IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 18, 2017, accepted August 14, 2017, date of publication September 18, 2017,
date of current version November 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2743985

A Clustering Validity Index Based on Pairing Frequency

HONGYAN CUI[1,2,3,5], (Senior Member, IEEE), KUO ZHANG[1,2,3], YAJUN FANG[7], STANISLAV SOBOLEVSKY[4,5,6], CARLO RATTI[5], (Fellow, IEEE), AND BERTHOLD K. P. HORN[7]

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Beijing Laboratory of Advanced Information Networks, Beijing 100876, China
[3]Key Laboratory of Network System Architecture and Convergence, Beijing 100876, China
[4]Center for Urban Science and Progress, New York University, Brooklyn, NY 10003 USA
[5]Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[6]Institute of Design and Urban Studies of the National Research University ITMO, 197101 Saint-Petersburg, Russia
[7]CSAIL Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Corresponding author: Hongyan Cui (yan555cui@163.com)

This work was supported in part by the National Natural Science Foundation of China under 61201153, in part by the National 973 Program of China under Grant 2012CB315805, and in part by the National Key Science and Technology Projects under Grant 2010ZX03004-002-02.

ABSTRACT Clustering is an important problem, which has been applied in many research areas. However, there is a large variety of clustering algorithms and each could produce quite different results depending on the choice of algorithm and input parameters, so how to evaluate clustering quality and find out the optimal clustering algorithm is important. Various clustering validity indices are proposed under this background. Traditional clustering validity indices can be divided into two categories: internal and external. The former is mostly based on compactness and separation of data points, which is measured by the distance between clusters' centroids, ignoring the shape and density of clusters. The latter needs external information, which is unavailable in most cases. In this paper, we propose a new clustering validity index for both fuzzy and hard clustering algorithms. Our new index uses pairwise pattern information from a certain number of interrelated clustering results, which focus more on logical reasoning than geometrical features. The proposed index overcomes some shortcomings of traditional indices. Experiments show that the proposed index performs better compared with traditional indices on the artificial and real datasets. Furthermore, we applied the proposed method to solve two existing problems in telecommunication fields. One is to cluster serving GPRS support nodes in the city Chongqing based on service characteristics, the other is to analyze users' preference.

INDEX TERMS Pairwise pattern, clustering validity, clustering analysis, fuzzy c-means.

I. INTRODUCTION

In hard clustering, each data point is assigned to exactly one cluster. A well-known example of hard clustering is k-means algorithm [1]. In fuzzy clustering, each of the data points can belong to multiple clusters. Fuzzy clusters can be easily converted to hard clusters by assigning the data point to the cluster with greatest probability. The most widely used fuzzy clustering algorithm is fuzzy c-means (FCM) [2]–[4]. FCM selects the centroid of each initial cluster randomly and computes initial fuzzy membership matrix, then tries to iteratively minimize an objective function until the algorithm converges. We will use FCM to generate flexible partitions in this paper.

Result from different clustering algorithms or even the same algorithm can be very different from each other on the same dataset, because the input parameters, which greatly decide the behavior of an algorithm, could be varied. The aim of CV is to find the partition result that best fits the input dataset. With the help of CV, parameters needed for the algorithm can be tuned more efficiently. For example, the number of clusters for a clustering process (represented by c) usually needs to be specified in advance, however people often do not have any specific criteria for choosing it, instead, they usually make an arbitrary choice based on common sense. In the proposed approach all the clustering parameters and results are evaluated by a clustering validity criterion, then the partition that best fits the dataset is produced as well as the corresponding value of c [5]–[7].

Clustering validity techniques are classified into two categories: external validation and internal validation.

External validation evaluates a partition by comparing it with the assumed correct partition result, while internal validation evaluates a partition by examining just the result. Obviously, the former one can only be applied in some limited scenarios, since in a real application the underlying structure of the dataset is unknown, and the ground truth is correct partition result is not available [8]. This paper will focus our research focuses on internal validation which often measures the compactness and separation of clusters.

Many internal indices have been proposed for clustering validation over the past few decades [9], [10]. In the following part of this paper we will call these Clustering Validity Indices (CVIs). Here are some typical CVIs for fuzzy clustering: Bezdek's partition coeffcient (PC) [11] and partition entropy (PE) [12], XB [13], FS [14], the fuzzy hypervolume validity (FHV) proposed by Gath and Geve [15], SC [16], PBM index [17], [18], Wu and Yang's PCAES index [19], Zhang's VW index [20], etc. In addition, CVIs for crisp partitions, like Dunn [2], Davies and Bouldin [21] are widely used. In common sense, a good clustering result should resemble high compactness and significance separation. However, those CVIs suffer several inherent shortcomings. For example, variance is a common measure of compactness, and it tends to prefer hyperspherical-shaped clusters. Another problem of the compactness measure is that these indices tend to monotonically decrease when the number of clusters tends toward the number of data points in the dataset. In addition, the calculation of separation measure between clusters is usually based on geometric centroid of each cluster but ignores other features of clusters such as shape, density and scatter features. The latest proposed OSI [22] uses a measure of multiple cluster overlap and a separation measure for each data point, both based on an aggregation operation [23], [24] of membership degrees. We can get a series of OSI indices using different aggregation operations, while how to choose the appropriate aggregation operations is always a challenge.

In this paper, we present Pairing Frequency Clustering Validity Index (PFCVI), a new clustering validity index aims to overcome the shortcomings of CVIs using compactness and separation measures. The proposed index was inspired by the following idea: For different values of c, different clustering results obtained by the same algorithm, If an arbitrary pair of data points in a dataset is always partitioned into the same cluster, then the optimal partition should also assign this very duo to the same cluster. We call the phenomenon, that a certain pair of data points always belongs to the same cluster across different value of c, pairing frequency. One advantage of PFCVI is that it is designed upon logical reasoning based on statistical analysis of pairwise patterns, rather than the frequently used compactness or separation measures. Moreover, PFCVI can be applied to both fuzzy clustering and hard clustering. Lastly, the computational cost of PFCVI does not depend on the dimension of the feature vector. A procedure for choosing the optimal value of parameter c (the number of clusters) from a range of alternatives using PFCVI is presented. At the end of this paper, evaluation of PFCVI is performed. Experiments on artificial and real datasets show that PFCVI is stable and efficient.

The rest of the paper is organized as follows. Section II describes our proposed clustering validity index in detail. In section III, we evaluate the new index with artificial and real datasets. Practical applications are showed in section IV. Concluding remarks are given in section V.

II. THE PROPOSED CLUSTERING VALIDITY INDEX BASED ON PAIRING FREQUENCY

In this section, we present the proposed clustering validity index based on pairing frequency (PFCVI). Unlike traditional CVIs introduced in section 6, our proposed PFCVI provides a new perspective to the issue of clustering validity.

A. THEORY OF PFCVI

Certain steps should be followed to determine the optimal value of c using traditional CVIs. First, perform a clustering algorithm several times with c varying in a user-defined range $[c_{\min}, c_{\max}]$. Second, compute CVI(c) for each partition result. Finally, set c_{opt} so that CVI(c_{opt}) is optimal within the predefined range, and the process of looking for c_{opt} uses each partition result independently. Unlike traditional CVIs, our proposed PFCVI takes many partition results together in order to take advantage of global information also and logical reasoning. When given different values of c in a clustering algorithm, different results will be produced. In this situation, if a pair of objects in a dataset are always assigned to the same cluster, then the optimal partition should also have the pair belong to a same cluster. PFCVI can not only tell us whether a pair of objects should be partitioned into the same cluster or not, but also it delivers a belief value, which indicates a degree of confidence that a pair of objects belong to the same cluster.

B. THE CALCULATION PROCEDURE OF PFCVI

This section describes the steps to calculate PFCVI. In next subsection.

First, we obtain the membership matrix $U = [u_{ij}]$ from the result of a clustering algorithm like FCM

$$U_{c\times n} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & \cdots & \cdots & u_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ u_{c1} & \cdots & \cdots & u_{cn} \end{bmatrix} \tag{1}$$

For each object j we define $I_j = \left[u_{ij}\right]_{\max}$, the notation $[\,]_{\max}$ acquires the value of i when u_{ij} $(1 \le i \le c)$ reaches its maximum. We can conclude that a pair of objects $\mathbf{x}_s, \mathbf{x}_k$ $(1 \le s, k \le n)$ belong to the same cluster under this value of c if $I_s = I_k$.

Second, for each value of c , we define:

$$F_c = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & \cdots & \cdots & f_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ f_{n1} & \cdots & \cdots & f_{nn} \end{bmatrix} \tag{2}$$

We call F_c a pattern matrix . The element of F_c denoted by f_{sk} indicates the degree to which $\mathbf{x}_s, \mathbf{x}_k$ belong to the same cluster or different clusters. It is calculated as follows:

$$f_{sk} = \begin{cases} 1 - \dfrac{c}{c-1} \times |\max \mathbf{u}_s - \max \mathbf{u}_k| & if\ I_s = I_k \\ -\dfrac{c}{2c-2} \times (\max \mathbf{u}_s + \max \mathbf{u}_k - \dfrac{2}{c}) & if\ I_s \neq I_k \end{cases} \tag{3}$$

where $\max \mathbf{u}_s = \max\{u_{is} : 1 \leq i \leq c\}$, $\max \mathbf{u}_k = \max\{u_{ik} : 1 \leq i \leq c\}$. Obviously, F_c is a symmetric matrix, and f_{sk} is a measurement of the extent to which two points belonging to the same or different clusters. Condition $I_s = I_k$ means that point s and point k are in the same cluster to some extent. A small value of $|\max \mathbf{u}_s - \max \mathbf{u}_k|$ means that s and k tend to belong to a certain cluster. For the second condition, $I_s \neq I_k$ means that member list for point s and point k are greatly different. And s and k should be assigned to different clusters shows a large value. In the following works, we must normalize those two values, because for fuzzy clustering, we have $\sum_{i=1}^{c} u_{ik} = 1 \forall k$. So $\frac{1}{c} \leq \max \mathbf{u}_s, \max \mathbf{u}_k \leq 1$, then we have $0 \leq |\max \mathbf{u}_s - \max \mathbf{u}_k| \leq \frac{c-1}{c}$ and $\frac{2}{c} \leq |\max \mathbf{u}_s + \max \mathbf{u}_k| \leq 2$. To make the nornalization meet the above requirement, we set the normalized formula as the Eq.(3). Overall, the case $0 < f_{sk} \leq 1$ suggests that $\mathbf{x}_s, \mathbf{x}_k$ share the majority of their membership in the same cluster. The closer f_{sk} is to 1, the stronger is their affinity for belonging together. Correspondingly, the case $-1 \leq f_{sk} < 0$ suggests that $\mathbf{x}_s, \mathbf{x}_k$ have little in common for this value of c. The closer f_{sk} is to -1, the stronger is the disassociation between $\mathbf{x}_s, \mathbf{x}_k$. For the case of hard clustering, the following formula is used to calculate f_{sk}

$$f_{sk} = \begin{cases} 1 & if\ s, k\ in\ the\ same\ cluster \\ -1 & if\ s, k\ in\ different\ clusters \end{cases} \tag{4}$$

Next, we combine a certain number of F_c to obtain the pairwise pattern the matrix which is denoted by P. P is defined as: $P = \sum_{c=2}^{c_{upper}} F_c$, where c_{upper} is a parameter will be discussed later in Section 2.4. Usually $c_{upper} \geq c_{\max}$, and P can be normalized under this condition as Q:

$$Q = P/(c_{upper} - 1) \tag{5}$$

matrix Q is the final global pairwise pattern matrix which indicates the likelihood of two data points to the same cluster (or different clusters). From definition from above we can see that matrix Q takes advantage of the information about all of the partitions obtained by FCM for the range of c used.

At last, our proposed clustering validity index PFCVI is defined as:

$$PFCVI(c) = S(Q \circ F_c) \tag{6}$$

where notation '$\circ$' represents Hadamard product, and S represents the *sum* of $Q \circ Fc$.

Let's look at element $(q_{sk} \cdot f_{sk})$ to fully understand PFCVI. If f_{sk} and q_{sk} share the same sign, namely pair (s, k) in a FCM's result for a specific value of c will be in accord with that in the global pattern matrix, so the pair (s, k) will contribute positively to $PFCVI(c)$, and vice versa. The " pattern of pair (s, k)" means the occurrence of (s, k) belonging to the same cluster or different clusters. Therefore, a larger value of $PFCVI(c)$ means that the clustering process with parameter c is more appropriate for a given dataset, and c producing the optimal value of PFCVI will be chose as the final result.

PFCVI does not compute Euclidean distances like many of the other CVIs, so the computation cost is independent of the dimension of feature vectors. So PFCVI is relatively efficient when dealing with high dimension data.

C. THE PROCEDURE OF SELECTING THE OPTIMAL VALUE OF c USING PFCVI

1) For each value of $c = 2, 3, \ldots, c_{upper}$, we carry out the corresponding clustering algorithm and compute F_c $(c = 2, 3, \ldots, c_{upper})$ using Eq.(2).
2) Compute the matrix Q using Eq.(5).
3) Compute $PFCVI(c)$ $(c = c_{\min}, \ldots, c_{\max})$ using Eq.(6).
4) $c_{opt} = \underset{c_{\min}, \ldots, c_{\max}}{\arg\max}$ (PFCVI(c))

D. COMMENTS ABOUT c_{upper}

Step 1 of this algorithm generates F_c for $c = 2$ to c_{uppper}, and when computing the global pattern matrix Q , the upper bound of the summation operator is c_{upper} instead of $c_{\max}$. c_{upper} is a main factor influencing the performance of $PFCVI$ for the following reasons. If the optimal but unknown number of clusters in a dataset (denoted by $c*$) is much larger than c_{upper} ($c_{upper} \ll c*$), there will be some pairs in all of the partitions that would eventually be split when c gets to $c*$, so the matrices F_c on hand will add incorrect information global pattern matrix Q. Let us give an example to illustrate this case. Suppose that $\mathbf{x}_s, \mathbf{x}_k$ belong to different clusters. The value of c ($c = 2, 3 \ldots, c_{upper}$) may be too small to split the pair $\mathbf{x}_s, \mathbf{x}_k$ into different clusters. We suppose $c_{upper} = 2$ and the actual number of clusters in a dataset $c* \gg 2$. In this case, many pairs of objects will be incorrectly paired because there are only two clusters, although most of them actually belong to different clusters. On the other hand, if the number of clusters in a dataset is much smaller than c_{upper} ($c_{upper} \gg c*$), some pairs of objects will be split into different clusters in the clustering process with a larger value of parameter c , although they are likely to belong to the same cluster. This also adds incorrect information into the global pattern matrix Q. To prevent this "splitting action" from happening, we try to make c_{upper} satisfy the formula: $c_{upper} \geq c_{\max}$.

The data sets used in our experiments are all labeled, and the maximum number of labelled subsets is less than 10, so we can simply set $c_{\max} = \min(10, \lfloor \sqrt{n} \rfloor)$ and $c_{upper} = \max(c_{\max}, \lfloor 0.5\sqrt{n} \rfloor)$, which worked well in our experiments.The optimal choice for c_{opt} falls in the middle of $(2, c_{upper})$.

When the number of samples n is too large, these heuristics become useless, so we can simply set $c_{max} = c_{upper}$.

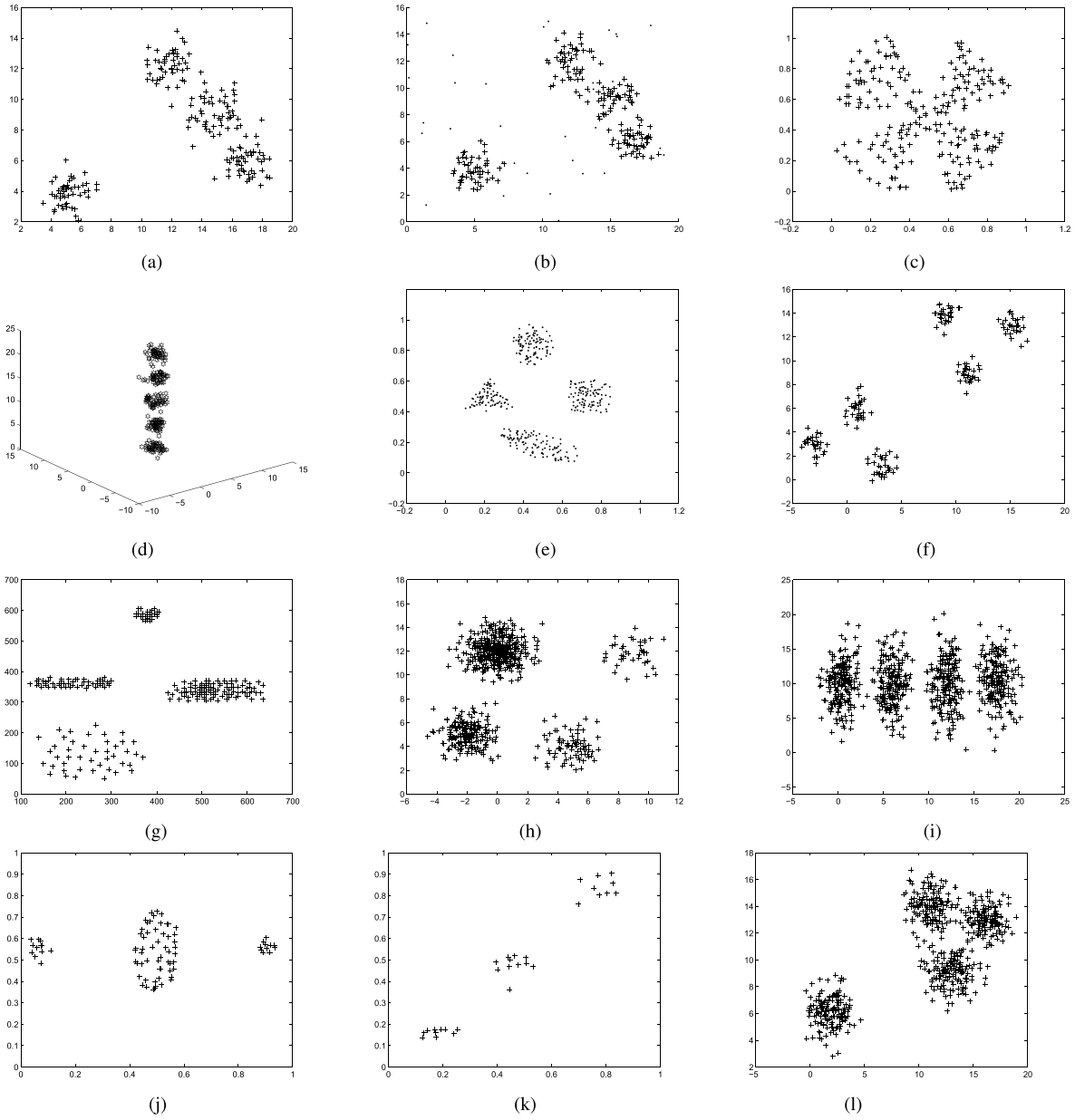

FIGURE 1. All the artificial datasets. (a) Over. (b) Over+Noise. (c) Bridge. (d) 3D-GD. (e) Shape. (f) Local Closeness. (g) Shape+Density. (h) Gaussian+Density. (i) Edge. (j) Size. (k) X30. (l) Local Overlap.

III. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed PFCVI by conducting extensive comparisons between eight CVIs (PC, PE, XB, FHV, PBMF, PCAES, VW, OSI) and FCM algorithm. A small note about the OSI is that the standard norms, namely, the min of t-norms and the max of t-conorms are used in our experiment for simplicity and representativeness. As in almost all papers dealing with fuzzy clustering validity, the fuzzifier exponent m is set to 2, the termination parameter for FCM for convergence is set to 10^{-3} and the Euclidean distance is used. The optimal number of clusters in range $[c_{\min} = 2, c_{\max}]$, with $c_{\max} = \min(10, \lfloor\sqrt{n}\rfloor)$ in order to ensure a good balance between the number of clusters and the number of points in dataset [12]. We set $c_{upper} = \max(c_{\max}, \lfloor 0.5\sqrt{n}\rfloor)$ in our experiments. In order to reduce the influence of random initialization for FCM, we run the FCM algorithm 20 times for each dataset and compute the corresponding CVIs 20 times too. Then we take the average as the final result. All the values of CVIs are normalized to fall in the interval interval [0, 1].

A. DATASETS

15 diverse characteristics datasets were used to evaluate our index, such as good separation, overlapping clusters, different shape of clusters, differences in density and additional noisy

TABLE 1. Datasets with different properties.

Data set	Cluster properties	n	p	$c*$
(a) Over	Overlapping	200	2	2 or 4
(b) Over+Noise	Overlapping with noisy points	230	2	2 or 4
(c) Bridge	Connected clusters	220	2	4
(d) 3D-GD	Three dimension Gaussian distribution	250	3	5
(e) Shape	Different shapes	400	2	4
(f) Local Closeness	Local closeness of more than one clusters	180	2	6
(g) Shape+Density	Different shapes and densities	225	2	4
(h) Gaussian+Density	Different densities by Gaussian Distribution	720	2	4
(i) Edge	Adjacent clusters with blurry edge	800	2	4
(j) Size	Different size and number of points	79	2	3
(k) X30	Well separated	30	2	3
(l) Local Overlap	Locally overlapping clusters	600	2	2 or 4
Iris		150	4	2 or 3
Breast Cancer		699	9	2
Pima		768	8	2

points. The first twelve datasets are artificial two-dimensional datasets such that ground truth can be visually assessed by examining their scatterplots in Fig. 1. Remaining three are real datasets from UCI Machine Learning Repository [31]. Table 1 presects the brief descriptions of all datasets. The dataset Over contains 200 points sampled from a mixture of $c = 4$ bivariate normal distributions of where 50 points for each component[see Fig.1(a)]. In Over+Noise dataset, 30 noise points sampled from a uniform distribution are added to the over dataset to simulate a noisy environment [see Fig.1(b)]. The dataset Bridge is composed of four connected clusters [see Fig.1(c)]. The dataset Local Closeness consists of 6 clusters which is likely to be partitioned into two clusters because of the local closeness [see Fig.1(f)]. The Iris dataset contains data from three types of Iris, named respectively as Setosa, Versicolor and Virginica, each of which contains 50 objects described by 4 dimensions (features). Although Iris has 3 labeled subsets, 2 of them are substantially overlapped. The consensus in most of the literature is that Iris has only 2 clusters which are optimal for most clustering models, but there are still some clustering algorithms can actually produce three clusters, so 2 or 3 are commonly regarded as the reasonable clustering results of Iris dataset.The Breast Cancer dataset contains 699 instances but 16 of them are removed because they are incomplete. The Pima dataset consists of 768 instances from two overlapping labels.

B. EXPERIMENT RESULT

Fig. 2 and Fig. 3 plot average values of all 8 indices for all the considered datasets (Fig. 3 for artificial datasets, Fig. 2 for real datasets). Optimal number of clusters are displayed on the horizontal axis and the y-value denotes normalized value of the CVIs. The c_{opt} for each CVI is shown in the filled dots and the PFCVI is denote as filled triangles. Table 2 summarizes the optimal number of clusters obtained from the tested CVIs on artificial and real datasets. The expected numbers of clusters are showned in $c*$ column, which is either the physical number of clusters given by an expert (real datasets) or can be figured out visually (artificial datasets). For datasets Over and Over+Noise, we can say they have 2 visually apparent clusters, or have 4 clusters with three of them being overlapped a little. We can see from Table 2 that PC,PE,OSI,VW suggest 2 as the result for c_{opt} while PFCVI shows that $c_{opt} = 4$ should be equal to 4, which indicates that these two dataset may have 2 well separated clusters and 4 fuzzy clusters. The dataset Bridge with linking points is a difficult problem for most of the indices. Here VW, PFCVI find the right number of clusers. The structures of the Shape and X30 datasets is relative easy, therefore, most of the presented indices, including the proposed one, correctly identify the right values 3 and 4. Data points from Edge dataset are sampled from a mixture of 4 gaussian distributions, however, the edge of the cluster is got blurred. As a result, PFCVI and XB think the optimal cluster numbers to be 4 and that is aligned with expectation. For the Size dataset, PC, OSI, PFCVI performs well on estimate the correct number of clusters. The dataset Local overlap has four clusters with three of them being overlapping, this is very similar with the dataset Over and Over+Noise. In the same way, we think 2 and 4 are the reasonable cluster number. On this view, FHV and PFCVI produced the correct number of clusters 4 while others tend

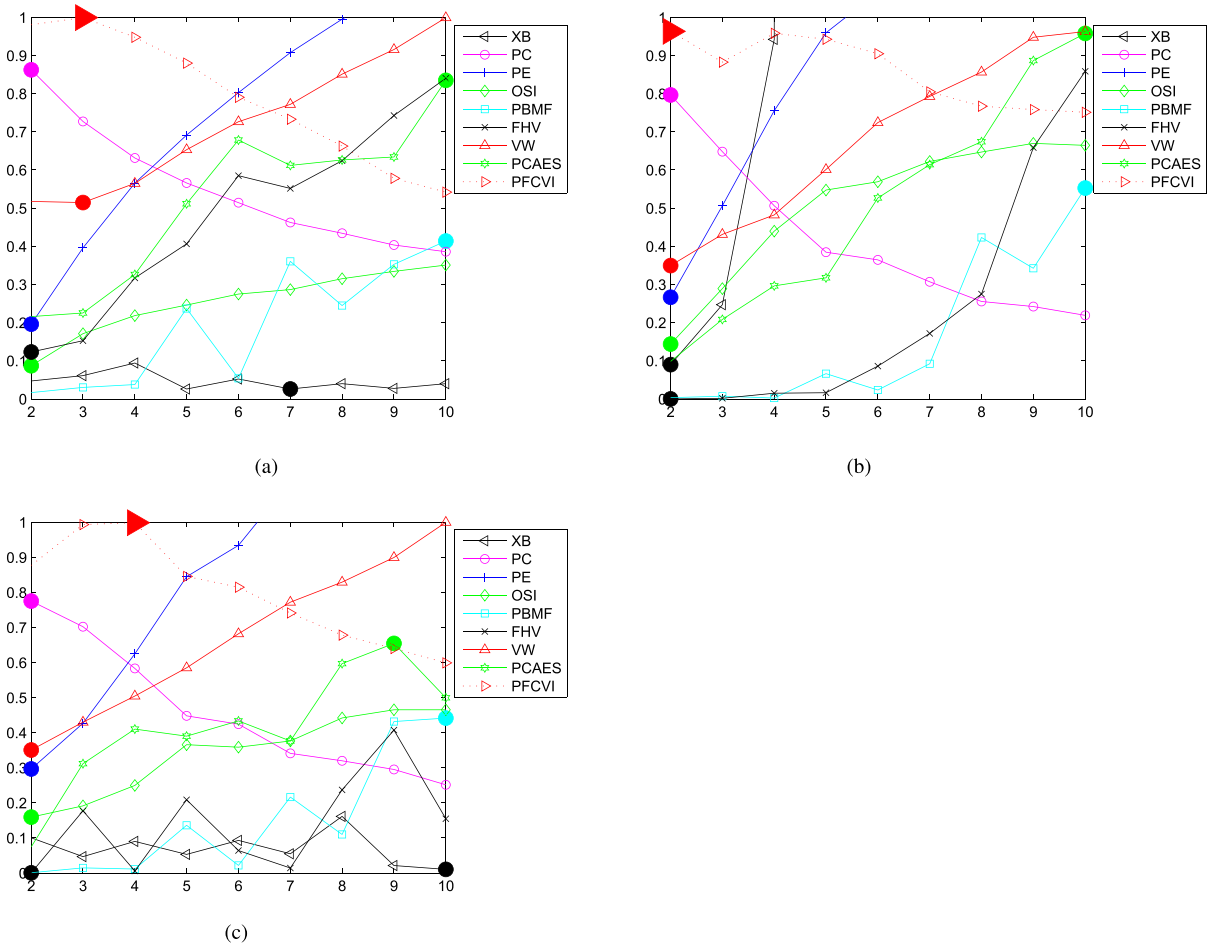

FIGURE 2. Value of all CVIs with c from c_{min} to c_{max} on 3 real datasets. (a) Iris dataset. (b) Breast Cancer dataset. (c) Pima dataset.

TABLE 2. Optimal number of clusters obtained using different CVIs on artificial and real datasets.

Dataset	$c$max	c_{upper}	XB	PC	PE	OSI	PBMF	FHV	VW	PCAES	PFCVI	c^*
(a) Over	10	10	8	2	2	2	9	3	2	9	4	2 or 4
(b) Over+Noise	10	10	5	2	2	2	10	3	2	10	4	2 or 4
(c) Bridge	10	10	10	2	2	5	10	2	4	10	4	4
(d) 3D-GD	10	10	5	2	2	5	10	2	5	10	5	5
(e) shape	10	10	10	4	4	4	10	4	4	10	4	4
(f) Local closeness	10	10	8	2	2	2	6	6	6	2	6	6
(g) Shape+Density	10	10	4	4	2	4	10	4	2	10	4	4
(h) Density	10	10	4	4	2	4	10	2	4	4	4	4
(i) Edge	10	10	4	2	2	2	10	2	2	10	4	4
(j) Size	8	8	6	3	2	3	6	4	4	10	3	3
(k) X30	5	5	4	3	3	3	5	4	4	3	3	3
(l) Local overlap	10	10	10	2	2	2	10	4	2	9	4	4
Iris	10	10	7	2	2	2	10	2	3	10	3	2 or 3
Breast Cancer	10	13	2	2	2	2	10	2	2	10	2	2
Pima	10	13	10	2	2	2	10	2	2	9	4	2

to get 2 as optimal number except for XB, PBMF, PCAES, which are far from the reasonable cluster number. It is generally accepted that the right number of clusters for the Iris dataset is two or three (the number of physical classes). Most of the indices indicate either $c* = 2$ or $c* = 3$. Interestingly, for c=2 and c=3, values of PFCVI is very close, which may also indicate that both of the results are rational. Almost all indices identify the right number of clusters for the Breast Cancer dataset except PBMF and PCAES. For the dataset Pima, our proposed PFCVI does not identify the right number

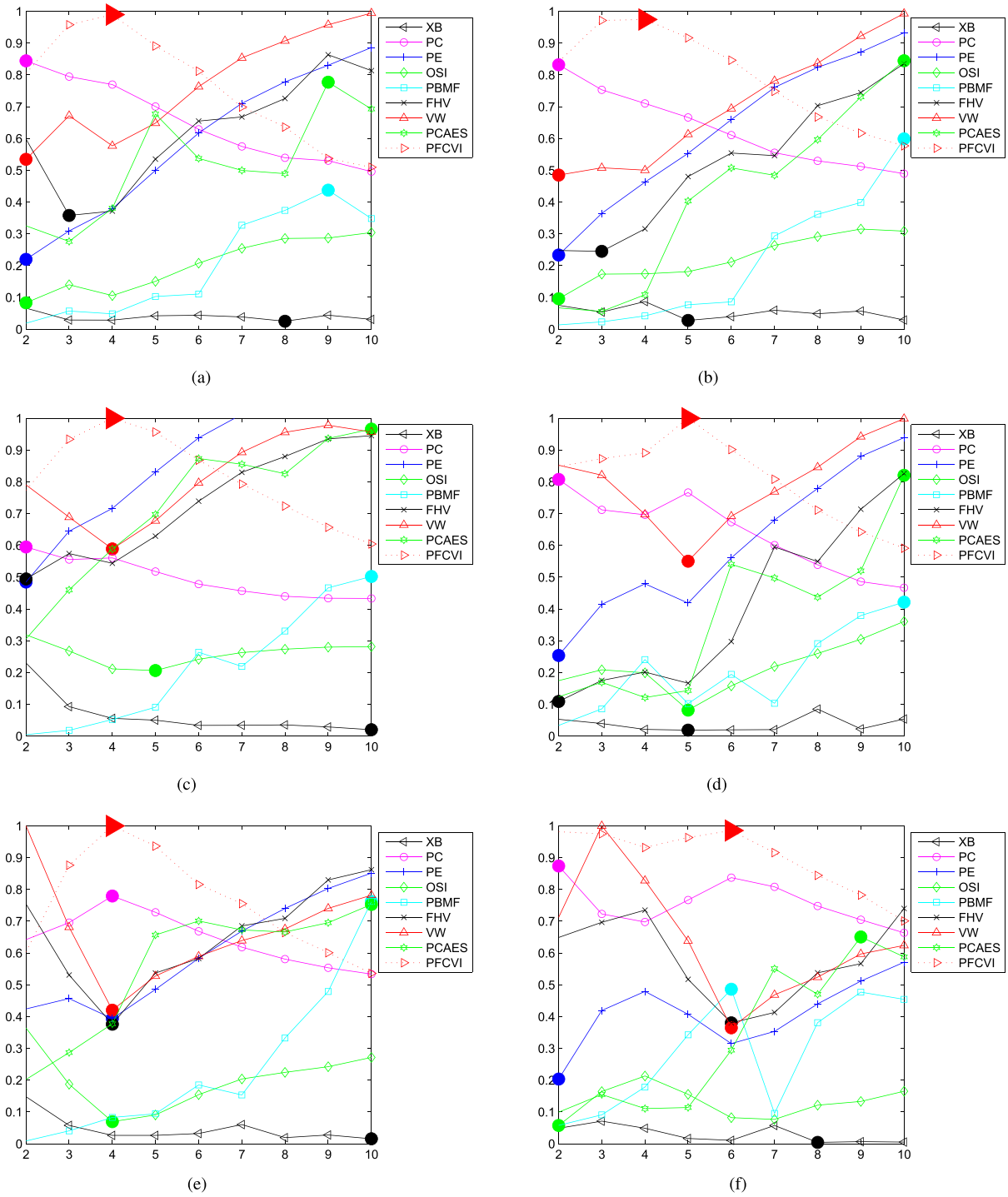

FIGURE 3. Average values of all CVIs with c from c_{min} to c_{max} on 12 artificial datasets. (a) Over. (b) Over+Noise. (c) Bridge. (d) 3D-GD. (e) Shape. (f) Local Closeness.

of labels infuenced by the 'splitting action phenomenon' phenomenon splitting action. In conclusion, none of the indices correctly recognizes the expected number for all the datasets. There will hardly be an index that is suitable for a large number of different datasets. As Pal and Bezdek [32] stated, ''no matter how good your index is, there is a dataset out there waiting to trick it (and you)''.

IV. PRACTICAL APPLICATION

In this section, we apply our proposed clustering validity index in two real tasks. The first one is to cluster Serving GPRS Support Nodes (SGSNs) in one main city of China based on service characteristics. The second one is to analyse user preferences for eight types of internet service based on their behavior records.

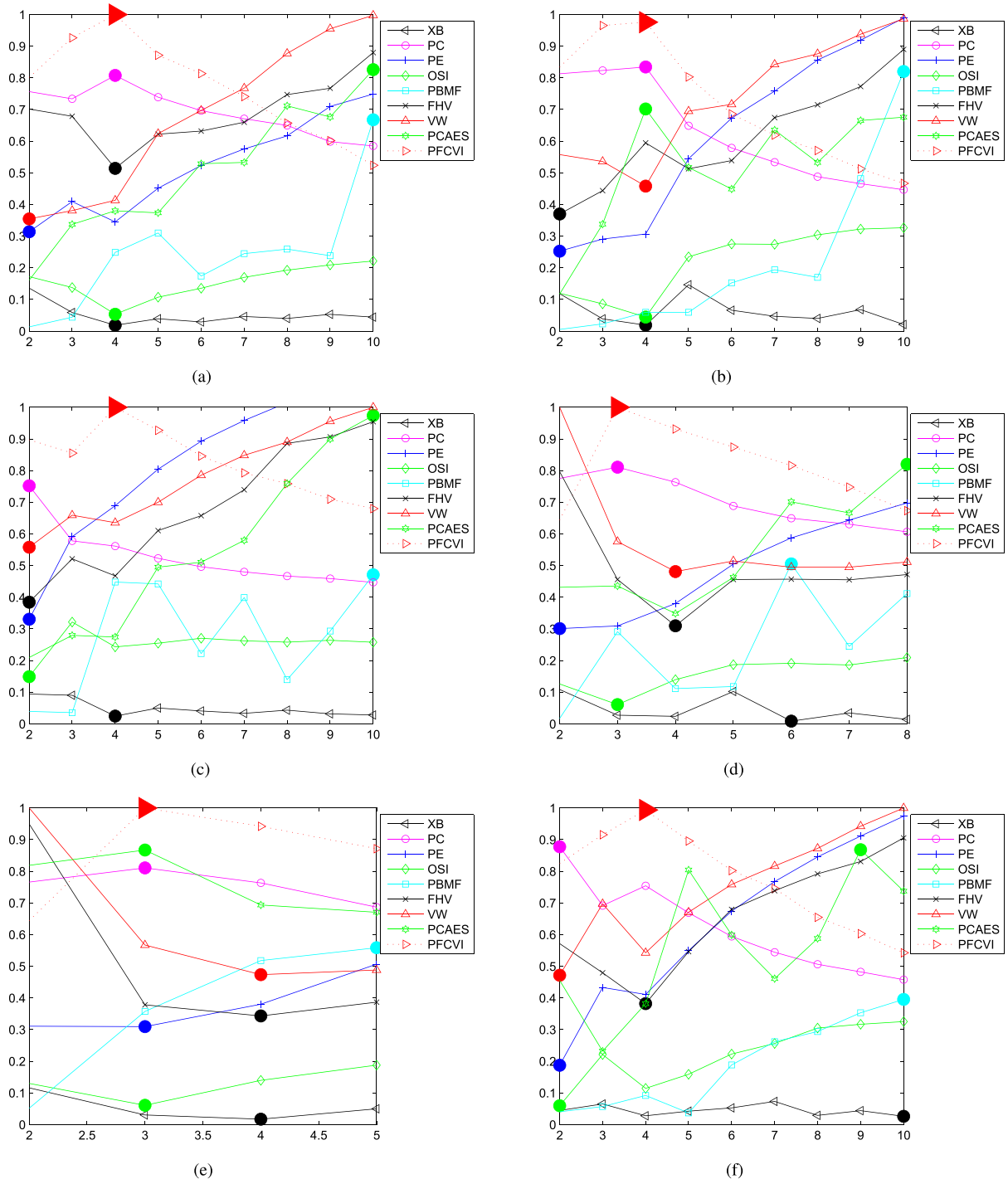

FIGURE 4. Average values of all CVIs with c from c_{min} to c_{max} on 12 artificial datasets. (a) Shape+Density. (b) Gaussian+Density. (c) Edge. (d) Size. (e) X30. (f) Local Overlap.

A. CLUSTER OF SGSNs

We use our proposed index FPCVI combined with FCM to analyze the SGSNs' and divide SGSNs based on the characteristics of customers' behavior. The dataset, which contains 3270860 usage records with connections to 921 SGSNs. Each SGSN acts as a data point, which contains eight feature items: the traffic of three IM (Instant Messaging) applications (QQ, MSN, WeChat), three streaming media applications (PPTV, PPS, KanKan), the records' amount and total traffic in each SGSN. So, FCM is applied to a set of n=921 objects, each represented by an 8 dimensional feature vector.

After choosing $c_opt = 8$ and hardening the corresponding fuzzy 8-partition of the data, Table 3 shows that the number of SGSNs in Cluster III is 833, and no more than 30 in the other seven clusters. The cluster center of traffic amount for

TABLE 3. SGSN characteristics of various types.

SGSN type	SGSN number	users number	records number	Mean flow of cluster centers
cluster 1	4	4.92E+03	6.00E+04	2.01E+09
cluster 2	27	7.02E+04	1.86E+06	2.99E+08
cluster 3	833	1.36E+05	2.78E+06	1.10E+07
cluster 4	5	5.75E+03	1.06E+05	7.59E+08
cluster 5	10	7.36E+05	9.43E+06	4.06E+09
cluster 6	11	1.20E+06	1.87E+07	7.93E+09
cluster 7	28	1.11E+06	1.73E+07	7.45E+08
cluster 8	3	2.80E+03	9.12E+04	1.57E+08

TABLE 4. c_optimal regarding to c_max ranging from 8,9,10,11,12.

c_max	8	9	10	11	12
c_optimal	2	6	5	4	6

Cluster III is 10^7, while other cluster centers are 10^8 or 10^9. That is to say, the magnitude of records'amount and users' amount in the Cluster III are 1-2 orders lower than other clusters. The relationship of SGSNs and real geographical positions can be used to explain the subscribers' usage pattern of traffics. When a subscriber uses traffic service in other cities or countries instead of his own city, the traffic expense will be high because of the cost of roaming. So it is easy to understand that the user would reduce his traffic usage in cellular networks because of the economic factor when he roams out of his own city, and he would turn to WiFi instead. The subscriber may prefer to use more data traffic in his own city. Now let us look at our clustering results. Those SGSNs with high traffic amounts are local SGSNs, like Cluster VI. On the contrary, those SGSNs with lower traffic amounts in the Cluster III spread all over the country, the real geographical positions of SGSNs in the Cluster III involve 25 different provinces, and the traffic amounts passing through these SGSNs are far lower than other SGSNs. From this, it can be concluded that the SGSNs in clusters V, VI, and VII are located in the home city.

B. CLUSTERS OF USERS GROUP

We applied this cluster index to cluster an operators's users in a certain city, and divided user groups of different traffic characteristics. Each user is represented by an 8-dimensional feature vector, which records the percentage of different types of services within an individual's total records. The services include Multimedia Message Service(MMS), Web, Instant Messaging (IM), Streaming media, E-mail, Phone call, File transfer / P2P and other types of services.

Here, c_{max} and c_{upper} are set equal, and take the values 8, 9, 10, 11, 12. Our cluster results are listed in Table 4, for $n = 313505$ user profiles. From the cluster results, we find that there is a big difference between the clusters when we divide the users into 4 groups. If users are clustered into more clusters, some clusters present similar patterns and could have been included into the 4 clusters so the result of $c\_optimal = 4$ when $c\_\max = 11$ is most reasonable.

In the case of 4 clusters, for the users in the 1st cluster and the 2nd cluster, Web and Instant messaging account for almost 50% of traffic record amounts. Users in the 1st cluster tend to use Web service more, while the users in the 2nd cluster have a tendency to use IM service more. Users in the 3rd cluster are mild users of Web and IM, while other unmentioned applications account for a big chunk of their usage. Users in the 4th cluster are heavy users of Web and IM, and Web usage weights far more than IM. We lists the characteristics of the cluster centers in Table 5. Because the percentages of other kinds of traffic uses are very small, so only Web, IM and Others out of 8 dimensions are listed here.

TABLE 5. Cluster centers of 4 clusters.

	cluster 1	cluster 2	cluster 3	cluster 4
Web	39.88%	21.17%	11.91%	65.03%
Instant Messaging	5.50%	34.21%	1.78%	6.18%
Others	54.01%	44.03%	85.74%	28.4251%

V. CONCLUSION

In this paper, a new clustering validity index called PFCVI has been proposed which is based on pairing frequency instead of compactness-to-separation ratio criteria employed by some classic CVIs. Another significant difference compared with other CVIs is that PFCVI, with the help of the global pattern matrix, takes advantage of the information from more than one clustering processes to compute the value of PFCVI(c). An extensive comparison with seven other widely used indices shows that our new index performs well for most of the datasets used in this study.

REFERENCES

[1] S. Theodoridis and K. Koutrombas, *Pattern Recognition*. London, U.K.: Academic, 2006, pp. 529–533.
[2] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973.
[3] J. C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-2, no. 1, pp. 1–8, Jan. 1980.
[4] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York, NY, USA, 1981.
[5] J. C. Bezdek, J. M. Keller, R. Krishnapuram, and N. R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Norwell, MA, USA: Kluwer, 1999.
[6] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.
[7] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2, pp. 107–145, 2001.
[8] M. Brun, C. Sima, J. Hua, and B. Lowey, E. Suh, and E. R. Dougherty, "Model-based evaluation of clustering validation measures," *Pattern Recognit.*, vol. 40, no. 3, pp. 807–824, 2007.
[9] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy C-means model," *IEEE Trans. Fuzzy Syst.*, vol. 3, no. 3, pp. 370–379, Aug. 1995.
[10] W. Wang and Y. Zhang, "On fuzzy cluster validity indices," *Fuzzy Sets Syst.*, vol. 158, no. 19, pp. 2095–2117, Oct. 2007.
[11] J. C. Bezdek, "Numerical taxonomy with fuzzy sets," *J. Math. Biol.*, vol. 1, no. 1, pp. 57–71, 1974.
[12] J. C. Bezdek, "Cluster validity with fuzzy sets," *J. Cybern.*, vol. 3, no. 3, pp. 58–74, 1973.
[13] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841–847, Aug. 1991.

[14] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy C-means method," in *Proc. 5th Fuzzy Syst. Symp.*, 1989, pp. 247–250.
[15] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 773–780, Jul. 1989.
[16] N. Zahid, M. Limouri, and A. Essaid, "A new cluster-validity for fuzzy clustering," *Pattern Recognit.*, vol. 32, no. 7, pp. 1089–1097, 1999.
[17] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognit.*, vol. 37, no. 3, pp. 487–501, 2004.
[18] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification," *Fuzzy Sets Syst.*, vol. 155, no. 2, pp. 191–214, 2005.
[19] K.-L. Wu and M.-S. Yang, "A cluster validity index for fuzzy clustering," *Pattern Recognit. Lett.*, vol. 26, no. 9, pp. 1275–1291, 2005.
[20] Y. Zhang, W. Wang, X. Zhang, and L. Yi, "A cluster validity index for fuzzy clustering," *Inf. Sci.*, vol. 178, no. 4, pp. 1205–1218, 2008.
[21] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[22] H. L. Capitaine and C. Frélicot, "A cluster-validity index combining an overlap measure and a separation measure based on fuzzy-aggregation operators," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 3, pp. 580–588, Jun. 2011.
[23] T. Calvo, A. Kolesárová, M. Komorníková, and R. Mesiar, *Aggregation Operators: Properties, Classes and Construction Methods*. Heidelberg, Germany: Physica-Verlag, 2002, pp. 3–106.
[24] M. Grabisch, J. Marichal, R. Mesiar, and E. Pap, *Aggregation Functions* (Encyclopedia of Mathematics and its Applications Series), vol. 127. Cambridge, MA, USA: Cambridge Univ. Press, 2009.
[25] J. Yu, Q. Cheng, and H. Huang, "Analysis of the weighting exponent in the FCM," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 634–639, Feb. 2004.
[26] D. Dembélé and P. Kastner, "Fuzzy C-means method for clustering microarray data," *Bioinformatics*, vol. 19, no. 8, pp. 973–980, 2003.
[27] M. Bouguessa, S. Wang, and H. Sun, "An objective approach to cluster validation," *Pattern Recognit. Lett.*, vol. 27, no. 13, pp. 419–430, 2006.
[28] E. P. Klement and R. Mesiar, *Logical, Algebraic, Analytic and Probabilistic Aspects of Triangular Norms*. New York, NY, USA: Elsevier, 2005.
[29] L. Mascarilla, M. Berthier, and C. Frélicot, "A k-order fuzzy OR operator for pattern classification with k-order ambiguity rejection," *Fuzzy Sets Syst.*, vol. 159, no. 15, pp. 2011–2029, 2008.
[30] S. H. Kwon, "Cluster validity index for fuzzy clustering," *Electron. Lett.*, vol. 34, no. 22, pp. 2176–2177, Oct. 1998.
[31] A. Asuncion and D. J. Newman. (2007). "UCI machine learning repository." School Inf. Comput. Sci., Univ. California, Irvine, CA, USA. Tech. Rep. [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html
[32] N. R. Pal and J. C. Bezdek, "Correction to 'on cluster validity for the fuzzy C-means model' [Correspondence]," *IEEE Trans. Fuzzy Syst.*, vol. 5, no. 1, pp. 152–153, Feb. 1997.

HONGYAN CUI (SM'14) researched in Massachusetts Institute of Technology as a visiting scholar since 2014, and in Australia CSIRO ICT Center in 2009. She is currently a Professor and a Ph.D. Supervisor in the School of Information and Communications Engineering, Beijing University of Posts and Telecommunications. She is a Founding Partner and a Deputy Director of the Specialties Committee of Smart Healthcare, China Ministry of Industry and Information Technology. She also acts as the Project Leader or a Primary Researcher for more than ten national research projects. She has published over 70 SCI/Ei papers and five books. She holds more than ten national patents. Her research interests include big data analysis and visualization, intelligent resource management in future networks, cloud technology, and social physics. She acts as the Publish Chair of IEEE UV'18, and the Track Chair of WPMC'13, Global Wireless Summit'14, ICC'15, the IEEE UV'16, and Globecom'16. She is a TPC member in IEEE WPMC'14, ICCC'15, ICC'15 and '16, and Bigdata'15, CIoT'16, BigData'15, Globlecom'16. She acts as a Reviewer for JSAC, Chaos, Transactions on NNLS, Globecom, ICC, WCNC, and WPMC.

KUO ZHANG received the master's degree in information and communications engineering from the Beijing University of Posts and Telecommunications, China, in 2014. He is currently pursuing the Ph.D. degree in computer science at Rutgers University, New Brunswick, NJ, USA.

He has been an Engineer with Baidu since 2014. His research interests are machine learning and big data analysis. He received the Excellent Students Scholarship Award of BUPT in 2011 and 2012. He has published 3 Ei papers and 1 SCI paper.

YAJUN FANG received the Ph.D. degree from the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT). She is currently a Research Scientist with MIT Intelligent Transportation and Research Center. She is also the Program Coordinator of the MIT Universal Village (an advanced version of Smart Cities) Program, and the Conference Chair of International Conference on Universal Village. Her research areas focus on machine vision, machine learning, big data analysis, system theory, and their applications for robotics/autonomous vehicles, intelligent transportation systems, intelligent healthcare, future smart cities/universal village.

STANISLAV SOBOLEVSKY received the Ph.D. degree in 1999 and the D.Sc. (Habilitation) degree in 2009 in mathematics in Belarus. He has been an Associate Professor of practice with the Center for Urban Science and Progress, New York University, since 2015. His former work experience includes research, faculty, and administrative positions at Massachusetts Institute of Technology, Belarusian State University, and Academy of Science of Belarus. He applies his fundamental quantitative background to studying human behavior in urban context and a city as a complex system through its digital traces-spatio-temporal big data created by various aspects of human activity. He has authored over hundreds of research papers in the top journals like PNAS, *Scientific Reports*, *Physical Review E*, PLoS ONE, *Royal Society Open Science*, *EPJ Data Science*, *Applied Geography*, *Environment and Planning B*, the *International Journal of GIS*, *Studies in Applied Mathematics*, and others. His research is conducted in close cooperation with city agencies and industrial partners from banking, telecom, defense, insurance, and other areas.

CARLO RATTI received the M.Sc. degree in engineering form the Politecnico di Torino, Italy, and the Ecole des Ponts, France, and the M.Phil. and Ph.D. degrees in architecture from the University of Cambridge, U.K. He is currently a Professor of practice of urban technologies with the Massachusetts Institute of Technology, USA, where he directs the Senseable City Laboratory. He is also a Founding Partner of the international design and innovation office Carlo Ratti Associati. He then moved to the Massachusetts Institute of Technology as a Fulbright Senior Scholar. His research interests include urban design, human–computer interfaces, electronic media, and the design of public spaces.

BERTHOLD K. P. HORN is currently an Academician and a Professor of computer science and engineering with the Massachusetts Institute of Technology. He is with the Computer Science and Artificial Intelligence Laboratory, Department of Electrical Engineering and Computer Science. He published the book *Robot Vision*, and has been a Translator for many national languages. He was elected as a fellow of the American Association for Artificial Intelligence for his significant contributions to the field of AI in 1990. He was elected to the National Academy of Engineering for his contributions to computer vision, including the recovery of three-dimensional geometry from image intensities in 2002. He received rich awards, including Rank Prize for pioneering work leading to practical vision systems (Rank Prize Funds) in 1989, and the Azriel Rosenfeld Lifetime Achievement Award (IEEE Computer Society) for pioneering work on early vision including optical flow and shape from shading in 2009.