BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **16**, No 6
Special issue with selection of extended papers from 6th International Conference on Logistic, Informatics and Service Science LISS'2016

Sofia • 2016 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2016-0091

An Internal Clustering Validation Index for Boolean Data

Liwei Fu, Sen Wu

Donlinks School of Economics and Management, University of Science and Technology Beijing, Beijing, China
Emails: Tavion_Fu@outlook.com wusen@manage.ustb.edu.cn

Abstract: *Internal clustering validation is recognized as one of the vital issues essential to clustering applications, especially when external information is not available. Existing measures have their limitations in different application circumstances. There are still some deficiencies for Internal Validation of Boolean clustering. This paper proposes a new Clustering Validation index based on Type of Attributes for Boolean data* (*CVTAB*)*. It evaluates the clustering quality in the light of Dissimilarity of two clusters for Boolean Data* (*DBD*)*. The attributes in the Boolean Data are categorized into three types: Type A, Type O and Type E representing respectively the attribute values 1,0 and not the same for all the objects in the set. When two clusters are composed into one, DBD applies the numbers of attributes with the types changed and the numbers of objects changed to measure dissimilarity of two clusters. CVTAB evaluates the clustering quality without respect to external information*

Keywords: *Clustering Validation index based on Type of Attributes for Boolean data* (*CVTAB*)*, Dissimilarity for Boolean Data* (*DBD*)*, internal clustering validation index, Boolean data, high dimensional data.*

1. Introduction

1.1. Background

Nowadays, data volume increases explosively along with the computer technology fully integrated into the social life. Internet applications, such as Micro-Blog, Social Network, and e-business, produce a large amount of data particularly in recent years. Data mining is the core of knowledge discovery in databases. Technologically, it is a process, to get implicit pattern from the various, incomplete, fuzzy, and random data. With abundant methods of data acquisition, data normally has two characteristics, high dimensionality and no label. Specifically, a great deal of research work has focused on unsupervised high dimensional data mining. In fact, some clustering algorithms, such as k-means [1, 2], are commonly used in practice.

Cluster analysis, as a main task of data mining, groups objects to clusters, so that objects in the same cluster are more similar to each other than to those in other clusters. It is also a common technique for statistical data analysis used in many fields, such as machine learning, image analysis, pattern recognition, information retrieval, bioinformatics and so on [3]. The result of cluster analysis depends on characteristics of the data set, but no matter what the data distribution pattern is the clustering algorithm can always give a result. So, the index evaluating the quality of clustering result is very significant [4, 5], particularly for high dimensional and large in size data, such as Time-series data [6].

The clustering validation index falls into three types [7]: External Index, Internal Index and Relative Index. External Index focuses on comparing clustering results with the external information; Internal Index focuses on evaluating the goodness of a clustering structure without respect to external information; Relative Index focuses on comparing the results of various algorithms of clustering. Among the 3 of them, only Internal Index can evaluate the clustering results by the interior information of data set without the information outside of the data set such as original category labels. So, Internal Index is more practical [8]. In practice, Internal Index can also be used to select the suitable algorithm and parameter of algorithm objectively.

Categorical data and Boolean data widely exist [9]. The difference between the Boolean data and categorical data is that the attributes of the Boolean data are only 0 and 1, but those of the categorical data are not. The multiple category attributes can be converted into binary attributes by using 0 and 1 to represent either a category absent or present [10]. This paper focuses on the evaluation of Boolean data clustering. It is also applicable to categorical data, which can be transformed into Boolean data harmoniously.

1.2. Related work

Clustering validation measures can be affected by various data characteristics [11], such as data type and noise. Internal Index evaluates the clustering result sensitive to the properties of data set and clusters. Specifically, compactness inside a cluster and separation among the clusters are closely related with Internal Index [12], and many researches focus on it.

For a dataset X, n is the number of objects. The dataset is divided into nc subset; $X=C_1 \cup C_2 \cup \ldots \cup C_{nc}$; c is the centroid of X, and c_i is the centroid of C_i; n_i is the number of objects in C_i; $d\ (x_i, x_j)$ is distance between x_i and x_j. But most of the Internal Indices lack pertinence to Boolean data. Some indices are shown in Table 1.

Table 1. Some of internal indices

Index	Equation of Definition
Calinski-Harabasz index (CH) [13]	$\mathrm{CH} = \dfrac{\frac{1}{nc-1}\sum_{i=1}^{nc} n_i d^2\left(c_i, c\right)}{\frac{1}{n-nc}\sum_{i=1}^{nc}\sum_{x\in C_i} d^2\left(c_i, x\right)}$
Dunn index (Dunn) [14]	$\mathrm{Dunn} = \min\left\{\min \dfrac{\min_{x\in C_i, y\in C_j} d\left(x, y\right)}{\max\left[\max_{x,y\in C_p} d\left(x, y\right)\right]}\right\}$
I index (*I*) [15]	$I = \left[\dfrac{1}{nc}\dfrac{\sum_{i=1}^{nc} d\left(x, c\right)}{\sum_{i=1}^{nc}\sum_{x\in C_i} d\left(c_i, x\right)}\max d\left(c_i, c_j\right)\right]^q$
Silhouette index (*S*) [16]	$S = \dfrac{1}{nc}\sum_{i=1}^{nc}\left\{\dfrac{1}{n_i}\sum_{x\in c_i}\dfrac{b\left(x\right)-a\left(x\right)}{\max\left[a\left(x\right), b\left(x\right)\right]}\right\}$

In Table 1, $a(x) = \dfrac{1}{n_i - 1}\sum_{x,y\in C_i, x\neq y} d(x, y)$, $b(x) = \min_{j, j\neq i}\left[\dfrac{1}{n_j}\sum_{x\in C_i, y\in C_j} d(x, y)\right]$, and $i, j, p = 1, 2, \ldots, nc$; q is a parameter of I index, and $q = 2$ in this paper.

Besides, dissimilarity, as an essential element of the Internal Index, is often used in many algorithms for Boolean data. In k-modes [17], the dissimilarity measure between two objects is defined by the total mismatches of the corresponding attribute. Formally,

$$d\left(X, Y\right) = \sum_{j=1}^{m}\frac{1}{N_j}\delta\left(x_j, y_j\right),$$

$$\delta\left(x_j, y_j\right) == \begin{cases} 0 & \left(x_j = y_j\right), \\ 1 & \left(x_j \neq y_j\right), \end{cases}$$

where X and Y denote the categorical objects, x_j and y_j denote the attributes of X and Y, $j = 1, 2, \ldots, m$; N_j denotes the coefficient. If $d(X, Y)$ gives equal importance to each category of an attribute, $N_j = 1$. But if the frequencies of categories in the data set are taken into account, N_j is half of the harmonic average of n_{xj} and n_{yj}, where n_{xj} and n_{yj} are the numbers of objects in X and Y respectively. The smaller the number of mismatches is, the more similar the two objects are. This measure can only evaluate the dissimilarity of the objects, not that of categories or clusters.

In CABOSFV [18], which is a clustering algorithm for high dimensional sparse data, the measure named Sparse Feature Dissimilarity (SFD) is proposed to represent the dissimilarity of the objects in a set, and it is defined as:

$$\mathrm{SFD} = \frac{e}{|X| \times a},$$

where a denotes the number of attributes that equal 1 for all objects, e denotes the number of attributes that equal 1 for some objects and equal 0 for other objects. $|X|$ indicates the number of objects in set X. The smaller the SFD is, the more similar the objects are. However, SFD can only measure the dissimilarity of the objects in a set or cluster rather than that between two sets or clusters.

Boolean data clustering is impacted negatively by lacking Internal Index which has pertinence to data with binary attributes. This paper proposes a new index to evaluate the effectiveness of the Boolean data clustering.

2. Concepts and definitions

To evaluate the dissimilarity of two clusters, the new index is proposed to measure the validation of clustering. Let A_1, A_2, …, A_m be m attributes of the dataset X, describing a space Ω; Ω is a Boolean space if all A_1, A_2, …, A_m are Boolean. A_j, whose acceptable values are only 1 and 0, is called a Boolean attribute, for j=1, 2, …, m.

2.1. The definition of attribute types

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of n objects with Boolean values and $X \subseteq \Omega$. For the Boolean dataset, the attributes can be categorized into three types. Type A is defined for attributes which values are 1 for all the objects in X; Type O is defined for attributes whose values are all 0; Type E is defined for attributes that have not the same value for all the objects in X. Then the space Ω would be divided into three subsets, J_A, J_O and J_E. For dataset X, if the attribute A_i belongs to Type A attributes, $A_i \in J_\mathrm{A}$; if the attribute A_i belongs to Type O attributes, $A_i \in J_\mathrm{O}$; and if the attribute A_i belongs to Type E attributes, $A_i \in J_\mathrm{E}$, for $1 \le i \le m$. Apparently, set J_E shows the differences and $J_\mathrm{A} \cup J_\mathrm{O}$ shows the similarity.

2.2. Dissimilarity of two Boolean sets

For Boolean data, to compare the dissimilarity of two clusters C_i and C_j, C_i and C_j can merge into one cluster C_U, and $C_U = C_i \cup C_j$. For C_U, J_A, J_O and J_E can be calculated as (1), (2) and (3):

(1) $$J_{\mathrm{A}U} = J_{\mathrm{A}i} \cap J_{\mathrm{A}j},$$

(2) $$J_{\mathrm{O}U} = J_{\mathrm{O}i} \cap J_{\mathrm{O}j},$$

(3) $$J_{\mathrm{E}U} = C_U(J_{\mathrm{A}U} \cup J_{\mathrm{O}U}),$$

where $J_{\mathrm{A}i}$ denotes the set J_A in C_i; $J_{\mathrm{O}i}$ denotes the set J_O in C_i; $J_{\mathrm{A}j}$ denotes the set J_A in C_j; $J_{\mathrm{O}j}$ denotes the set J_O in C_j; $J_{\mathrm{A}U}$ denotes the set J_A in C_U; $J_{\mathrm{O}U}$ denotes the set J_O in C_U; C_U (•) denotes calculating the complementary set. Composing C_i and C_j will result in some attributes of Type A or O altered to E, but Type E attributes in C_i or C_j cannot be changed in C_U obviously. Fig. 1 shows this process.

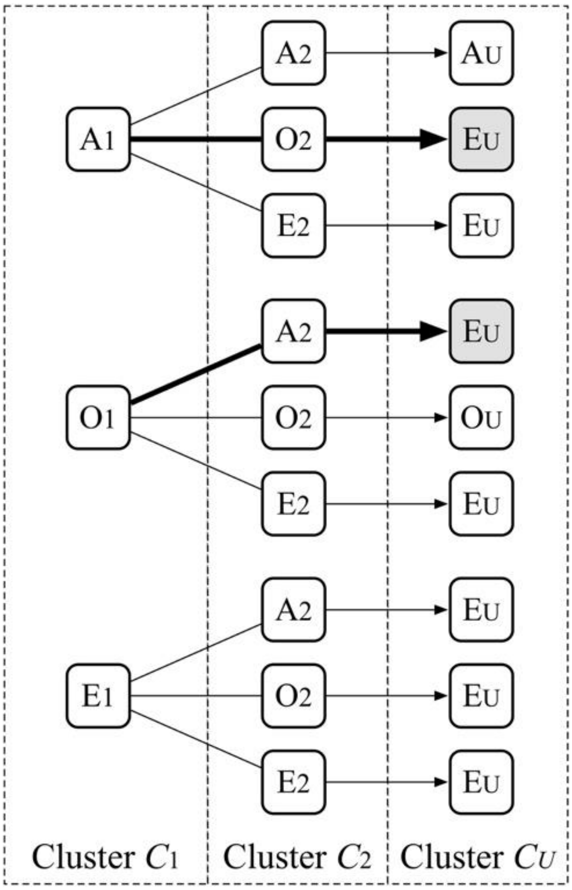


Fig. 1. The altering process of attributes

According to Fig. 1, the attribute of Type A or Type O meeting the attribute of Type E before composing will be altered into Type E in C_U. This means that when two different clusters are put into one, the differences will be more, but similarity will be less without considering the number of the objects in clusters. Therefore, the situation Type A meeting Type O should be focused on, and that is marked by the bold arrows in Fig. 1. and the number of the Type A attributes meeting Type O attributes is used to measure the dissimilarity from the attributes altered as

(4) $$\text{AlterA}_i\text{O}_j = J_{\text{A}i} \cap J_{\text{O}j},$$

where $\text{AlterA}_i\text{O}_j$ denotes attributes altered to E by Type A in C_i meeting Type O in C_j; $J_{\text{A}i}$ denotes the set J_A in C_i; $J_{\text{O}i}$ denotes the set J_O in C_i; $J_{\text{A}j}$ denotes the set J_A in C_j; $J_{\text{O}j}$ denotes the set J_O in C_j.

Furthermore, 1 is usually paid more attention than 0, particularly for sparse data, and the dissimilarity from attributes altering can be indicated by DFAA (Dissimilarity From Altered Attributes) as

(5) $$\mathrm{DFAA} = \frac{n_i}{n_i + n_j}\left|\mathrm{AlterA}_i\mathrm{O}_j\right| + \frac{n_j}{n_i + n_j}\left|\mathrm{AlterA}_j\mathrm{O}_i\right|,$$

where |•| denotes the number of elements in the set; n_i denotes the number of objects in C_i; n_j denotes the number of objects in C_j. According to (5), AlterA$_i$O$_j$ denotes the Type A attributes in C_i altering to Type E by meeting Type O in C_j, and AlterA$_j$O$_i$ denotes that in C_i; $n_i/(n_i + n_j)$ and $n_j/(n_i + n_j)$ are the weights. In symmetric variables, 0 and 1 are equally important. So the weights in (5) will not be taken into consideration, and DFAA is the number of Type A meeting Type O, and it is

(6) $$\mathrm{DFAA} = \left|\mathrm{AlterA}_i\mathrm{O}_j \cup \mathrm{AlterA}_j\mathrm{O}_i\right|$$

On the other hand, the number of objects in C_i and C_j respectively can also effect the dissimilarity of the two clusters. The value of this dissimilarity is the number of objects in C_U minus the harmonic average of those in C_i and C_j, and it is

(7) $$\mathrm{DFN} = n - \frac{2}{1/n_i + 1/n_j},$$

where DFN (Dissimilarity From Number) denotes the dissimilarity from the number of objects in cluster; n denotes the number of objects in C_U, and $n = n_i + n_j$.

Thus the Dissimilarity for Boolean Data (DBD) between C_i and C_j is composed by DFAA and DFN, and it is shown as

(8) $$\mathrm{DBD} = \mathrm{DFAA} \times \mathrm{DFN}.$$

In this paper, DBD (C_i, C_j) indicates the DBD of the cluster C_i and C_j.

2.3. CVTAB for clustering validation

CVTAB (Clustering Validation index based on Type of Attributes for Boolean data) based on DBD can evaluate the validation of clustering for Boolean data. CVTAB is the average of the DBD between each two clusters.

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of n objects and $A_1, A_2, \ldots, A_m$ be m attributes of the dataset, then $X \subseteq \Omega$. After clustering, X is partitioned into k subsets, and $X = C_1 \cup C_2 \cup \ldots \cup C_k$. In C_i, there are n_i objects for $1 \le n_i \le k$, and CVTAB is

(9) $$\mathrm{CVTAB} = \frac{\sum_{i=1}^{k}\sum_{j=1, i\neq j}^{k}\mathrm{DBD}\left(C_i, C_j\right)}{k \times (k-1)},$$

CVTAB is positively associated with the effects of the clustering. The higher CVTAB, the better effects of clustering. Because higher CVTAB denotes more differences between each two clusters and more similarities in each cluster. The bigger value of CVTAB means the better effect of clustering.

3. Steps and examples

Let $X = \{x_1, x_2, \ldots x_7\}$, and X is a Boolean dataset. $x_1, x_2, \ldots, x_7$ are seven objects in a dataset, and $A_1, A_2, \ldots, A_{16}$ are the attributes. Let's assume that $C_1 = \{x_1, x_2\}$, $C_2 = \{x_3, x_4, x_5\}$, $C_3 = \{x_6, x_7\}$. The dataset X is as given in Table 2.

Table 2. The dataset X

Cluster	Object	Sequence of attributes															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C_1	X_1	0	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0
	X_2	0	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0
C_2	X_3	1	1	1	0	0	1	0	0	1	0	0	1	0	0	0	1
	X_4	1	1	0	0	1	1	0	0	1	0	0	1	1	0	0	0
	X_5	1	1	1	1	0	1	0	0	0	0	1	1	0	0	0	0
C_3	X_6	1	0	1	0	1	0	1	1	0	0	1	1	1	0	0	1
	X_7	1	1	0	0	1	0	1	1	1	0	0	1	1	1	0	1

3.1. Example of DBD

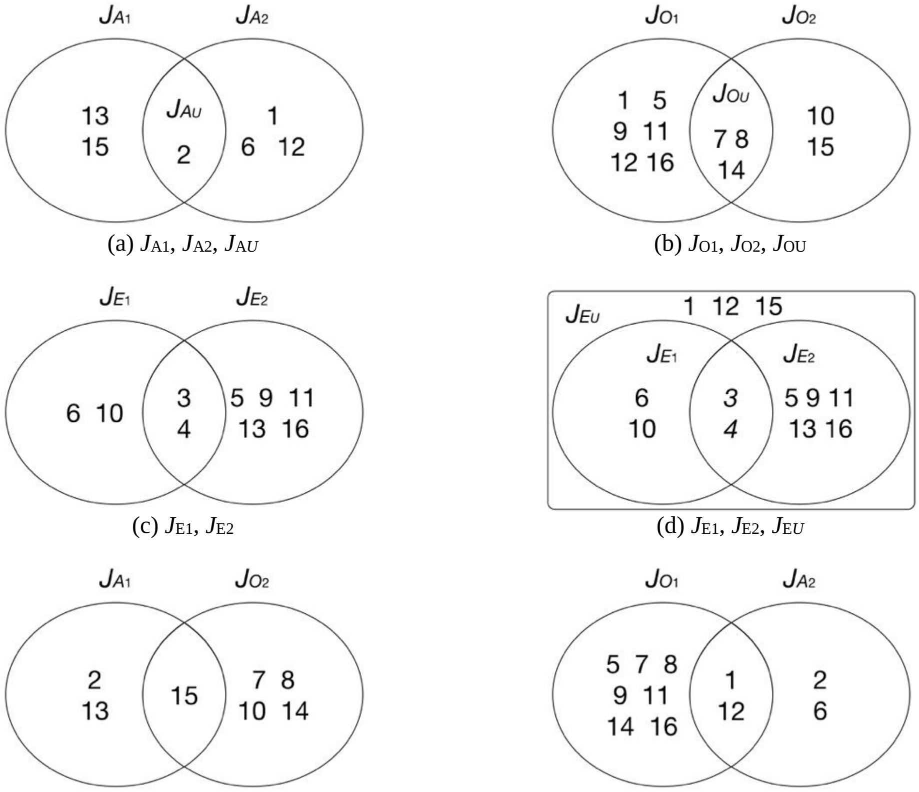
According to Fig. 1, after C_1 and C_2 merged to C_U, the types of attributes are transformed as given in Table 3. Further on, the sets J_A, J_O and J_E of cluster C_1, cluster C_2, and the composed cluster C_U are shown in Table 4. Based on Table 4, the AlterAO, DFAA, DFN, and CVTAB can also be calculated, and this process can be indicated with Venn diagram, shown on Fig. 2.

Table 3. The types of attributes in C_1, C_2 and C_U

Cluster	Sequence of attributes															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C_1	O	A	E	E	O	E	O	O	O	E	O	O	E	O	E	O
C_2	A	A	E	E	E	A	O	O	E	O	E	A	E	O	O	E
C_U	E	A	E	E	E	E	O	O	E	E	E	E	E	O	E	E

Table 4. The sets of J_A, J_O and J_E

Cluster	Type of sets	Sequence of attributes
C_1	$J_{\text{A}1}$	2, 13, 15
	$J_{\text{O}1}$	1, 5, 7, 8, 9, 11, 12, 14, 16
	$J_{\text{E}1}$	3, 4, 6, 10
C_2	$J_{\text{A}2}$	1, 2, 6, 12
	$J_{\text{O}2}$	7, 8, 10, 14, 15
	$J_{\text{E}2}$	3, 4, 5, 9, 11, 13, 16
C_U	$J_{\text{A}U}$	2
	$J_{\text{O}U}$	7, 8, 14
	$J_{\text{E}U}$	1, 3, 4, 5, 6, 9, 10, 11, 12, 13, 15, 16



(a) J_{A1}, J_{A2}, J_{AU}

(b) J_{O1}, J_{O2}, J_{OU}

(c) J_{E1}, J_{E2}

(d) J_{E1}, J_{E2}, J_{EU}

(e) Type A altered to E in J_{A1}

(f) Type A altered to E in J_{A2}

Fig. 2. Venn diagram of composing C_1 and C_2

According to Table 4 and Fig. 2, there are 12 elements in set J_{EU}, and 9 of them are from J_{E1} and J_{E2} which are not transformed, but three of them are from J_{A1} and J_{A2}. Specially, Attribute 15 is in the intersection of J_{A1} and J_{O2}. Attribute 1 and attribute12 are in intersection of J_{O1} and J_{A2}. According to (5), (7) and (8), DBD is calculated as:

$$\left|\mathrm{AlterA_1O_2}\right| = \left|\{15\}\right| = 1,$$

$$\left|\mathrm{AlterA_2O_1}\right| = \left|\{1, 12\}\right| = 2,$$

$$\mathrm{DFAA} = \frac{2}{5}\times 1 + \frac{3}{5}\times 2 = 1.6,$$

$$\mathrm{DFN} = (2+3) - \frac{2}{1/2 + 1/3} = 2.6,$$

$$\mathrm{DBD} = \mathrm{DFAA}\times\mathrm{DFN} = 4.16.$$

In this example, DBD of C_1 and C_2 is 4.16. Similarly, DBD of C_1 and C_3 is 7.00, and DBD of C_2 and C_3 is 3.64.

3.2. Example of CVTAB

As above, DBD of each two clusters can be calculated, and the results will be a symmetric matrix as given in Table 5, and there are 6 values in it. The average of these values is CVTAB. According to (9), the CVTAB is calculated as:

$$\text{CVTAB} = \frac{4.16 + 7.00 + 4.16 + 3.64 + 7.00 + 3.64}{3 \times (3-1)} = 4.9333.$$

Table 5. Symmetric matrix of DBD

Cluster	C_1	C_2	C_3
C_1	–	4.16	7.00
C_2	4.16	–	3.64
C_3	7.00	3.64	–

4. Experiments

4.1. Experiment design

In this experiment, k-modes algorithm on two UCI data sets (Table 6) are implemented in Matlab R2015b to measure the effectiveness of clustering. To compare the effectiveness of various clustering validation measures, the selected data set has external information, original clustering label. For data of zoo, after eliminating the repeated objects, there are 59 objects in it. Another 6 indices Calinski-Harabasz index (CH), Dunn index (D), I index (I), Silhouette index (S), Normalized Mutual Information (NMI) [19], and Accuracy Index(ACC) will also be used as comparative indices to evaluate the validation properties and performances of CATAB. Among these indices, CH, Dunn, I, and S are internal clustering validation measures, focusing on the data set, but NMI and ACC are external clustering validation measures.

Table 6. Data sets for experiments

Data set	Instances	Attributes	Categories
Zoo	101	16	7
Small Soybean	47	35	4

Tests of k-modes would be carried out 100 times to eliminate the effects of randomness, and each time involves different parameters of *k* whose range is 2-11.

4.2. Results and analysis

Table 7 and Table 8 show averages of 100 times clustering results by k-modes. There are some null values in ACC column because ACC cannot evaluate the result while calculated number of clusters is larger than that of actual clusters.

Table 7. Results of data Zoo by k-modes

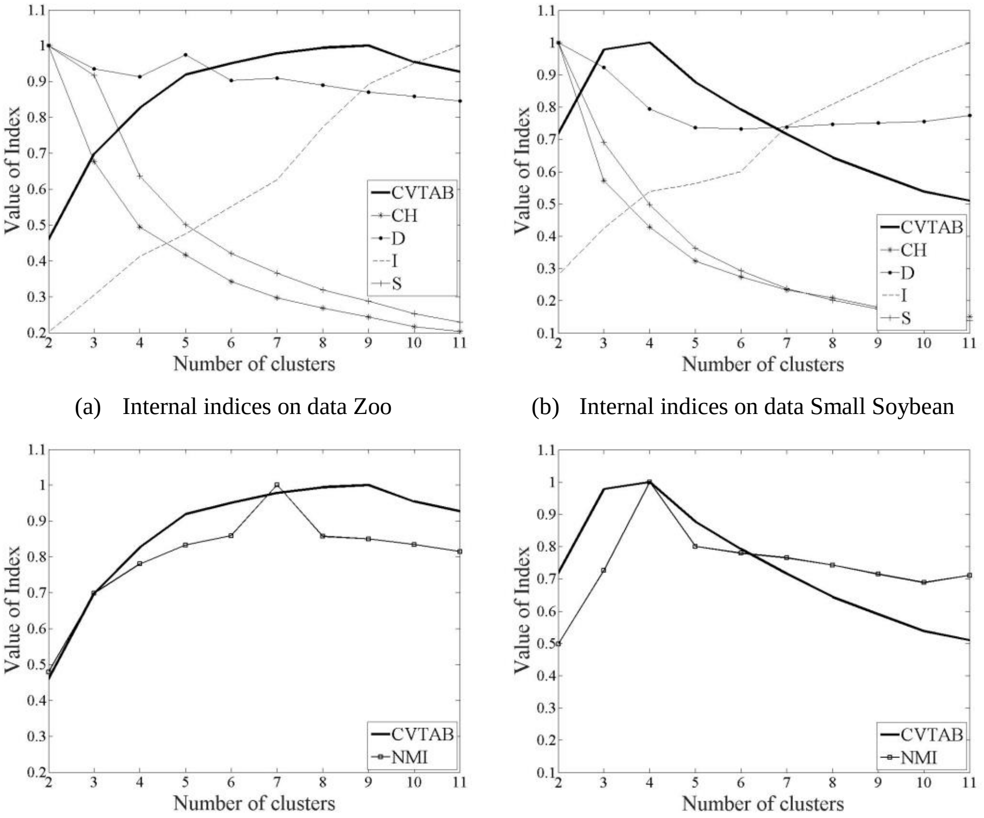
k	CVTAB	CH	D ($\times 10^{-1}$)	S ($\times 10^{-3}$)	I ($\times 10^{-1}$)	NMI	ACC
2	5.99	49.18	4.69	21.75	19.29	0.39	0.51
3	9.52	31.18	4.27	32.18	17.22	0.56	0.65
4	11.14	23.37	4.31	42.94	12.22	0.63	0.71
5	12.47	18.96	4.44	47.70	9.60	0.67	0.72
6	13.41	16.66	4.15	58.57	8.29	0.69	0.75
7	13.60	13.95	4.12	69.23	7.02	0.82	0.75
8	13.90	12.22	4.02	81.43	6.14	0.68	–
9	13.32	11.50	4.04	93.11	5.39	0.68	–
10	13.17	10.21	4.00	105.27	4.96	0.67	–
11	12.92	9.65	4.00	105.98	4.38	0.67	–

Table 8. Results of data Small Soybean by k-modes

k	CVTAB	CH	D ($\times 10^{-1}$)	S ($\times 10^{-2}$)	I ($\times 10^{-1}$)	NMI	ACC
2	32.10	39.04	7.22	2.86	4.61	0.45	0.57
3	43.73	22.34	6.67	4.34	3.18	0.65	0.78
4	44.69	16.73	5.74	5.49	2.30	0.90	0.87
5	39.21	12.63	5.32	5.74	1.67	0.72	–
6	35.39	10.72	5.29	6.13	1.35	0.70	–
7	31.99	9.14	5.33	7.56	1.10	0.69	–
8	28.79	8.14	5.39	8.25	0.93	0.67	–
9	26.38	7.06	5.42	8.94	0.81	0.64	–
10	24.04	6.26	5.45	9.65	0.71	0.62	–
11	22.80	5.91	5.59	10.20	0.64	0.64	–

According to Table 7 and Table 8, the Fig. 3 shows the trends of the results under various numbers of clusters (*k*) to compare these indices after data standardization. ACC is ignored on Fig. 3 considering the null values. In fact, determining the number of clusters is one of the most important clustering validation problems [20].

According to Fig. 3a and b show the results of internal validation indices. CH, *S*, and *I* have a significant monotonic relationship with *k*. Among them, CH and *S* show monotonic increase, but index of *I* shows monotone decrease. This means that obviously, CH, *S*, and *I* are affected by *k* on Boolean data. Because of being sensitive to *k*, they cannot evaluate the results of clustering objectively and validly. As for Dunn index, it shows stationarity both in (a) and (b) for being sensitive to noise. And this will add the uncertainty to evaluation. So, CH, *I*, *S* and *D* are not suitable for Boolean data, and all of them cannot determine the best *k* for k-modes algorithm. However, on data Zoo, CVTAB increases rapidly from *k*=2 to *k*=5, and increases smoothly from *k*=5 to *k*=9, then decreases. On data Small Soybean, CVTAB shoots up until *k*=3, and increases smoothly from *k*=3 to *k*=4, then falls rapidly. In summary, CVTAB has the obvious peak value as the increase of *k* values, and it suggests the recent, even "correct", cluster number on Boolean data, while CH, *I*, *S* and *D* cannot do it.



(a) Internal indices on data Zoo

(b) Internal indices on data Small Soybean

(c) CVTAB and NMI on data Zoo

(d) CATAB and NMI on data Small Soybean

Fig. 3. Clustering results by k-modes

On Fig. 3c and d is shown the trend of CVTAB and NMI as the increase of *k* values. As an internal evaluation index, CVTAB shows consistency with the external clustering validation measures, and the normalized value of CVTAB is near that of NMI. NMI is more accurate than CVTAB on data Zoo, but on data Small Soybean, both of them show the same superiority. Apparently, since NMI is an external validation measure, it knows the "true" cluster number in advance and usually is more precise than an internal measure. However, internal validation measures are the only option for cluster validation without external information, and the conditions when the external information is not available are more common in practice. From this perspective, CVTAB is more applicable.

5. Conclusion

CVTAB is an effective internal clustering validation measure for high dimensional Boolean data. It is also suitable for categorical data which can be translated into Boolean data. Experimental results show that compared with some internal validation indices (CH index, *I* index and *S* index), CVTAB shows consistency with the external clustering validation measure (NMI); CVTAB can point the best clustering result instead of showing monotonicity with the parameter of algorithm. Meanwhile, CVTAB is not as sensitive to noise as Dunn index (internal validation index). Compared with NMI which is external validation index, CVTAB evaluates the clustering validation without external information. From this perspective, CVTAB is more applicable.

In addition, CVTAB can optimize clustering algorithm by determining the parameter of the clustering algorithm, or by selecting optimal results from many experiments to avoid the negative impacts from randomness.

***Acknowledgments*: This work is supported by National Natural Science Foundation of China (NSFC) under Grant No 71271027, and the Research Fund for the Doctoral Program of Higher Education under Grant No 20120006110037 .

References

1. Elangasinghe, M. A., N. Singhal, K. N. Dirks et al. Complex Time Series Analysis of PM 10, and PM 2.5, for a Coastal Site Using Artificial Neural Network Modelling and k-Means Clustering. – Atmospheric Environment, Vol. **94**, 2014, pp. 106-116.
2. Ferrandez, S. M., T. Harbison, T. Weber et al. Optimization of a Truck-Drone in Tandem Delivery Network Using k-Means and Genetic Algorithm. – Journal of Industrial Engineering & Management, Vol. **9**, 2016, No 2, pp. 374-388.
3. Guan, N., D. Tao, Z. Luo et al. NeNMF: An Optimal Gradient Method for Nonnegative Matrix Factorization. – IEEE Transactions on Signal Processing, Vol. **60**, 2012, No 6, pp. 2882-2898.
4. Niennattrakul, V., C. A. Ratanamahatana. On Clustering Multimedia Time Series Data Using k-Means and Dynamic Time Warping. – International Conference on Multimedia and Ubiquitous Engineering, IEEE, 2007, pp. 733-738.
5. Niennattrakul, V., C. A. Ratanamahatana. On Clustering Multimedia Time Series Data Using k-Means and Dynamic Time Warping. – International Conference on Multimedia and Ubiquitous Engineering, IEEE, 2007, pp. 733-738.
6. Rani, S., G. Sikka. Recent Techniques of Clustering of Time Series Data: A Survey. – International Journal of Computer Applications, Vol. **52**, 2012, No 15, pp. 1-9.
7. Liu, Y. Research on Internal Clustering Validation Measures. University of Science and Technology Beijing, 2012, pp. 16-20.
8. Kremer, H., P. Kranen, T. Jansen et al. An Effective Evaluation Measure for Clustering on Evolving Data Streams. – In: Proc. of 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 868-876.
9. Feng, X., S. Wu, Y. Liu. Imputing Missing Values for Mixed Numeric and Categorical Attributes Based on Incomplete Data Hierarchical Clustering. – In: Proc. of International Conference on Knowledge Science, Engineering and Management, Springer Verlag, 2011, pp. 414-424.
10. Ralambondrainy, H. A Conceptual Version of the k-Means Algorithm. – Pattern Recognition Letters, Vol. **16**, 1995, No 11, pp. 1147-1157.
11. Liu, Y., Z. Li, H. Xiong et al. Understanding and Enhancement of Internal Clustering Validation Measures. – IEEE Transactions on Systems Man & Cybernetics Part B. Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, Vol. **43**, 2012, No 3, pp. 982-994.
12. Kraus, J. M., C. Müssel, G. Palm et al. Multi-Objective Selection for Collecting Cluster Alternatives. – Computational Statistics, Vol. **26**, 2011, No 2, pp. 341-353.
13. Zhang, G. X., L. Q. Pan. School of Electrical Engineering, University S. J., Chengdu. A Survey of Membrane Computing as a New Branch of Natural Computing. – Chinese Journal of Computers, Vol. **33**, 2010, No 2, pp. 208-214.

14. Busi, N. Using Well-Structured Transition Systems to Decide Divergence for Catalytic P Systems. – Theoretical Computer Science, Vol. **372**, 2007, No 2-3, pp. 125-135.
15. Nishida, T. Y. An Approximate Algorithm for NP-Complete Optimization Problems Exploiting P Systems. – In: Proc. of 8th World Multi-Conference on Systems, Cybernetics and Information, 2004, pp. 109-112.
16. Huang, L. Research on Membrane Computing Optimization Methods. – Zhejiang University, 2007.
17. Huang, Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. – Research Issues on Data Mining & Knowledge Discovery, 1998, pp. 1-8.
18. Wu, S., X. Gao. CABOSFV Algorithm for High Dimensional Sparse Data Clustering. – Journal of University Science & Technology Beijing, Vol. **11**, 2004, No 3, pp. 283-288.
19. Knops, Z. F., J. B. Maintz, M. A. Viergever et al. Normalized Mutual Information Based Registration Using k-Means Clustering and Shading Correction. – Medical Image Analysis, Vol. **10**, 2006, No 3, pp. 432-439.
20. Chen, L. F., Q. S. Jiang, S. R. Wang. A Hierarchical Method for Determining the Number of Clusters. – Journal of Software, Vol. **19**, 2008, No 1, pp. 62-72.