found among the instances [5]. Clustering is a task that requires a way of measuring grouping quality, if it does not have previous information it has to guarantee that the result is optimum. For this, it has been developed different validation techniques that allow to measure from several points of view the quality and structure of the groups obtained as a result of the clustering task [6]. It is important to analyze the behavior of these techniques in an on-line environment and with large size data sets.

This article is divided into five sections: the first one explains the clustering and validation methods. Next, introduce a review of previous work on this subject; after that, in section three the proposal is described, and in section four and five are shown the experiments and the results, respectively. Finally, conclusions and study open lines are exposed in section six.

II. RELATED WORKS

In literature there are a lot of proposals and works about diversity and usage of indexes of cluster's validation. Some of them are focused in analyzing the origin or constitution of some of these indexes, like the case of [7] who exposes mathematically the origin of the F-Measure. Others analyze one of the most diffused indexes as Rand Index and it's adjustment (Adjusted Rand Index) as it is the case of [8] who propose the ARI as a metric for evaluating the supervised classification and features' selection. The same way [9] exhibits ARI and exemplifies its implementation. Meanwhile [10] proposes a fuzzy extension of Rand Index and from this formulation drifts into other indexes as ARI, Jaccard, Fowlkes and Mallos, among others.

978-1-4799-2605-3/13 $31.00 © 2013 IEEE
DOI 10.1109/MICAI.2013.30

Other authors exhibit validation indexes based on an existing one for evaluating an improvement or a new clustering algorithm or index proposal, it's the case exhibit by [11] who proposes to resume clustering division and refractionation algorithms based on models and analyzes the clusters agreement obtained and original groups comparing results from Fowlkes-Mallows validation index and based, also, in the number of entries in blank from confusion matrix.

On the other hand Xuan [12] analyzes some of the most diffused measurements on information theory and proposes an adjustment based on the forecast proposed by Hubert and Arabie from mutual information and Information variation distance (VI) from which its normalization takes place based on normalization index. This adjustments help to maintain the index close to zero.

In its proposal Vendramin [13] exhibits a methodology for comparing, on an efficient way, 24 existing validation indexes in literature and lot aided, proposes a different analysis from Milligan and Cooper that allows identifying the agreement to the reference group (original group) and, as concordance among results of indexes, cluster's quality. On its side Wanger [6] exhibits a succession of indexes based on three ideas, pair remembering, the sum of overlapping clusters and mutual information. From this division and on, it is explained each of the indexes for, after this, propose a succession of axioms that allows to define if a certain validation index is okay taking into account that it is doing a dynamic clustering.

III. Clustering

The clustering is a data mining task designed for identifying similar features between the data and put similar patterns together in the same group, while among the groups exist a difference between the features of their patterns. This task starts from an unlabeled data set, that is to said, there is no specific information indicating the classification of the patterns and it is expected that real data does not have prior information [14].

There are two principal approaches of clustering and different kinds of algorithms for doing this task. Murty [15] expose a categorization of these types, the first one corresponds to soft clustering which handles the cluster overlapping through fuzzy and genetic techniques and hard clustering where is not permitted the group overlapping. This last approach is divided into hierarchical and partition algorithms.

The hierarchical algorithms are divided into divisive and agglomerative algorithms. The first one use a top-down strategy to divide the data, and the second does the opposite, taking each point as a group and putting together the similar ones to make a big group; both using a tree known as dendrogram. The partition algorithms are based in the square error and mainly in prototypes; this prototypes serve as a reference on how a pattern should look like to belong to a cluster. In this final classification is where the clustering algorithm used here belongs.

A. Clustering algorithm

As seen, there are different kinds of clustering algorithms, but the one analyzed and implemented here is the Batchelor and Wilkins algorithm [16]. For applying the clustering algorithm the number of groups to obtain does not need to be previously known and it just needs one free parameter to calculate dynamically the radius of the groups.

This algorithm takes as a first center any of the objects and takes as a second center the furthest from the first, from these two centers are calculated the groups threshold and select again the furthest object from the two centers, its distance is compared with the nearest center. If the threshold is exceeded a new center is created, if not, then the object is assigned to the nearest center, this is done until there are no more centers. Finally, the assignation of objects is verified with its nearest centers. The main pseudo-code is shown below [16]:

```
Inputs:
  X set de M patterns $\{X_1, X_2, ..., X_M\}$
  $\theta \in [0,1]$ Fraction of the average distance between clusters
Outputs:
  $A \in N$ Number of found clusters
  $S1, S2, ..., S_A$  A pattern sets (clusters)
  $Z_1, Z_2, ..., Z_A$ Cluster centers
Auxiliar variables:
  L Unprocessed patterns
  T Pairs set (nearest pattern-center)
Initialization
  $L \leftarrow X$
  $Z_1 \leftarrow X_1$  $A \leftarrow 1$  $S_1 \leftarrow \{X_1\}$  $L \leftarrow L - \{X_1\}$
  Let $m : \delta(X_m, Z_1) = \max_{X_i \in L}\{\delta(X_i, Z_1)\}$
  $Z_2 \leftarrow X_m$  $A \leftarrow 2$  $S_2 \leftarrow \{X_m\}$  $L \leftarrow L - \{X_m\}$
Do cluster?
  While
    $T \leftarrow \emptyset$
    For each $X \in L$
      $m \leftarrow$ NearestCluster($X$)  {save pair in $T$}
      $T \leftarrow T \cup \{(X, Z_m)\}$  {formed by $X$ y $Z_m$
    End-for // each $X \in L$
    Let $n : \delta(X_n, Z) = \max_{(X,Z) \in T}\{\delta(X, Z)\}$
    threshold $\leftarrow$ CalculateThreshold($\theta$)
    If $\delta(X_n, Z)$>threshold      {Do cluster}
      $Z_{A+1} \leftarrow X_n$  $A \leftarrow A+1$  $S_A \leftarrow \{X_n\}$  $L \leftarrow L - \{X_n\}$
      end $\leftarrow$ FALSE
    If-not // In this case $\delta(X_n, Z) \leq$ threshold
      end $\leftarrow$ TRUE
    End-if // If $\delta(X_n, Z)$>threshold
  Until end = TRUE
Free clustering
  For each $X \in L$
    $m \leftarrow$ NearestCluster ($X$)
    $S_A \leftarrow S_m \cup \{X\}$
    $Z_m \leftarrow$ RecalculateCenter($m$)
  End − for//each $X \in L$
End of algorithm
```

B. Cluster validation

One of the clustering features is that normally does not know the number of groups that will be formed from a data set. To evaluate clustering results in a quantity way, there are cluster validation methods that can help to know the assignation of parameters and restrictions ways that have been established. In general, cluster validation can be shown into three big approaches according to [5]; external criteria, internal criteria and relative criteria. The first criteria correspond to the validation through independence structures, it means, it reflects the intuition that it had about grouping structure, also randomness of them.

The Internal criteria evaluates the quality of clustering based on shape and internal distances, it is no necessary any prior knowledge about the right partition that has to be got, thus it is easier to implement in real tasks. It is important to mention that the cases can be used is when the clustering structure is hierarchical or from just one clustering. The test applications based on internal criteria can be reviewed in detail in [5]. The last criteria value the clustering through the comparison of the algorithms and their parameters, for example, applying different parameters in one algorithm.

According to [17] there is no criteria that ensure 100% the clustering quality because of the origin of the task, there is not any information about the result that should obtain, and it makes task to value the optimum number of clusters and their more complex quality, it mentions that external criteria assumes it has the knowledge about the right partition (ground truth o standard gold) that must get like a result, situation that is less probable in real tasks, but it can perform in experimental environments.

This study applies internal and external validation criteria. For their operation it is recommended to apply the contingency Matrix, this matrix compare the result of both partitions (in this case, real partition and the clustering result), where rows represent a cluster $X=\{C_1, \ldots, CK\}$ and columns the other cluster $P=\{C'_1, \ldots, C'_K\}$, in this case the actual clustering, where C represents a particular cluster. Whereby, each cell contains the number of patterns that both have in common (Fig. 1). While the sum of the columns represents the number of instances that should be assigned to the cluster and the sum of the rows contains the number of actual instances assigned to the cluster.

	C_1	C_2	...	C_K	
C'_1	n_{11}	n_{12}	...	n_{1K}	$n_{1.}$
C'_2	n_{21}	n_{22}	...	n_{1K}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
$C'_{K'}$	$n_{K'1}$	$n_{K'2}$	...	$n_{K'K}$	$n_{K'.}$
	$n_{.1}$	$n_{.2}$	...	$n_{.K}$	$n_{..}=N$

Figure 1. Contingency Matrix for counting-pairs. Extracted from [6].

Counting-pairs. It belongs to the internal criteria. From contingency matrix, can be calculated various index to understand the structure and quality of the clustering got. For this validation method it takes two vectors from the data set, and it compares if both are in the same cluster C and in the same group P, if it's the value is identified by SS (eq. 1). If both selected vectors are in the same cluster C but different group P then the value corresponds to the SD (eq. 2) value. If both vectors are in different cluster C and different group P then the DD (eq. 4) value must be increased. Finally, if both vectors are in different cluster and same group P then the value that should be increased is DS (eq. 3). Following, the equations are shown for these four values from the contingency matrix, and the validation index is listed.

$$SS = \left(\frac{1}{2}\right)\sum_{i=1}^{k}\sum_{j=1}^{k'} n_{ij}^2 - (N/2) \tag{1}$$

$$SD = \left(\frac{1}{2}\right)\sum_{j=1}^{k'} n_j^2 - \left(\frac{1}{2}\right)\sum_{i=1}^{k}\sum_{j=1}^{k'} n_{ij}^2 \tag{2}$$

$$SD = \left(\frac{1}{2}\right)\sum_{j=1}^{k'} n_j^2 - \left(\frac{1}{2}\right)\sum_{i=1}^{k}\sum_{j=1}^{k'} n_{ij}^2 \tag{3}$$

$$DD = \frac{N(N+1)}{2} - \left(\frac{1}{2}\right)\left[\sum_{i=1}^{k} n_i^2 + \sum_{j=1}^{k'} n_j^2\right] \tag{4}$$

Where, n_{ij} corresponds to the number of instances in common between the clusters obtained with the algorithm and the original partitions. The following index can be obtained:

- Rand statistic. (RI, Rand Index) "Represent the fraction of pairs of cases in the same state in both partitions" [18]. i.e. the proportion of patterns equally classified. This index takes value 1 if partitions are identical.

$$R = \frac{(SS + DD)}{(SS + SD + DS + DD)} \tag{5}$$

- Adjusted Rand Index (ARI). It is the adjustment of the previous index and takes into account the hyper geometric space. According to [9] in one Milligan's previous study, ARI is recommended as a good index to determine the similitude between two clustering with different number of groups. This index was created to solve RI differences, they consist of increasing their value to more than one according to the groups quantity increase or the value is not constant. Therefore the ARI result is always between 0 and 1.

$$ARI = \frac{\binom{n}{2}(SS + DD) - [(SS + SD)(SS + DS) + (DS + DD)(SD + DD)]}{\binom{n}{2}^2 - [(SS + SD)(SS + DS) + (DS + DD)(SD + DD)]} \tag{6}$$

Where, $\binom{n}{2}$ is the number of possible pair combinations.

- Jaccard coefficient. Ignores the cases that have been assigned to different clusters in both partitions. This is convenient when there are large clusters quantities and this value can be very high.

$$J = \frac{SS}{(SS + SD + DS)} \tag{7}$$

- Fowlkes y Mallows index. It also ignores the patterns assigned to different groups in both partitions.

$$FM = \frac{SS}{\sqrt{m_1 m_2}} = \sqrt{\frac{SS}{SS + SD}\frac{SS}{SS + DS}} \tag{8}$$

- F measure. It helps to determine how much the clustering resulting groups resemble to those that could have been achieved through manual sorting, therefore requires information on the actual grouping of the cluster [18]. This measure combines the precision and recall measures. Thus we have a set of clusters $C=\{C_1, \ldots, C_k\}$ and the actual classification $C = \{C'_1, \ldots, C'_k\}$, the precision and recall measures are as follows:

$$F_{i,j} = \frac{2}{\frac{1}{prec(i,j)} + \frac{1}{rec(i,j)}} \qquad (9)$$

Where $prec(i, j) = |C_j \cap C'_i|/|C_j|$ y $rec(i, j) = |C_j \cap C'_i|/|C'_i|$

- Calinski-Harabasz (C-H). Also known as Percentage Variation Criterion (VRC). It evaluates the quality of the clusters through the use of variance of the patterns within the cluster and between the clusters [13]. This distance uses the centers [19]. Its performance is obtained comparing the resulting calculation of the index by varying the algorithm parameters and selecting as best grouping that has the highest value, which can be seen as a peak in the resulting graph. Even if the results have a linear trend, up or down then there is no reason to prefer one solution over another.

$$CH(C) = \frac{(N - |P|)inter_{CH}(P)}{(|P| - 1)intra_{CH}(P)} \qquad (10)$$

Where $inter_{CH}(P) = \sum_{C \in P} |C|\, d(C, \bar{X})$ e $intra_{CH}(P) = \sum_{C \in P} \sum_{x \in C} d(x, \bar{C})$

- Davies-Boulin (D-B). It is based on the within-group and between-group ratio to evaluate a particular data partition, that is to say, quantifies the proportion of dispersion [19].

$$DB(C) = \frac{1}{|P|} \sum_{C_k \in P} \max_{C_l \in P/C_k} \left\{ \frac{S(C_k) + S(C_l)}{d(\bar{C}_k, \bar{C}_l)} \right\} \qquad (11)$$

Where, $S(C) = 1/|C| \sum_{x \in C} d(x, \bar{C})$

IV. Set Up

As seen there are many measures that permit to validate the result of clustering task, although according to [6] there is no specific way to compare the clustering result. According to [12] there is no established measure recognized as the best one. Thus, in this article is exposed the using and way of analysis of the clustering results obtained through the implementation of the exposed methodology in [21], it means that corresponds to the online processing of the dataset according to the RAM availability in accordance with their arrival and comparing these results with those which are obtained offline, namely processing the entire data set at a time.

To perform the cluster validation on data sets that have information on their actual grouping, or reference set, raises the using of the counting-pairs index as well as the F-measure. For the data sets lacking of information or reference set, were used the Davies-Boulin and Kalinsky-Harabasz validation index, also the precedent obtained on the data sets that has actual grouping information.

The proposal is to calculate the result of each validation index and taking into account the suggestions for their analysis, comparing these values with each other and with the number of groups known beforehand to determine if the clustering quality is good as well as knowing the most appropriated value of the clustering algorithm parameter to obtain a grouping with quality close to the original. In the situation of those sets that have their actual grouping. Table 1 shows the datasets used for testing. These sets were taken from the repositories of the SIPU (Speech and Image Processing Unit, http://cs.joensuu.fi/sipu/datasets/).

The parameter used for the clustering algorithm, θ, influences the calculation of the threshold to determine when there are new groups, therefore it is important to choose a value that is as optimal as possible to get a quality clustering, for which θ was assigned values of 0.1, 0.2, 0.3, 0.5, 0.7 y 0.9. For cluster validation of artificial dataset D31 was implemented the counting-pairs through contingency matrix to calculate the index obtained from it, which are: RI, ARI, Jaccard index, Fowlkes & Millows index and the F-measure. Likewise shows the results obtained from index C-H and D-H.

V. Experimental Results

A. D31

For data base D31, according to validation indexes, Calinski-Harabasz results are ascending (Fig. 2), which indicates that any value selected is as good as any other. This behavior is almost the same in the rest of the data base excepting that in a D31 with the values θ = 0.5, 0.7 o 0.9 D-B index is higher than C-H, meanwhile the expected ones are a high value from C-H and a value beneath D-B, compared with the rest of the resultant values of θ.

So, data showed next, takes as a reference the value of θ and 0.2 is one of the values that shows a close group, in number as the original, in the case of D31 data base. Another reason for choosing it is that it also has as a result, in the task of classification, a high precision without getting far from the original group. In Table 2 shows the values obtained in the counting-pairs, online and offline.

TABLE I. Data Description

Features	D31	Joensuu	MOPSI
No. Of vectors	3100	6014	8589
No. Of clusters	31	-	-
Dimension	2	2	2
File size	49.5KB	110KB	155KB

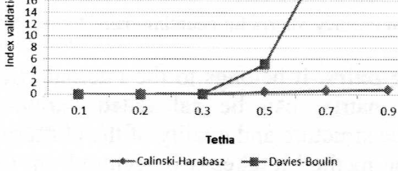


Figure. 1. Cluster validation indexes CH, DB for D31.

TABLE II: VALUE COUNTING-PAIRS OF THE PROCESSING ONLINE AND OFFLINE

	Online	offline
SS	7965.1242	75519
SD	7611.8491	68882
DS	7849.465	77931
DD	469424.39	4581118

In an intuitive way, the *SS* and *DD* high values indicates that exists similarities between divisions obtained with clustering algorithm, which is quite similar to the original, although some objects were integrated in different groups which were assigned in the partition, this is because the number of groups determined by the algorithm is lower than the original division, as it is explained ahead.

It can be observed, in Table 2, that groups formed, both online as offline are well defined because *DD* is higher, although in online processing the SS index is lower than DS, it means that within the groups are assigned patterns from a different one. This may be due to two situations, the first, the influence of θ parameter on the threshold to formed groups, then the greater the value of θ the greater the radius of groups and consequently the number of resulting groups can be different to real number, as in this case, which causes that patterns are assigned to one group that is not their own, according to the real data.

The second situation is that the patterns are so similar, that the algorithm ranks them as part of a different group from the original. This situation is observed in much smaller extent, because setting the θ parameter can be approached, very accurately the original result, this can be seen with the value of $\theta = 0.2$, where the number of groups and their allocation is almost equal to that of the original set, consisting of 31 groups and the closest number the algorithm gets is 37.

From the counting-pairs results were calculated the index in Table 3, which shows the result of the index used to assess the quality of the clustering. To begin, generally can be seen in the Table 3 that results of offline and online processing are quite similar, the difference among them is about one hundredth, besides having the same behavior. This indicates that online processing of the dataset according to RAM available generate a clustering very similar to the offline clustering.

The Rand statistic calculation indicates the fraction from those pairs that are grouped the same way as in the clustering result as in the original partition. While if is closer to one the value of this index, the more will be similar between each other. Rand statistic focuses basically on calculating the coincidence between compared partitions; it is shown in Table 3, that concordance between data set D31 is very high.

TABLE III: RESULTS OF THE VALIDATION INDEXES PROCESSING THE DATA SET D31, OFFLINE AND ONLINE WITH $\Theta = 0.2$

Indexes	Offline	Online
Rand Statistic	0.969435926	0.9685451
ARI	0.491336496	0.4895717
Jaccard index	0.3396677	0.3390697
Fowlkes y Millows	0.50732666	0.5060091
F-measure	0.88135201	0.896825

On the other hand the ARI, which is an adjustment from last index, quantifies the coincidence between compared partitions but takes into account the obtained and expected index, in such a way that if the value is zero, the partitions are independent and approaches to one, are the same. So, the obtained value indicates that the group isn't close to the original, but it has to be taken into account that according to [6] the significance of the measure can be affected by the supposition that is made about the distribution.

Jaccard coefficient doesn't take into account data pair that doesn't match (*DD*), and as a result of this, reflects the proportion of data that has been assigned, so, according to this index, the similarity between partitions is very low. While Fowlkes and Millows index calculus indicates de probability that the elements are assigned the same way in the group and in the real data, thus the obtained value indicates that probability is 50%.

On the other side, F-Measure calculates in a more accurate way how much clusters are similar between each and the original cluster, obtaining a balanced average of the patterns assigned correctly respect to the total of them and those who should have been assigned, thus the result obtained in Table 3 indicates the accuracy of the clustering is good, being the patterns assigned in a correct way in its most.

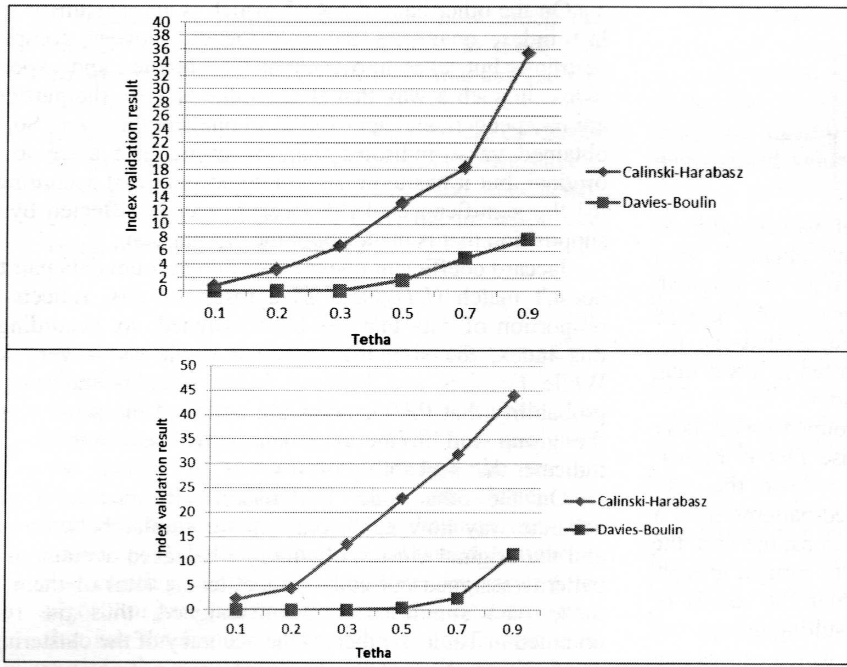
B. *Joensuu and MOPI*

Following will be described the results for the datasets Joensuu and MOPI. For their validation were implemented the Calinski-Harabasz and Davies-Boulin index, from the results of which is determined what is θ more convenient parameter value for processing the data. It is important to point, as it has been mentioned, that in all the results there is a growing tendency to the θ value, in both index, so that is taken as main reference the D-B index score, for which the lower value the better clustering result.

Therefore, as the majority results of D-B, the last three θ values increase significantly with respect to the first three, these last are discarded, so most of the graphics included here are shown just the most relevant area for the analysis, but at the same time shows the behavior of the clustering through the index value. Consequently the optimal values of θ are determined, from which is selected the best based on the C-H result and the number of generated groups with the corresponding θ value.

In the case of the Joensuu and MOPI datasets, and the lack of information about their clustering, their online results are compared with their offline processing results. In Fig 3 can be seen the graphics from the resulting clustering of Joensuu dataset, offline (a) and online (b). According to the graphs, the general behavior in both processing techniques take to the same tendency, even the value of $\theta = 0.2$ the D-B value is still small, the C-H value is significantly higher than with $\theta = 0.1$ and there is more similarity in the number of generated clusters. So the more balanced result of clustering is generated with $\theta = 0.2$.

For MOPI dataset the results are similar to those of Joensuu, as likewise discards the last three values of θ for which the D-B result is bigger than 1. The value of $\theta = 0.3$ is



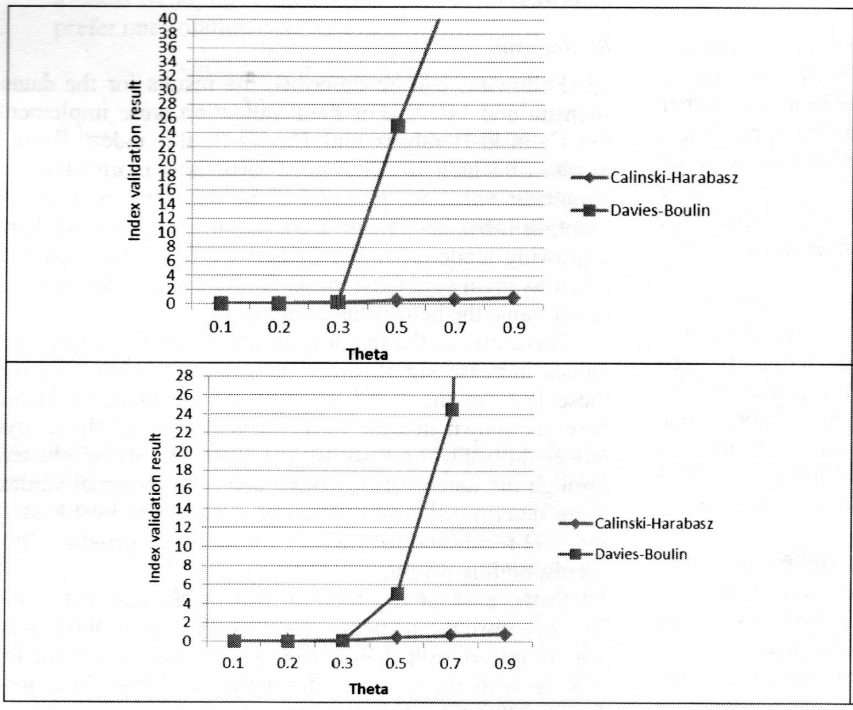
(a)

C-H	D-B	Groups
0.782539156	0.000292808	75
3.145888624	0.006643861	29
6.639860423	0.055440428	15
13.12607999	1.599240769	6
18.37597624	4.948387189	4
35.38254563	7.762014636	3

(b)

C-H	D-B	Groups
2.25388551	0.00058929	57
4.45963377	0.00352498	25
13.5032178	0.03621329	12
22.7958235	0.47361588	5
31.7780149	2.49684401	3
43.939682	11.5414901	2

Figure. 2. Cluster validation indexes CH, DB to Joensuu. (a) offline, (b) online.



(a)

C-H	D-B	Groups
0.019295739	0.000593965	19
0.059647962	0.021311501	11
0.137855536	0.244321076	5
0.538611023	25.09372648	3
0.666525355	46.93172956	2
0.867211814	1068.606773	2

(b)

C-H	D-B	Groups
0.01913797	0.00023392	14
0.07432147	0.00805482	9
0.12976198	0.0851258	4
0.47157364	5.10235998	3
0.66398435	24.5350898	2
0.79557315	190.210248	2

Figure. 3. Cluster validation indexes CH, DB to MOPI. (a) offline, (b) online.

discarded because the prior value is increased considerably, thus the first two θ values could be selected randomly, because, according to the result of index validation, these takes its expected value, but θ = 0.2 generates a number of groups near to offline processing data sets (Fig.4).

Based on the previous analysis of the results is possible to see that no all the validation index quantify the clustering quality in the same way, and this quantity sometimes does not completely describe the clustering quality. Due to mention before, taking into account that the results for D31 dataset are known, could be determined that in the most of the experiments, with all the θ values for this dataset, the F-measure is closer to the right description of the cluster quality, as it was mentioned it considers the number of assigned patterns correctly to respect all these and the ones that really should be assigned.

Furthermore, in the datasets which only knows the number of groups, the Calinski-Harabazs and Davies-Boulin index with the real number of groups can be used to determine which the largest quality clustering, thus is finding the corresponding θ value for that clustering. From both index have a growing trend, the clustering analysis based on both allows finding more reliable clustering of better quality. As for the datasets that lack of any information describing the actual grouping, both indexes are used again to define the quality.

VI. Conclusions

In this article we checked some of the most used cluster validation index to determine the clustering quality. This quality was analyzed based on the result of all the indexes, to know its quality and finally which of them quantifies the actual, being the F-measure the one that offers the most reliable results. The results presented correspond to those obtained from the clustering step of the methodology presented in [21] so detailed presentation and analysis.

The clustering result is quite close to the expected and through this analysis can be seen that for these datasets an optimal θ value can be specified as follows: $0.1 < \theta <= 0.2$. Consequently, for determining the quality of clustering in datasets with information about the real grouping, the F-measure gives the clearest value about the quality of the clustering result. As seen in the case of datasets with lack of information no index is the best, and then both are used.

The Future lines for this issue are oriented to the proposal to implement different validation index and analyze their behavior, as well as work with other datasets to parse through the index an optimal θ value, and compare if is the same as in this paper.

Acknowledgment

We would like to thank to the Instituto Tecnológico de Toluca and also acknowledge the PROMEP/103.5/12/4783 project for their financial support.

References

[1] E. Alppaydin, Introduction to Machin Learning 2nd ed. Cambridge, MA: The MIT Press, 2010.

[2] A. Schwaighofer, J. Q. Candela, T. Borchert, T. Graepel, and R. Herbrich, "Scalable clustering and keyword suggestion for online advertisements," in Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising, ACM, 2009, pp. 27-36.

[3] J. Beringer, E. Hullermeyer, Online clustering of data streams. Technical Report 31, Department of Mathematics and Computer Science. Philipps-University Marburg: Germany, 2003.

[4] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in Proceedings of the 26th International Conference on Machine Learning, 2009.

[5] S. Theodoridis, K. Koutroumbas, Pattern Recognition 4th ed. Estados Unidos: Elseiver Academic Press, 2006.

[6] S. Wagner, and D. Wagner, Comparing clusterings: an overview, Universität Karlsruhe, Fakultät für Informatik, 2007.

[7] Y. Sasaki, Y. "The truth of the F-measure" in Teach Tutor mater, 2007, pp. 1-5.

[8] J. M. Santos, and M. Embrechts, "On the use of the adjusted rand index as a metric for evaluating supervised classification," ICANN, Artificial Neural Networks, ICANN, 2009, pp. 175-184.

[9] K. Y. Yeung, W. L. Ruzzo, "Details of the adjusted rand index and clustering algorithms," in Bioinformatics, vol. XVII, 2001, pp. 763–774.

[10] R. J. Campello, "A fuzzy extension of the rand index and other related indexes for clustering and classification assessment," in Pattern Recognition Letters, vol. XXVIII no. 7, 2007, pp. 833-841.

[11] J. Tantrum, A. Murua, and W. Stuetzle, "Hierarchical model-based clustering of large datasets through fractionation and refractionation," in Information Systems, vol. XXIX no. 4, 2004, pp. 315-326.

[12] V. N. Xuan, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?," in Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009.

[13] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka, On the comparison of relative clustering validity criteria, Sparks, 2009.

[14] J. Hernández, M. J. Ramírez, C. Ferri, Introducción a la Minería de Datos. Madrid: Person Educación, 2004.

[15] M. N. Murty, and V. S. Devi, Pattern recognition: An algorithmic approach. Springer, 2012.

[16] B. F. Cortijo, Técnicas no supervisadas: Métodos de agrupamiento [*online*]. 2001. Disponible en: http://www-etsi2.ugr.es:8080/depar/ccia/RF0708/material.htm.

[17] G. O. Arbelaitz, R. J. Muquerza, "Aportaciones a la clasificación no supervisada y a su validación. Aplicación a la seguridad informática,". Tesis doctoral, Universidad del País Vasco, Facultad de Informática, 2010.

[18] D. A. Ingaramo, M. L. Errecalde, and P. Rosso, "Medidas internas y externas en el agrupamiento de resúmenes científicos de dominios reducidos," in Procesamiento de Lenguaje Natural vol. XXXIX , 2007, pp. 55-62.

[19] J. Lewis, M. Ackerman, and V. De Sa, "Human cluster evaluation and formal quality measures," in Proc. 34th Annual Conference of the Cognitive Science Society, 2012.

[20] A. B. Garay, R. P. Escarcina, and Y. T. Valdivia, "Validación de clusters usando IEKA y SL-SOM".

[21] M. Santibáñez, R. M. Valdovinos E. Rendón, R. Alejo, and J. R. Marcial-Romero, "Optimización de Recurso para el Tratamiento de Grandes Volúmenes de Datos," in Research in Computing Science Avances en Inteligencia Artificial vol. *62*, 2013, pp. 15-24.