Data Min Knowl Disc (2015) 29:1560–1597
DOI 10.1007/s10618-014-0387-5

Cluster validity functions for categorical data: a solution-space perspective

Liang Bai · Jiye Liang

Received: 26 April 2013 / Accepted: 18 September 2014 / Published online: 2 October 2014

Abstract For categorical data, there are three widely-used internal validity functions: the k-modes objective function, the category utility function and the information entropy function, which are defined based on within-cluster information only. Many clustering algorithms have been developed to use them as objective functions and find their optimal solutions. In this paper, we study the generalization, effectiveness and normalization of the three validity functions from a solution-space perspective. First, we present a generalized validity function for categorical data. Based on it, we analyze the generality and difference of the three validity functions in the solution space. Furthermore, we address the problem whether the between-cluster information is ignored when these validity functions are used to evaluate clustering results. To the end, we analyze the upper and lower bounds of the three validity functions for a given data set, which can help us estimate the clustering difficulty on a data set and compare the performance of a clustering algorithm on different data sets.

Keywords Cluster analysis · Cluster validity function · Generalization · Effectiveness · Normalization

Responsible editor: G. Karypsis.

L. Bai · J. Liang (✉)
Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China
e-mail: ljy@sxu.edu.cn

L. Bai
Key Laboratory of Network Data Science and Technology, Institute of Computing Technology Chinese Academy of Sciences, Beijing 100190, China
e-mail: sxbailiang@hotmail.com

1 Introduction

Clustering is an unsupervised classification technique that is used to group a set of unlabeled objects into clusters so that the objects in the same cluster have high similarity but are very dissimilar with objects in other clusters. Many types of clustering algorithms have been developed in the literature (e.g., Jain and Dubes 1988 and references therein), which have extensive applications in various domains, including information retrieval, image processing and biological engineering.

Cluster validation is an important part of cluster analysis (Jain and Dubes 1988). For a data set, there are a number of possible partitions. This has led to the use of cluster validity functions to evaluate the quality of the partitions and select the one that best fits the data to be as the final clustering result. Generally speaking, there are two types of cluster validation techniques, i.e., external and internal validity functions (Halkidi et al. 2001; Halkidi and Vazirgiannis 2001; Zhao and Karypis 2004; Steinbach et al. 2000). Their difference is whether or not the external information (class labels) is used (Xiong et al. 2009). External validity functions are mainly to compare or measure the similarity between two clustering results. People often use them to measure the clustering output with the "true" partition determined by the class label information. Internal validity functions (Liu et al. 2010, 2013) are mainly to evaluate the quality of the internal structure of a partition without external information. Internal validity functions can be further divided into two types. The one does not participate in clustering process but is used to evaluate clustering results and determine the parameters, e.g., the number of clusters k. The other can be as objective functions and help users find out the best clustering result in clustering process. The focus of this paper will be on the second type.

In the literature (e.g., Dunn 1973; Berry and Linoff 1996; Xiong et al. 2009; Wu et al. 2010; Yu 2005 and references therein), there have been considerable research efforts on studying the cluster validity functions for numerical data. However, further investigation is needed to develop validity functions for clustering categorical data, where records are made up of nonnumerical data, since this task is of great practical relevance in several fields ranging from statistics to psychology (Aggarwal et al. 2002; Barbara and Jajodia 2002; Baxevanis and Ouellette 2001; Gowda and Diday 1991; Wrigley 1985). Due to the lack of intuitive geometric properties between categorical values, the techniques used in cluster validation for numerical data are not suitable for categorical data (Huang 1997; Chen et al. 2008; Chen and Liu 2005, 2009). For categorical data, there are three well-known internal validity functions: the k-modes objective function (Huang 1997), the category utility function (Gluck and Corter 1985) and the information entropy function (Barbara et al. 2002). Many algorithms have been developed to use these validity functions as objective functions and find their (local) optimal solutions. The brief reviews of these functions and relevant algorithms can be found in Sect. 2. While the above functions are used to evaluate the clustering results, the three issues are needed to discuss:

(1) *How do we discover generality and difference of these validity functions?* The different validity functions are often defined based on the different assumptions. Thus, these validity functions for a given clustering result maybe have the same

or different evaluation results. Therefore, we need to analyze the generality and difference of these validity functions. The obtained difference can help users select the most suitable function for cluster analysis. The obtained generality can promote the mutual learning of clustering algorithms with different objective functions.

(2) *Is the between-cluster information ignored when these validity functions are used to evaluate clustering results?* A good cluster result should have high within-cluster similarity and low between-cluster similarity. However, the three validity functions are based on the within-cluster information only. Therefore, we need to discuss the relation between the within-cluster and between-cluster information in using these validity functions to evaluate the clustering results.

(3) *How do we normalize these validity functions for a given data set?* There should be two application scenarios of the validity functions, which was introduced by Luo et al. (2009). First, the validity functions can be used to compare the performances of different clustering algorithms in a given data set. In this case, the validity functions mainly help users select the one that best fits the data set from a large number of clustering algorithms. Second, the validity functions can be used to compare the performances of a specific clustering algorithm in different data sets. In this case, the validity functions mainly help a specific clustering algorithm find out the data sets that fit it. For example, in order to analyze the effects of the data characteristics (e.g., the dimensionality, scalability and distribution of the data) on the performance of a clustering algorithm, we often sample several data sets with different characteristics to respectively cluster each data set and compare their clustering results. However, different data characteristics often lead to different clustering difficulty. The larger the clustering difficulty on the data set is, the more possible it is that a clustering algorithm produces a clustering result with the poorer values of the validity functions. When comparing the performance of a clustering algorithm on different data sets, if we judge its effectiveness by only the values of the validity functions, the evaluation may be biased. Therefore, we need the normalization of the validity functions to obtain the relative position of each validity function value between the minimal and maximal values while a data set is provided. However, a critical challenge for the normalization is to obtain the ranges of the optimal solutions of the validity functions for a given data set, since the minimization or maximization of these functions in the relevant constraints is a class of nonconvex optimization problems whose solutions are unknown.

Therefore, we will address the three issues from a solution-space perspective. The major contributions are as follows:

(1) A generalized validity function is presented for evaluating the clustering results of categorical data. Furthermore, we apply the generalized validity function to analyze the generality and difference of the k-modes objective function, the category utility function and the information entropy function in the solution space.

(2) Due to the fact that the three existing validity functions are only based on the within-cluster information, a theoretical analysis is provided to answer whether

these validity functions can effectively evaluate the clustering results by using the within-cluster information only.

(3) A theoretical method is provided to obtain the upper and lower bounds of the three validity functions for a given data set. Furthermore, we present a normalization method based on the obtained bounds to reduce the effects of data characteristics on the performance of the clustering algorithm.

The organization of this paper is as follows: In Sect. 2, we review the notations of categorical data and the three cluster validity functions. Section 3 introduces a generalized validity function and discusses the relations of the three functions. Section 4 shows the effectiveness of the within-cluster information in evaluating the clustering results. In Sect. 5, we discuss the normalization issue for the validity functions. In Sect. 6, the experimental results illustrate the effectiveness of the above analysis. Finally, concluding remarks are given in Sect. 7.

2 Preliminaries

In this section, we will firstly introduce the notations of categorical data and then review the three widely-used validity functions for categorical data.

2.1 Categorical data

Huang (1997) provided the notations of categorical data which were introduced as follows: Let $U = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a set of n objects and $A = \{a_1, a_2, \ldots, a_m\}$ be a set of m attributes which are used to describe U. Each attribute a_j describes a domain of values, denoted by D_{a_j}, associated with a defined semantic and a data type. Here, only consider two general data types, numerical and categorical, and assume other types used in database systems can be mapped to one of these two types. The domains of attributes associated with these two types are called numerical and categorical, respectively. A numerical domain consists of real numbers. A domain D_{a_j} is defined as categorical if it is finite and unordered, i.e., $D_{a_j} = \{a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(n_j)}\}$ where n_j is the number of categories of attribute a_j for $1 \le j \le m$. For any $1 \le p \le q \le n_j$, either $a_j^{(p)} = a_j^{(q)}$ or $a_j^{(p)} \ne a_j^{(q)}$. For $1 \le i \le n$, an object $\mathbf{x}_i \in U$ is represented as a vector $[x_{i1}, x_{i2}, \ldots, x_{im}]$, where $x_{ij} \in D_{a_j}$, for $1 \le j \le m$. If each attribute in A is categorical, U is called a categorical data set.

2.2 Cluster validity functions

(1) *The k-modes objective function* was proposed by Huang (1997) which is an extension of the k-means objective function (MacQueen 1967) by using a simple matching dissimilarity measure for categorical objects, modes instead of means for clusters. The objective function attempts to minimize the dispersion of objects from the center in each cluster. The k-modes algorithm begins with an initial set of cluster centers and uses the alternating minimization method to obtain a local minimal solution for

the objective function. Several modified k-modes algorithms have been developed in the literature (Huang and Ng 1999; He et al. 2005; San et al. 2004; Ng et al. 2007; Huang et al. 2005; Bai et al. 2011, 2013). The k-modes objective function is defined as follows (Huang 1997):

$$F(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} d(\mathbf{z}_l, \mathbf{x}_i) \tag{1}$$

subject to

$$\begin{cases} w_{li} \in \{0, 1\}, & 1 \le l \le k, 1 \le i \le n, \\ \sum_{l=1}^{k} w_{li} = 1, & 1 \le i \le n, \\ 0 < \sum_{i=1}^{n} w_{li} < n, & 1 \le l \le k, \end{cases} \tag{2}$$

where

- n is the number of objects in U, $k(\le n)$ is a known number of clusters;
- $W = [w_{li}]$ is a k-by-n $\{0, 1\}$ matrix, w_{li} indicates whether $\mathbf{x}_i$ belongs to the lth cluster, $w_{li} = 1$ if $\mathbf{x}_i$ belongs to the lth cluster and 0 otherwise;
- $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\} \subseteq R$, where $R = D_{a_1} \times D_{a_2} \times \cdots \times D_{a_m}$ and $\mathbf{z}_l = [z_{l1}, z_{l2}, \ldots, z_{lm}]$ is the lth cluster prototype with categorical attributes $a_1, a_2, \ldots, a_m$;
- $d(\mathbf{z}_l, \mathbf{x}_i)$ is the simple matching dissimilarity measure between object $\mathbf{x}_i$ and the prototype $\mathbf{z}_l$ of the lth cluster which is defined as

$$d(\mathbf{z}_l, \mathbf{x}_i) = \sum_{j=1}^{m} \delta(z_{lj}, x_{ij}), \tag{3}$$

where

$$\delta(z_{lj}, x_{ij}) = \begin{cases} 1, & z_{lj} \ne x_{ij}, \\ 0, & z_{lj} = x_{ij}. \end{cases} \tag{4}$$

(2) *The category utility function* was introduced by Gluck and Corter (1985), which attempts to maximize the probability that two data objects in the same cluster obtain the same attribute values. This function has been applied in Cobweb (Fisher 1987), a tool for incremental clustering with categorical features, and related systems. In addition, the original framework has been expanded to both nonincremental clustering and mixed scale data. For instance, Mirkin (2001) provided extensions of the scoring function to situations with differently standardized and mixed scale data. CU is shown in the following equation (Gluck and Corter 1985):

$$CU(W) = \sum_{l=1}^{k} \frac{|c_l|}{n} \sum_{j=1}^{m} \sum_{q=1}^{n_j} \left[P\left(a_j^{(q)} | c_l\right)^2 - P\left(a_j^{(q)}\right)^2 \right], \tag{5}$$

subject to the same conditions as those in (2), where $P(a_j^{(q)}|c_l) = \frac{|c_{ljq}|}{|c_l|}$, $|c_{ljq}| = \sum_{i=1,x_{ij}=a_j^{(q)}}^{n} w_{li}$, $|c_l| = \sum_{i=1}^{n} w_{li}$, $P(a_j^{(q)}) = \frac{|a_j^{(q)}|}{n}$, $|a_j^{(q)}| = |\{\mathbf{x}_i|x_{ij} = a_j^{(q)}, \mathbf{x}_i \in U\}|$ and c_l is the lth cluster.

(3) *The information entropy function* E_{log} was proposed (Barbara et al. 2002), which uses the information-theoretic principles and the notion of entropy to measure the clustering results. The basic intuition is that groups of similar objects have lower entropy than those of dissimilar ones. Several algorithms have been developed (Andritsos et al. 2004; Barbara et al. 2002; Chen and Liu 2005, 2009), which are aimed at finding the optimal data partition that minimizes the information entropy function. Besides, Li et al. (2004) introduced the relations of the information entropy function, probabilistic mixture models and dissimilarity coefficients. The information entropy function is defined as follows (Barbara et al. 2002):

$$E_{log}(W) = -\sum_{l=1}^{k} |c_l| \sum_{j=1}^{m} \sum_{q=1}^{n_j} P\left(a_j^{(q)}|c_l\right) \log P\left(a_j^{(q)}|c_l\right), \tag{6}$$

subject to the same conditions as those in (2).

3 The generalization issue

3.1 The generalized validity function

Before discussing generalization of the validity functions for categorical data, we introduce a set of variables $V = \left[v_{lj}^{(q)}\right]$ that is a three-dimensional $\left(k \times m \times \max_{j=1}^{m} n_j\right)$ array. $v_{lj}^{(q)}$ is the representability of the qth categorical value of the jth attribute in the lth cluster, for $1 \le l \le k$, $1 \le j \le m$, $1 \le q \le n_j$. The larger $v_{lj}^{(q)}$ is, the more representability the categorical value $a_j^{(q)}$ has in the lth cluster. We constrain $\sum_{q=1}^{n_j} v_{lj}^{(q)} = 1$ and $0 \le v_{lj}^{(q)} \le 1 (1 \le q \le n_j)$ for each attribute $a_j (1 \le j \le m)$. For each cluster $(1 \le l \le k)$, we use $\mathbf{v}_l = \{v_{l1}, v_{l2}, \ldots, v_{lm}\}$, where $v_{lj} = \{v_{lj}^{(1)}, v_{lj}^{(2)}, \ldots, v_{lj}^{(n_j)}\}$, to summarize and characterize the lth cluster. For a categorical data set, V is a clustering model which can be used to predict the likelihood of an unseen object being a cluster member. If V has good predictive ability, it is thought to be good.

A clustering result should be composed of two parts: W (which is defined in (2)) and V. While analyzing the effectiveness of a clustering result, we need to evaluate not only the effectiveness of W but also that of V. Therefore, we will define a generalized validity function $F_g(W, V)$ whose values depend on the two set of variables W and V. The function $F_g(W, V)$ should have the following properties:

- Given V, a good W should be possible to make objects within clusters have very high similarity.

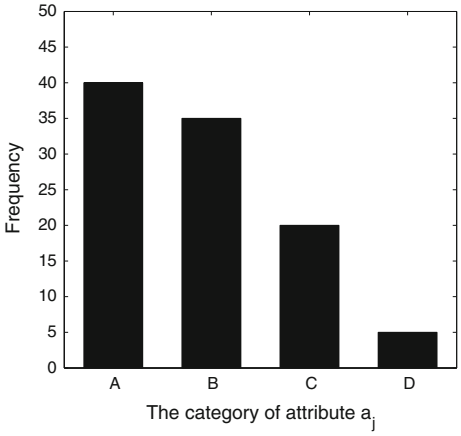

Fig. 1 An example of an attribute distribution in the cluster, where each *bar* corresponds to each categorical value

- Given W, a good V should be possible to objectively reflect the characteristics of each cluster.

Using the function $F_g(W, V)$ to find the best clustering result on a data set is like a dynamic game between W and V. The game scenario is described as follows: When V is given, it is wished that W can make each object belong to a cluster whose $\mathbf{v}_l$ has the best representative to the object. That can enhance the purity within clusters. When W is given, V should not blindly overestimate the purity within clusters but stimulate more categorical values to contribute to the identification of clusters and effectively avoid losing information. Let us consider the following example to demonstrate the problem. We suppose that there is a categorical attribute a_j which has four categorical values: 'A', 'B', 'C' and 'D', and a cluster c_l which contains 40 'A', 35 'B', 20 'C' and 5 'D' in attribute a_j. Figure 1 shows the categorical attribute distribution in cluster c_l. If we only select 'A' from the attribute domain to represent cluster c_l, other 60 % categorical values will be ignored. Therefore, we should use more categorical values to identify the cluster.

The generalized validity function and optimization problem can be written as follows:

$$\min F_g(W, V) = \sum_{l=1}^{k} \sum_{\mathbf{x}_i \in c_l}^{n} d_g(\mathbf{x}_i, c_l) + T \sum_{l=1}^{k} |c_l| \sum_{j=1}^{m} \sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2 \tag{7}$$

subject to

$$\begin{cases} w_{li} \in \{0, 1\}, & \sum_{l=1}^{k} w_{li} = 1, 1 < \sum_{i=1}^{n} w_{li} < n, \\ v_{lj}^{(q)} \in [0, 1], & \sum_{q=1}^{n_j} v_{lj}^{(q)} = 1, \end{cases} \tag{8}$$

where $T(\geq 0)$ is a parameter and $d_g(\mathbf{x}_i, c_l)$ is a dissimilarity measure between the object $\mathbf{x}_i$ and the lth cluster c_l defined as follows:

$$d_g(\mathbf{x}_i, c_l) = \sum_{j=1}^{m} \phi_{a_j}(\mathbf{x}_i, c_l) \tag{9}$$

with

$$\phi_{a_j}(\mathbf{x}_i, c_l) = 1 - v_{lj}^{(r)}, \quad if\ x_{ij} = a_j^{(r)}, \quad 1 \le r \le n_j. \tag{10}$$

Here, $\phi_{a_j}(\mathbf{x}_i, c_l)$ depends on $v_{lj}^{(r)}$, which is the representability of $a_j^{(r)}$ in the lth cluster. The larger $v_{lj}^{(r)}$ is, the more representability $a_j^{(r)}$ has in the lth cluster, the smaller the dissimilarity between $\mathbf{x}_i$ and c_l in the attribute a_j is. When the representability of $a_j^{(r)}$ is 1, $\phi_{a_j}(\mathbf{x}_i, c_l) = 0$.

In the generalized validity function (7), $\sum_{l=1}^{k} \sum_{\mathbf{x}_i \in c_l} d_g(\mathbf{x}_i, c_l)$ is the sum of the within-cluster dispersions that we want to minimize. $T \sum_{l=1}^{k} |c_l| \sum_{j=1}^{m} \sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2$ is used to stimulate more categorical values to contribute to the identification of clusters. In the following, we will analyze how the second term in (7) works. Let

$$\Gamma = T \sum_{l=1}^{k} |c_l| \sum_{j=1}^{m} \sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2$$

which are nonnegative and independent. Given W, minimizing (maximizing) the quantity is equivalent to minimizing (maximizing) each inner sum. We write the $l,\ j$th inner sum ($1 \le l \le k$ and $1 \le j \le m$) as

$$\psi_{lj} = |c_l| \sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2. \tag{11}$$

Since ψ_{lj} is a strictly convex function, the K–K–T necessary optimality condition is also sufficient. Thus, $\hat{v}_{lj}$ is an optimal solution of min ψ_{lj} subject to $\sum_{q=1}^{n_j} v_{lj}^{(q)} - 1 = 0$ if and only if there is some $\hat{\lambda}$ together with $\hat{v}_{lj}$ satisfying the following system of equations:

$$\begin{aligned} &\nabla_{v_{lj}} \tilde{\varphi}(v_{lj}, \lambda) = 0, \\ &1 - \sum_{q=1}^{n_j} v_{lj}^{(q)} = 0, \end{aligned} \tag{12}$$

where

$$\tilde{\varphi}(v_{lj}, \lambda) = |c_l| \sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2 + \lambda \left(\sum_{q=1}^{n_j} v_{lj}^{(q)} - 1 \right). \tag{13}$$

Note that

$$\frac{\partial \tilde{\varphi}(v_{lj}, \lambda)}{\partial v_{lj}^{(q)}} = 2|c_l| \left(v_{lj}^{(q)}\right) + \lambda, 1 \le q \le n_j. \tag{14}$$

From (12) and (14), we obtain that

$$\hat{v}_{lj}^{(q)} = \frac{1}{n_j}, \quad 1 \le q \le n_j. \tag{15}$$

The above analysis shows that, when $v_{lj}^{(q)}$ are the same for $1 \le q \le n_j$, ψ_{lj} achieves its minimum value given by

$$\min \sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2 = \frac{1}{n_j}.$$

We also know that

$$\sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2 \le \left(\sum_{q=1}^{n_j} v_{lj}^{(q)}\right)^2 = 1.$$

If only one of the $v_{lj}^{(q)}$ for $1 \le q \le n_j$ is nonzero, ψ_{lj} achieves the maximum value, i.e.,

$$\max \sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2 = 1.$$

Note that the smaller $\sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2$ is, the more categories the weights are assigned to. While giving W, we wish to minimize Γ to make more categorical values identify clusters.

Therefore, the two terms $\sum_{l=1}^{k} \sum_{\mathbf{x}_i \in c_l}^{n} d_g(\mathbf{x}_i, c_l)$ and $T \sum_{l=1}^{k} |c_l| \sum_{j=1}^{m} \sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2$ are integrated to the generalized validity function so that we can simultaneously minimize the within cluster dispersion and stimulate more categorical values to contribute to the identification of clusters.

The parameter T is used to balance which part plays a more important role in the minimization process of (7). The larger T is, the more the last term contributes in the optimization process and the smoother or fuzzier of the resulting V is. However, the values of T should not be too large. The reason is that when T is very large so that each element in v_{lj} is close to $1/n_j$. Next, we will discuss how to set the parameter T.

Let

$$\vartheta_{lj} = \sum_{q=1}^{n_j} |c_{ljq}| \left(1 - v_{lj}^{(q)}\right) + T|c_l| \sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2$$

for $1 \le l \le k$, $1 \le j \le m$. Then

$$F_g(W, V) = \sum_{l=1}^{k} \sum_{j=1}^{m} \vartheta_{lj}. \tag{16}$$

Thus, given $W = \hat{W}$, minimizing the function is equivalent to minimizing each ϑ_{lj}. Since ϑ_{lj} is a strictly convex function, the well-known K–K–T necessary optimization condition is also sufficient. Therefore, $\hat{v}_{lj}$ is an optimal solution if and only if there exists $\hat{\lambda}$ together with $\hat{v}_{lj}$ satisfying the following system of equations:

$$\begin{aligned} &\nabla_{v_{lj}} \tilde{\vartheta}_{l,j}(v_{lj}, \lambda) = 0, \\ &\sum_{q=1}^{n_j} v_{lj}^{(q)} = 1, \end{aligned} \tag{17}$$

where

$$\tilde{\vartheta}_{lj}(v_{lj}, \lambda) = \sum_{q=1}^{n_j} |c_{ljq}| \left(1 - v_{lj}^{(q)}\right) + T|c_l| \sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2 + \lambda \left(\sum_{q=1}^{n_j} v_{lj}^{(q)} - 1 \right). \tag{18}$$

We have

$$\frac{\partial \tilde{\theta}_{l,j}(\mathbf{v}_{lj}, \lambda)}{\partial v_{lj}^{(r)}} = 2T|c_l|v_{lj}^{(r)} - |c_{ljr}| + \lambda, 1 \le q \le n_j. \tag{19}$$

From (17) and (19), we obtain that F_g is minimized iff

$$v_{lj}^{(r)} = \frac{1}{2T} \frac{|c_{ljr}|}{|c_l|} + \frac{2T - 1}{2Tn_j} \tag{20}$$

for $1 \le l \le k$, $1 \le j \le m$, $1 \le r \le n_j$.

According to (20), we can rewrite

$$v_{lj}^{(r)} = f\left(\frac{|c_{ljr}|}{|c_l|}\right) = a\frac{|c_{ljr}|}{|c_l|} + b, \tag{21}$$

where $a = \frac{1}{2T}$ and $b = \frac{2T-1}{2Tn_j}$. We can see that $v_{lj}^{(r)}$ is linear related to $\frac{|c_{ljr}|}{|c_l|}$. If the relative frequency of a categorical value in a cluster is large, its representability in the cluster is high. Here, a and b are constants when given T. While setting $T = 1/2$,

$$v_{lj}^{(r)} = \frac{|c_{ljr}|}{|c_l|}, \tag{22}$$

which reduces the effect of a and b.

3.2 The relation of the validity functions

We know

$$\min F_g(W, V) = \min_V \min_W F_g(W, V) = \min_W \min_V F_g(W, V). \tag{23}$$

Therefore, let $Q(W) = \arg\min_V F_g(W, V)$ and $G(W) = F_g(W, Q(W))$. The optimization problem of the generalized validity function can be rewritten as follows:

$$\min F_g(W, V) = \min_W G(W). \tag{24}$$

While obtaining $Q(W)$ according to (22) and plug it into $F_g(W, V)$, we have

$$\begin{aligned} G(W) &= \min_V \sum_{l=1}^{k}\sum_{j=1}^{m}\sum_{q=1}^{n_j} |c_{ljq}| \left(1 - v_{lj}^{(q)}\right) + \frac{1}{2}\sum_{l=1}^{k} |c_l| \sum_{j=1}^{m}\sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2 \\ &= \sum_{l=1}^{k}\sum_{j=1}^{m}\sum_{q=1}^{n_j} |c_{ljq}| \left(1 - \frac{|c_{ljq}|}{|c_l|}\right) + \frac{1}{2}\sum_{l=1}^{k} |c_l| \sum_{j=1}^{m}\sum_{q=1}^{n_j} \left(\frac{|c_{ljq}|}{|c_l|}\right)^2 \\ &= \sum_{l=1}^{k}\sum_{j=1}^{m}\sum_{q=1}^{n_j} |c_{ljq}| - \sum_{l=1}^{k} |c_l| \sum_{j=1}^{m}\sum_{q=1}^{n_j} \left(\frac{|c_{ljq}|}{|c_l|}\right)^2 + \frac{1}{2}\sum_{l=1}^{k} |c_l| \sum_{j=1}^{m}\sum_{q=1}^{n_j} \left(\frac{|c_{ljq}|}{|c_l|}\right)^2 \\ &= mn - \frac{1}{2}\sum_{l=1}^{k} |c_l| \sum_{j=1}^{m}\sum_{q=1}^{n_j} \left(\frac{|c_{ljq}|}{|c_l|}\right)^2. \end{aligned} \tag{25}$$

According to (5), we rewrite the category utility function as follows:

$$CU(W) = \frac{1}{n}\sum_{l=1}^{k} |c_l| \sum_{j=1}^{m}\sum_{q=1}^{n_j} \left(\frac{|c_{ljq}|}{|c_l|}\right)^2 - P, \tag{26}$$

where

$$P = \frac{1}{n}\sum_{l=1}^{k}\sum_{j=1}^{m}\sum_{q=1}^{n_j} P\left(a_j^{(q)}\right)^2.$$

Given a data set U, P is a constant, which means that maximizing CU is equal to maximizing the first term.

From (25) and (26), we obtain

$$\min_V F_g(W, V) = mn - \frac{n}{2}\left(CU(W) + P\right). \tag{27}$$

Remark Equation (27) tells us that when setting $T = 1/2$, minimizing $F_g(W, V)$ is equivalent to maximizing $CU(W)$.

Next, we will discuss the relation between the generalized validity function and the information entropy function. Liang et al. (2002) applied the complementary entropy to measure the uncertainty of information tables and showed the equivalence of the complementary entropy and the Shannon entropy. Thus, we will replace the Shannon entropy with the complementary entropy. The information entropy function is redefined as follows:

$$E_c(W) = \sum_{l=1}^{k} |c_l| \sum_{j=1}^{m} \sum_{q=1}^{n_j} \frac{|c_{ljq}|}{|c_l|}\left(1 - \frac{|c_{ljq}|}{|c_l|}\right). \tag{28}$$

We have

$$E_c(W) = mn - \sum_{l=1}^{k} |c_l| \sum_{j=1}^{m} \sum_{q=1}^{n_j} \left(\frac{|c_{ljq}|}{|c_l|}\right)^2. \tag{29}$$

From (25) and (29), we obtain

$$\min_V F_g(W, V) = \frac{1}{2}E_c(W) + \frac{1}{2}mn. \tag{30}$$

Remark Equations (27) and (30) tell us that when setting $T = 1/2$, minimizing $F_g(W, V)$ is equivalent to maximizing $CU(W)$ and minimizing $E_c(W)$.

While we restrict $v_{lj}^{(q)} \in \{0, 1\}$ for $1 \le l \le k, 1 \le j \le m, 1 \le q \le n_j$, $\sum_{q=1}^{n_j} \left(v_{lj}^{(q)}\right)^2$ is equal to 1. Thus, the generalized validity function becomes

$$\begin{aligned} F_g(W, V) &= \sum_{l=1}^{k} \sum_{j=1}^{m} \left[\sum_{q=1}^{n_j} |c_{ljq}| \left(1 - v_{lj}^{(q)}\right) + \frac{1}{2}|c_l| \right] \\ &= \sum_{l=1}^{k} \sum_{j=1}^{m} \left(|c_l| - |c_{ljr}|\right) + \frac{1}{2}mn, \end{aligned} \tag{31}$$

where $1 \le r \le n_j$.

According to (31), we know that given $W = \hat{W}$, F is minimized iff

$$v_{lj}^{(q)} = \begin{cases} 1, & |c_{ljr}| = \max\limits_{q=1}^{n_j} |c_{ljq}|, \\ 0, & otherwise. \end{cases} \tag{32}$$

for $1 \le l \le k$, $1 \le j \le m$, $1 \le r \le n_j$.

When V is computed by (32), we find that for each attribute, only one categorical value with the relatively maximum frequency has the representability in the cluster. This means that (32) is equivalent to the technique for computing cluster representatives in the original k-modes algorithm. The dissimilarity measure d_g is equivalent to the simple matching dissimilarity measure, i.e.,

$$d_g(\mathbf{x}_i, c_l) = d(\mathbf{x}_i, \mathbf{z}_l). \tag{33}$$

Remark If $v_{lj}^{(q)} \in \{0, 1\}$ for $1 \le l \le k$, $1 \le j \le m$, $1 \le q \le n_j$, the generalized validity function is equivalent to the k-modes objective function, i.e.,

$$\min_V F_g(W, V) = \min_Z F(W, Z) + \frac{1}{2}mn. \tag{34}$$

Thus, we have

$$\min_V F_g(W, V) \le \min_Z F(W, Z) + \frac{1}{2}mn. \tag{35}$$

This indicates that if we assume that the optimal solution of F_g is the best clustering result, the optimal solution obtained by the k-modes algorithm can not guarantee to be the best clustering result.

By the above analysis, we know that while setting $T = 1/2$ for F_g, the relations can be obtained as follow:

$$\begin{aligned} G(W) &= \min_V F_g(W, V) = mn - \frac{n}{2}(CU(W) + P) \\ &= \frac{1}{2}E_c(W) + \frac{1}{2}mn \le \min_Z F(W, Z) + \frac{1}{2}mn. \end{aligned} \tag{36}$$

Equation (36) tells us that the category utility function is equivalent to the information entropy function in evaluating the clustering results, and the obtained optimal solution of the k-modes objective function on a data set is the upper bound of the optimal solution of the category utility function.

4 The effectiveness issue

Generally speaking, a good clustering result should have the following two characteristics:

- The within-cluster similarity is high.
- The between-cluster similarity is low.

However, the k-modes objective function, the category utility function and the information entropy function are based on the within-cluster information only, which are used to measure the within-cluster compactness. The between-cluster information, i.e., the between-cluster separation, is not considered. Therefore, we need analyze the effectiveness of these validity functions and address whether only using the within-cluster information can effectively evaluate the clustering results. In 1973, Duda and Hart (1973) showed that the total covariance matrix S_T (which is fixed given the data) is a combination of the within class covariance matrix S_W and the between class covariance matrix S_B, i.e.,

$$S_T = S_W + S_B, \tag{37}$$

where $S_T = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^{'}(\mathbf{x}_i - \bar{\mathbf{x}})$, $S_W = \sum_{l=1}^{k}\sum_{\mathbf{x}_i \in c_l}(\mathbf{x}_i - \mathbf{a}_l)^{'}(\mathbf{x}_i - \mathbf{a}_l)$, $S_B = \sum_{l=1}^{k}|c_l|(\mathbf{a}_l - \bar{\mathbf{x}})^{'}(\mathbf{a}_l - \bar{\mathbf{x}})$, $\bar{\mathbf{x}} = \sum_{i=1}^{n}\mathbf{x}_i/n$ and $\mathbf{a}_l = \sum_{\mathbf{x}_i \in c_l}\mathbf{x}_i/|c_l|$. The conclusion (37) is mainly aimed at the numerical data clustering. Besides, in (37), the between class covariance matrix S_B does not directly measure the difference between any two clusters but measure the difference between each cluster center and the center of the data set. The $tr(S_B)$ value dose not increase as the dissimilarity between clusters increases. Thus, S_B can not effectively reflect the difference between the distributions of any two clusters.

Next, we will analyze the effectiveness of the k-modes objective function, the category utility function and the information entropy function, and discuss the relation between them and the between-cluster information. According to Sect. 3, we have known the relation of the three validity functions. Thus, we will use the function $G(W)(T = 1/2)$ instead of them in the analysis below.

We have the sum of all pairwise dissimilarity of data objects in a data set X as follows:

$$\begin{aligned} TD(X) &= \sum_{i=1}^{n}\sum_{j=1}^{n} d(\mathbf{x}_i, \mathbf{x}_j) \\ &= \sum_{l=1}^{k}\sum_{\mathbf{x}_i \in c_l}\sum_{\mathbf{x}_j \in c_l} d(\mathbf{x}_i, \mathbf{x}_j) + 2\sum_{1 \le l < h \le k}\sum_{\mathbf{x}_i \in c_l}\sum_{\mathbf{x}_j \in c_h} d(\mathbf{x}_i, \mathbf{x}_j), \end{aligned}$$

where the function d is the simple matching dissimilarity measure mentioned in Sect. 2. We know that $TD(X)$ is a constant for a given data set.

We have

$$\begin{aligned} \sum_{\mathbf{x}_i \in c_l}\sum_{\mathbf{x}_j \in c_l} d(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{\mathbf{x}_i \in c_l}\sum_{\mathbf{x}_j \in c_l}\sum_{p=1}^{m}\delta(x_{ip}, x_{jp}) \\ &= \sum_{p=1}^{m}\sum_{\mathbf{x}_i \in c_l}\sum_{\mathbf{x}_j \in c_l}\delta(x_{ip}, x_{jp}) \end{aligned}$$

$$= \sum_{p=1}^{m} \sum_{q=1}^{n_j} |c_{lpq}| \left(|c_l| - |c_{lpq}| \right)$$

$$= m|c_l|^2 - \sum_{p=1}^{m} \sum_{q=1}^{n_p} |c_{lpq}|^2. \tag{38}$$

From (38), we know that

$$G(W) = \frac{1}{2} \sum_{l=1}^{k} \frac{1}{|c_l|} \sum_{\mathbf{x}_i \in c_l} \sum_{\mathbf{x}_j \in c_l} d(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{2} mn. \tag{39}$$

Furthermore, we can show that

$$\begin{aligned}
2 \sum_{\mathbf{x}_i \in c_l} \sum_{\mathbf{x}_j \in c_h} d(\mathbf{x}_i, \mathbf{x}_j) &= 2 \sum_{\mathbf{x}_i \in c_l} \sum_{\mathbf{x}_j \in c_h} \sum_{p=1}^{m} \delta(x_{ip}, x_{jp}) \\
&= 2 \sum_{p=1}^{m} \sum_{\mathbf{x}_i \in c_l} \sum_{\mathbf{x}_j \in c_h} \delta(x_{ip}, x_{jp}) \\
&= 2 \sum_{p=1}^{m} \sum_{q=1}^{n_j} |c_{lpq}| \left(|c_h| - |c_{hpq}| \right) \\
&= 2m|c_l||c_h| - 2 \sum_{p=1}^{m} \sum_{q=1}^{n_s} |c_{lpq}||c_{hpq}| \\
&= |c_l||c_h| \left(m - \sum_{p=1}^{m} \sum_{q=1}^{n_s} \frac{|c_{lpq}|^2}{|c_l|^2} \right) \\
&\quad + |c_l||c_h| \left(m - \sum_{p=1}^{m} \sum_{q=1}^{n_s} \frac{|c_{hpq}|^2}{|c_h|^2} \right) \\
&\quad + |c_l||c_h| \sum_{p=1}^{m} \sum_{q=1}^{n_s} \frac{|c_{lpq}|^2}{|c_l|^2} + |c_l||c_h| \sum_{p=1}^{m} \sum_{q=1}^{n_s} \frac{|c_{hpq}|^2}{|c_h|^2} \\
&\quad - 2 \sum_{p=1}^{m} \sum_{q=1}^{n_s} |c_{lpq}||c_{hpq}| \\
&= \frac{|c_h|}{|c_l|} \sum_{\mathbf{x}_i \in c_l} \sum_{\mathbf{x}_j \in c_l} d(\mathbf{x}_i, \mathbf{x}_j) + \frac{|c_l|}{|c_h|} \sum_{\mathbf{x}_i \in c_h} \sum_{\mathbf{x}_j \in c_h} d(\mathbf{x}_i, \mathbf{x}_j) \\
&\quad + |c_l||c_h| \sum_{s=1}^{m} \sum_{q=1}^{n_s} \left(\frac{|c_{lsq}|}{|c_l|} - \frac{|c_{hsq}|}{|c_h|} \right)^2.
\end{aligned} \tag{40}$$

According to (39) and (40), we have

$$TD(X) = nG(W) - \frac{mn^2}{2} + S(W), \tag{41}$$

where

$$S(W) = \sum_{1 \le l < h \le k} |c_l||c_h| \sum_{s=1}^{m} \sum_{q=1}^{n_s} \left(\frac{|c_{lsq}|}{|c_l|} - \frac{|c_{hsq}|}{|c_h|} \right)^2. \tag{42}$$

From the above equations, we can see that $G(W)$ is inversely proportional to $S(W)$. Next, we will analyze whether $S(W)$ contains the between-cluster information. According to (22), we can rewrite $S(W)$ as follows.

$$S(W, V) = \sum_{1 \le l < h \le k} |c_l||c_h| \sum_{s=1}^{m} \sum_{q=1}^{n_s} \left(v_{ls}^{(q)} - v_{hs}^{(q)} \right)^2, \tag{43}$$

subject to the same conditions as those in (8). In this equation, we can see that $\sum_{s=1}^{m} \sum_{q=1}^{n_s} \left(v_{ls}^{(q)} - v_{hs}^{(q)} \right)^2$ takes use of the difference between the clustering models of clusters c_l and c_h to reflect the between-cluster separation. For the two clusters, the closer the representability of each categorical value in c_l is to that in c_h, the larger the similarity between c_l and c_h is. $|c_l||c_h|$ is the weight of the two-clusters separation. Let us consider the following example to demonstrate how the function S reflects the between-cluster separation. Assume that there are three clusters (c_1, c_2 and c_3) and an attribute "color"(including three categorical values: 'Red', 'Yellow' and 'Blue'). Table 1 shows the representability of these categorical values in different clusters. According to Table 1, we can see that the representability of the categorical values in c_1 is very close to c_2, compared to that in c_3. Thus, in terms of the attribute "color", the similarity between c_1 and c_2 is higher than c_2 and c_3.

According to the above analysis, we see that the function S can effectively reflect the difference between clusters. For a clustering result, the larger its function S value is, the larger its between-cluster separation is. According to (41), we also know that minimizing $G(W)$ is equivalent to maximizing $S(W)$, due to the fact that $TD(X)$ is a constant for a given data set. This tells us that while using the function G to obtain a clustering result with the high within-cluster similarity, we don't ignore the between-cluster separation. The obtained result has both high within-cluster similarity and between-cluster dissimilarity.

Table 1 The representability of these categorical values in different clusters

	Red	Yellow	Blue
c_1	0.7	0.1	0.2
c_2	0.75	0	0.25
c_3	0.1	0.8	0.1

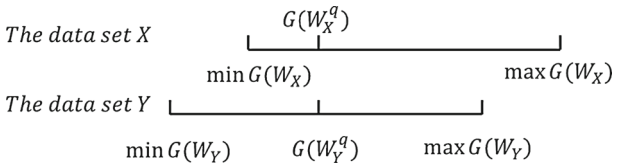

Fig. 2 Comparing clusterings of different data sets

5 The normalization issue

According to Sect. 3, we have known the relation of the k-modes objective function, category utility function and information entropy function. Therefore, we will make use of the function $G(W)(T = 1/2)$ to discuss the normalization issue of the three validity functions. Different data sets have different data characteristics, and thus have different clustering difficulties. In general, while we select a validity function to evaluate a clustering result, the more it is difficult in clustering the data set, the more it is possible that a clustering algorithm generates a clustering result with a bad value. If we judge its effectiveness by only the value of the validity function, the evaluation may be biased. Let us consider the following example to demonstrate the problem. We suppose that there are two data sets: X and Y. A specifical algorithm is used to cluster X and Y and obtain the clustering results W_X^q and W_Y^q, respectively. $minG(W_X)$ and $maxG(W_X)$ are the minimum and maximum values of the validity function G on the data set X. $minG(W_Y)$ and $maxG(W_Y)$ are the minimum and maximum values of the validity function G on the data set Y. According to Fig. 2, we see that $G(W_X^q)$ is equal to $G(W_Y^q)$. However, this doesn't illustrate that the specifical algorithm has the same performance on the data sets X and Y. We can find that $\frac{G(W_X^q)-minG(W_X)}{maxG(W_X)-minG(W_X)}$ is smaller than $\frac{G(W_Y^q)-minG(W_Y)}{maxG(W_Y)-minG(W_Y)}$. This indicates that the performance of the algorithm on X is better than on Y.

Thus, normalization is critical when it is used to compare clusterings of different data sets by a specific clustering algorithm. It represents the relative position of the original function value between the minimal and maximal values. However, the minimization and maximization of the function G in the constraints in (2) are unknown. Therefore, we will analyze the lower $\underline{G}(W)$ and upper $\overline{G}(W)$ bounds of the function G and use them to compute $NG(W)$, instead of the minimal and maximal solutions of the function $G(W)$. The normalized function G is defined as follows:

$$NG(W) = \frac{G(W) - \underline{G}(W)}{\overline{G}(W) - \underline{G}(W)}. \tag{44}$$

In the following, we will introduce how to obtain the bounds of the function G for a data set.

5.1 The lower bound of the function G

Let $S = [s_{ij}]$ be a n by n matrix, where $s_{ij} = m - d(\mathbf{x}_i, \mathbf{x}_j)$ and $H = [h_{il}]$ be a n by k matrix, where

$$h_{il} = \begin{cases} \frac{1}{\sqrt{|c_l|}}, & \mathbf{x}_i \in c_l, \\ 0, & otherwise. \end{cases} \tag{45}$$

We have

$$\frac{1}{|c_l|} \sum_{\mathbf{x}_i \in c_l} \sum_{\mathbf{x}_j \in c_l} d(\mathbf{x}_i, \mathbf{x}_j) = m|c_l| - h_l^{'} S h_l.$$

Thus,

$$\begin{aligned} G(W) &= \frac{1}{2} \sum_{l=1}^{k} \frac{1}{|c_l|} \sum_{\mathbf{x}_i \in c_l} \sum_{\mathbf{x}_j \in c_l} d(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{2} mn \\ &= mn - \frac{1}{2} \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} h_{il} h_{jl} \\ &= mn - \frac{1}{2} tr(H^{'} S H). \end{aligned} \tag{46}$$

According to (46), we know that minimizing $G(W)$ becomes maximizing $tr(H^{'} S H)$. If we relax the restriction that h_{ij} must take discrete values, and allow the entries of the matrix H to take arbitrary real values. Then the relaxed problem becomes:

$$\max tr(H^{'} S H)$$

subject to

$$H^{'} H = I.$$

Let $eig(S) = \{\lambda_1(S), \lambda_2(S), \ldots, \lambda_n(S)\}$ be the n eigenvalues of the matrix S. These eigenvalues are assumed to be in decreasing order, that is, $\lambda_1(S) \geq \lambda_2(S) \geq \cdots \geq \lambda_n(S)$ where $\lambda_l(S)$ denotes the lth largest eigenvalue of S. Therefore, the optimal value of the function G satisfies the lower bound

$$\underline{G}(W) = nm - \frac{1}{2} \sum_{l=1}^{k} \lambda_l(S) \leq G(W). \tag{47}$$

According to (36) and (47), we can directly obtain the lower bounds of the function CU, E_c and F by the value of $\underline{G}$.

Table 2 The data sets from UCI

Data set	Objects	Attributes	Classes
Soybean	47	35	4
Lung cancer	32	56	3
Zoo	101	16	7
Hepatitis	155	19	2
Heart disease	303	8	2
Voting	435	16	2
Credit approval	690	8	2
Dermatology	366	33	6
Breast cancer	699	9	2
Letters (E,F)	1,543	16	2
DNA	3,190	60	3
Mushroom	8,124	22	2

Table 3 Notation for the contingency table for comparing two partitions

$C \backslash P$	p_1	p_2	$\cdots$	$p_{k'}$	*Sums*
c_1	n_{11}	n_{12}	$\cdots$	$n_{1k'}$	b_1
c_2	n_{21}	n_{22}	$\cdots$	$n_{2k'}$	b_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_k	n_{k1}	n_{k2}	$\cdots$	$n_{kk'}$	b_k
Sums	d_1	d_2	$\cdots$	$d_{k'}$	

5.2 The upper bound of the function G

We can rewrite the function G as follows:

$$G(W) = mn - \frac{1}{2}\sum_{l=1}^{k}\sum_{j=1}^{m}\sum_{q=1}^{n_j}\frac{f_{ljq}^2}{y_l} \tag{48}$$

where

$$f_{ljq} = |c_{ljq}|,\ 1 \le l \le k, 1 \le j \le m, 1 \le q \le n_j,$$
$$y_l = |c_l|,\ 1 \le l \le k.$$

According to (48), we know that maximizing $G(W)$ becomes minimizing the second term. If we relax the restriction that f_{ljq} and y_l must take discrete values, then the relaxed problem becomes:

$$\min \sum_{l=1}^{k}\sum_{j=1}^{m}\sum_{q=1}^{n_j}\frac{f_{ljq}^2}{y_l} \tag{49}$$

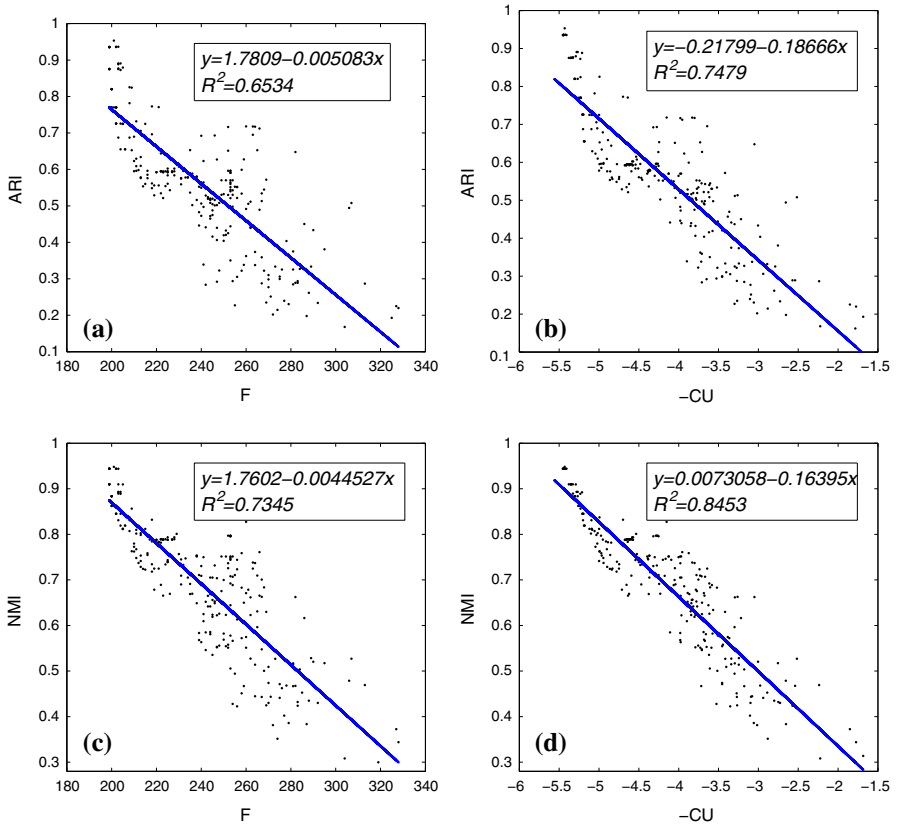

Fig. 3 On the soybean data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

subject to

$$\begin{cases} \sum_{l=1}^{k} f_{ljq} = |a_j^{(q)}|, 1 \le j \le m, 1 \le q \le n_j, \\ \sum_{q}^{n_j} f_{ljq} - y_l = 0, 1 \le l \le k, 1 \le j \le m, \\ \sum_{l=1}^{k} y_l = n, \\ f_{ljq} \ge 0, y_l > 0, 1 \le l \le k, 1 \le j \le m, 1 \le q \le n_j. \end{cases} \tag{50}$$

From (49) and (50), we know that the relaxed problem is an linearly constrained convex programming problem. We can obtain the minimum solutions, $\hat{f}_{ljq}$ and $\hat{y}_l$ for $1 \le l \le k, 1 \le j \le m, 1 \le q \le n_j$, by the existing optimization algorithms which are available from the Matlab software package. Therefore, the value of the function G satisfies the upper bound

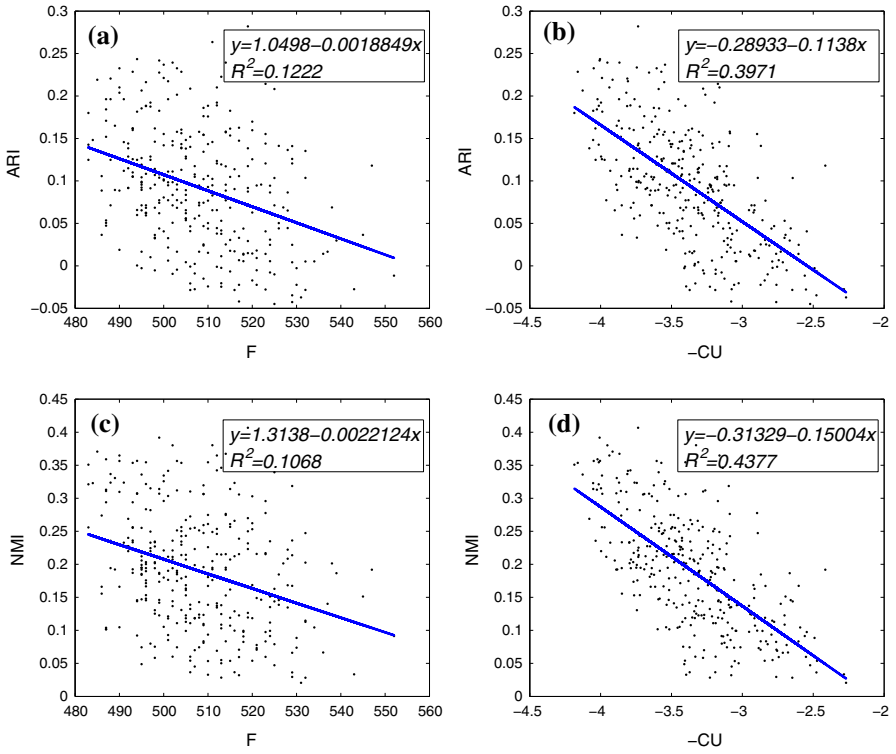

Fig. 4 On the lung cancer data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

$$G(W) \leq \overline{G}(W) = mn - \frac{1}{2}\sum_{l=1}^{k}\sum_{j=1}^{m}\sum_{q=1}^{n_j}\frac{\hat{f}_{ljq}^2}{\hat{y}_l}. \tag{51}$$

According to the relation of the functions CU, E_c and G, we can directly obtain the upper bound of CU and E_c by the upper bound of the function G. However, since the k-modes objective function F is not equivalent to the function G, we need to discuss how to obtain the upper bound of the function F.

For the k-modes objective function F, we have

$$\begin{aligned}\min_{Z} F(W, Z) &= \sum_{l=1}^{k}\sum_{j=1}^{m}\left(|c_l| - \max_{q=1}^{n_j}|c_{ljq}|\right)\\ &= mn - \sum_{l=1}^{k}\sum_{j=1}^{m}\max_{q=1}^{n_j}|c_{ljq}|.\end{aligned} \tag{52}$$

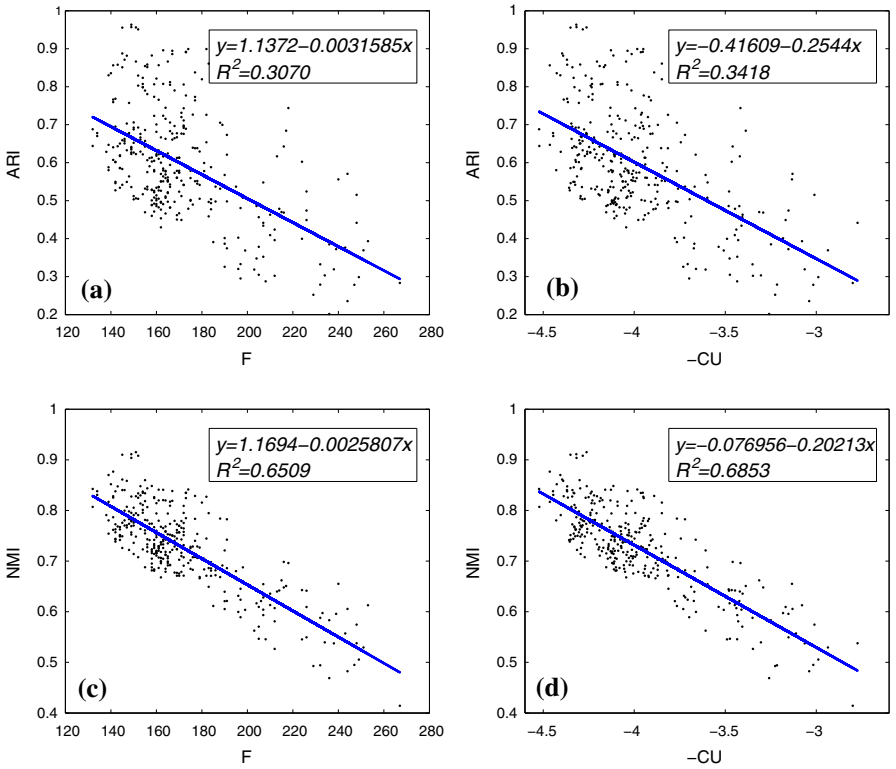

Fig. 5 On the zoo data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

According to (52), we know that maximizing $\min_Z F(W, Z)$ is equivalent to minimizing $\sum_{l=1}^{k}\sum_{j=1}^{m}\max_{q=1}^{n_j}|c_{ljq}|$. Hence, we can transform the computation of the upper bound of $\min_Z F(W, Z)$ into solving the following problem:

$$\min \sum_{l=1}^{k}\sum_{j=1}^{m}\max_{q=1}^{n_j} f_{ljq} \tag{53}$$

subject to the same constraints in (50). Because of $\sum_{q=1}^{n_j}\frac{f_{ljq}}{y_l} = 1$ for $1 \le l \le k$ and $1 \le j \le m$, we know

$$\sum_{q=1}^{n_j}\left(\frac{f_{ljq}}{y_l}\right)^2 \le \max_{q=1}^{n_j}\frac{f_{ljq}}{y_l}. \tag{54}$$

Therefore, we have

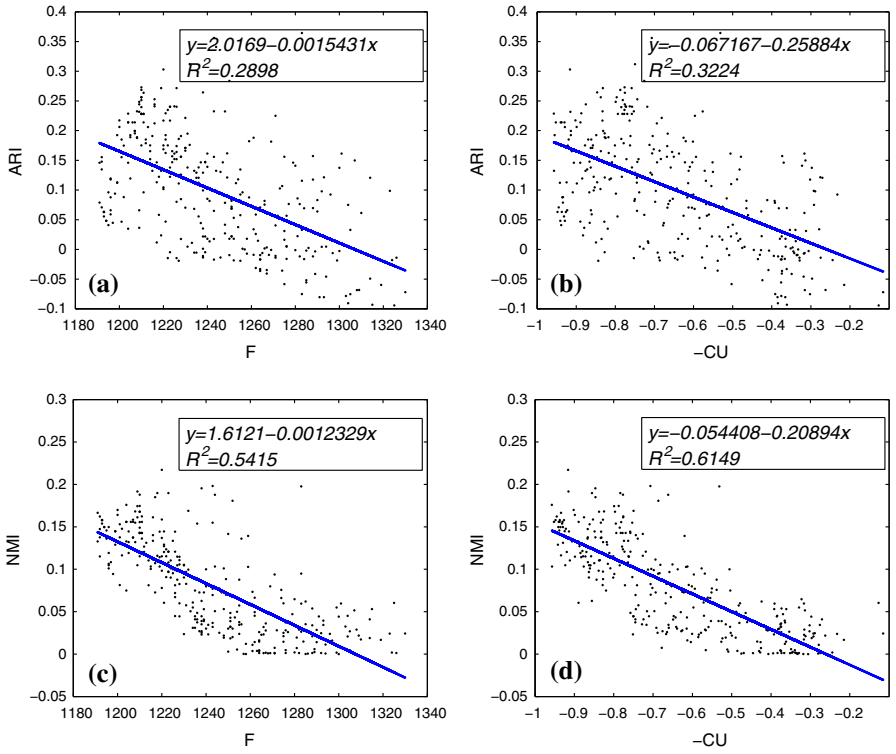

Fig. 6 On the hepatitis data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

$$\sum_{l=1}^{k}\sum_{j=1}^{m}\sum_{q=1}^{n_j}\frac{f_{ljq}^2}{y_l} \le \sum_{l=1}^{k}\sum_{j=1}^{m}\max_{q=1}^{n_j} f_{ljq}. \tag{55}$$

From (55), we see that the upper bound of the k-modes objective function can be also obtained by using the minimum solution of the problem (49).

6 Experimental analysis

In this section, we will use 12 data sets from the UCI Machine Learning Repository (UCI 2012) Kindly provide link (shown in Table 2). In our experiments, we apply the k-modes algorithm for clustering the input data sets and set the number of clusters k as the "true" cluster number for each given data set. Furthermore, we employ the adjusted rand index (ARI) and the normalized mutual information (NMI) (Yang 2004) to test the effectiveness of the above validity functions. Both ARI and NMI are external validity functions which attempts to measure similarity between the clustering results and the "true" partition determined by the class label information. Given a set U

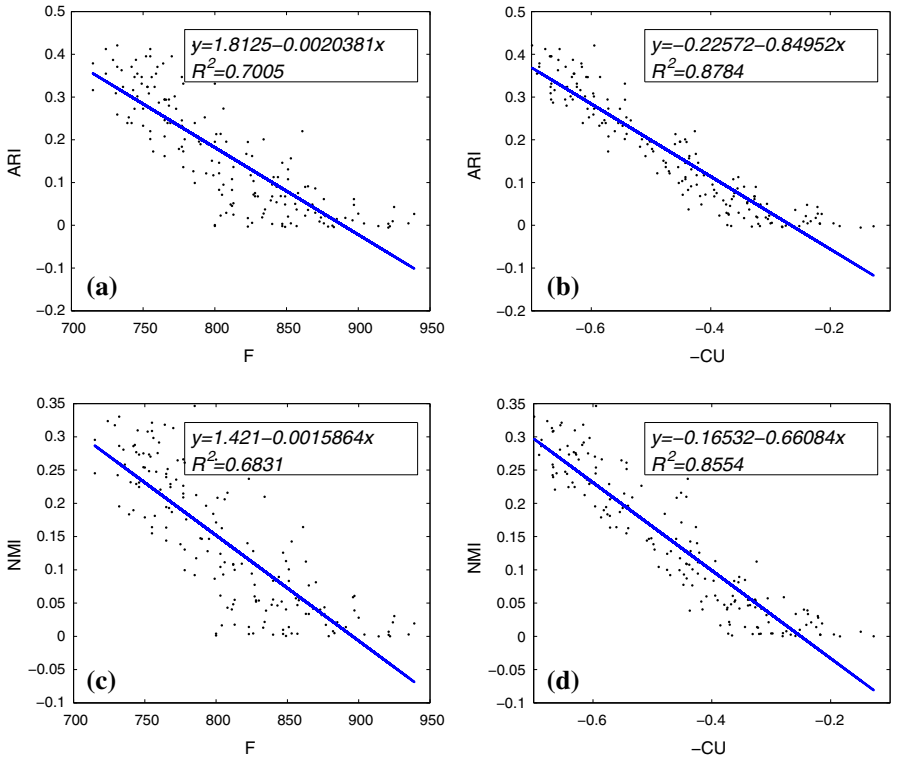

Fig. 7 On the heart disease data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

of n data objects and two groupings (e.g. clusterings) of these objects, namely $C = \{c_1, c_2, \ldots, c_k\}$ and $P = \{p_1, p_2, \ldots, p_{k'}\}$, the overlappings between C and P can be summarized in a contingency table where n_{ij} denotes the number of common objects of groups c_i and p_{lj}: $n_{ij} = |c_i \cap p_{lj}|$.

The adjusted rand index is defined as $AdjustedIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex}$, more specifically,

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{b_i}{2}\sum_j \binom{d_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{b_i}{2} + \sum_j \binom{d_j}{2}\right] - \left[\sum_i \binom{b_i}{2}\sum_j \binom{d_j}{2}\right]/\binom{n}{2}}$$

where n_{ij}, b_i, d_j are values from the contingency table (Table 3). The normalized mutual information is defined as

$$NMI = \frac{2\sum_{i=1}^{k}\sum_{j=1}^{k'}\frac{n_{ij}}{n}\log\frac{n_{ij}n}{b_i d_j}}{\sum_{i=1}^{k} -\frac{b_i}{n}\log\frac{b_i}{n} + \sum_{j=1}^{k'} -\frac{d_j}{n}\log\frac{d_j}{n}}$$

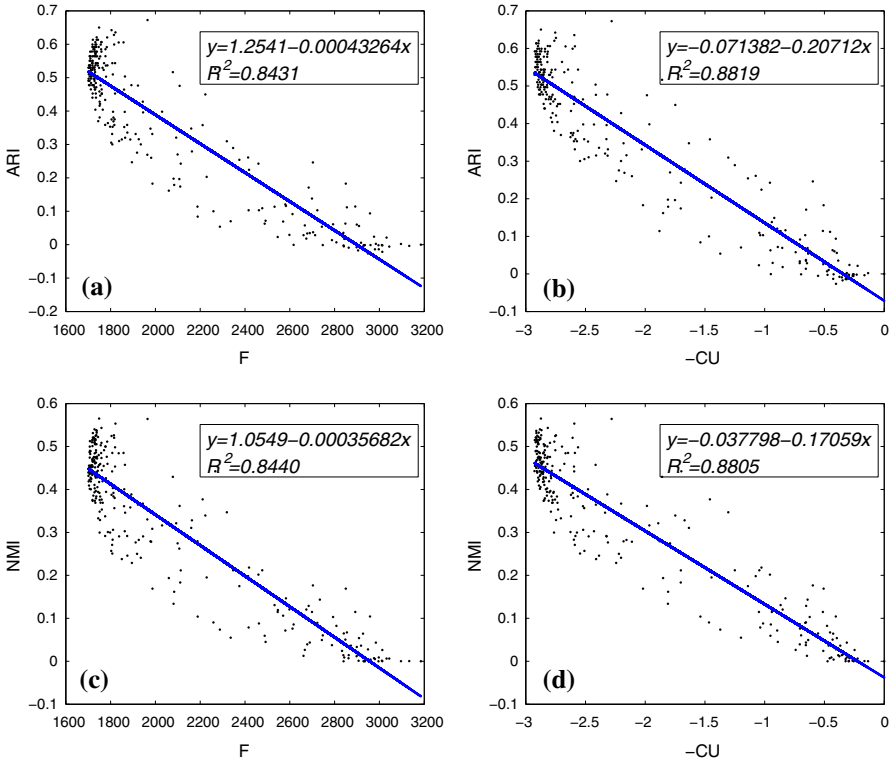

Fig. 8 On the voting data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

Both ARI and NMI take the values within the interval [0,1]. The closer a clustering result is to the "true" partition, the higher the ARI and NMI values are. The two validity functions have been widely used to evaluate the performance of clustering algorithms. For an internal validity function, we can not directly test its effectiveness in evaluating the clustering results, due to its lack of the class label information. Thus, we need to employ the external validity functions to indirectly test its effectiveness. If the evaluation results of an internal validity function are close to those of ARI and NMI, it is thought to be effective in evaluating the clustering results. Therefore, we will make use of the linear regression to analyze the correlations or consistencies between the external and internal validity functions. The coefficient of determination R^2 is employed to reflect the "goodness-of-fit" of linear regression. For a linear regression fitting line about an internal validity function and an external validity function, the larger the R^2 value is, the higher their correlation or consistency in evaluating clustering results is, and the more the internal validity function can effectively evaluate clustering results.

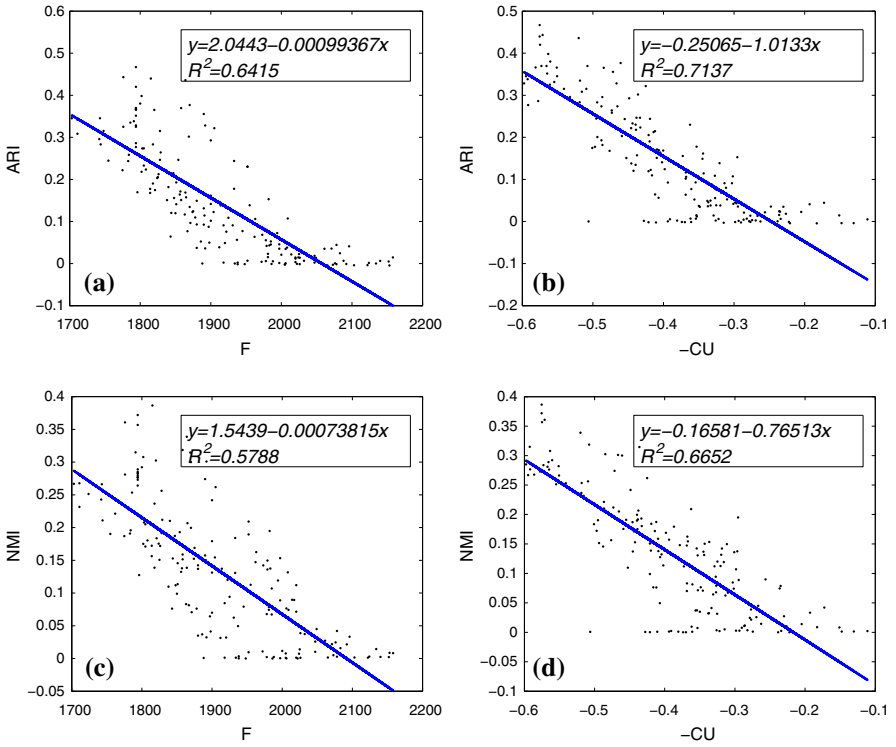

Fig. 9 On the credit approval data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

6.1 The comparison of the validity functions

Due to the fact that the category utility function CU is equivalent to the information entropy function E_{log} in evaluating clustering results, we only test the performance of the k-modes objective function F and the category utility function CU. In the experiment, we use the 12 data sets shown in Table 2 and compare the effectiveness of these functions from the two aspects.

The one is to use the linear regression analysis to illustrate the performance of these internal validity functions. We first employ the k-modes algorithm to randomly produce the 100 clustering results on each given data set, respectively. Based on the 100 clustering results on each data set, we analyze the correlations between the external validity functions (ARI and NMI) and the internal validity functions (F and CU). Figures 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14 illustrate the linear regression analysis of ARI and F, ARI and $-CU$, NMI and F, NMI and $-CU$ on each of the 12 data sets, respectively. In these figures, we see that the values of F and $-CU$ are inversely proportional to those of ARI and NMI in these linear regression fitting lines. We also observe that the R^2 values of F are always smaller than those of $-CU$ on these data

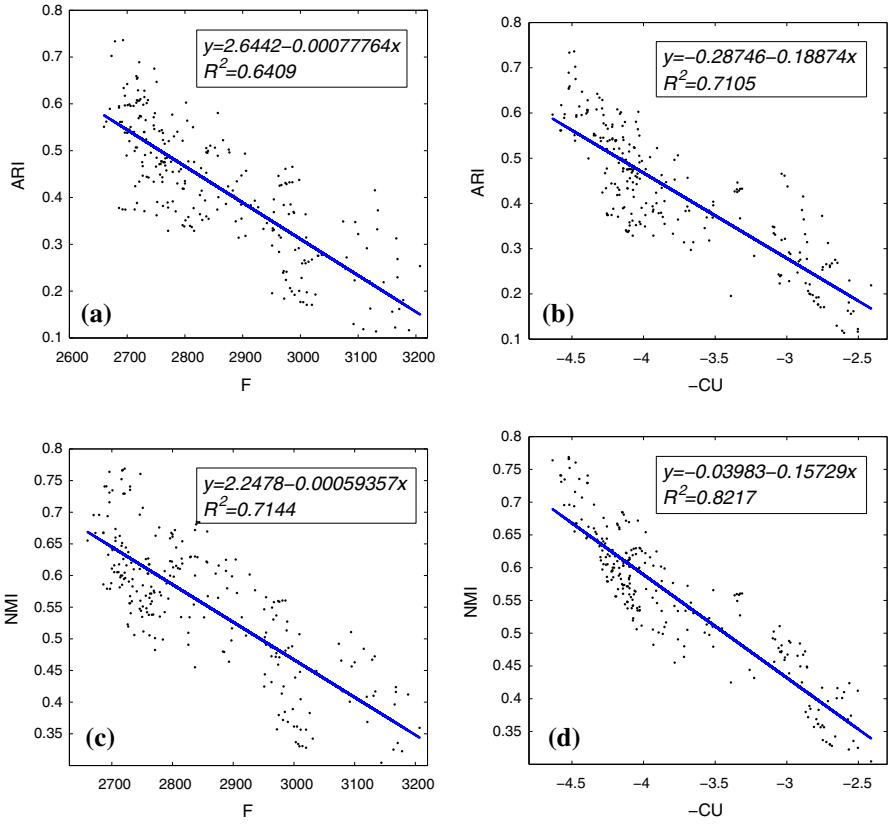

Fig. 10 On the dermatology data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

sets. This illustrates that the function CU has the more consistencies with the external validity functions ARI and NMI than the function F in evaluating the clustering results. Besides, according to Fig. 13, we not only see that the R^2 values of F are smaller than those of $-CU$ on the DNA data set, but also observe that the correlations between the internal and external validity functions are very weak on it. In the case, we cannot determine which one of these internal validity functions is bad. We only conclude that they may be not appropriate to find out the cluster structure of the true class labels on the data set. The case is inevitable. Because the internal validity functions are defined based on users' subjective assumptions, and the "true" class labels of a data set are obtained based on users experiences.

The other is to compare the effectiveness of clustering results produced by optimizing the functions F and CU, respectively. The comparison is carried out on the 12 data sets. We employ the alternating optimization (AO) method to optimize the functions F and CU. The AO method can obtain the local optimal solutions for an optimization problem. It requires an initial solution to start and end up with different

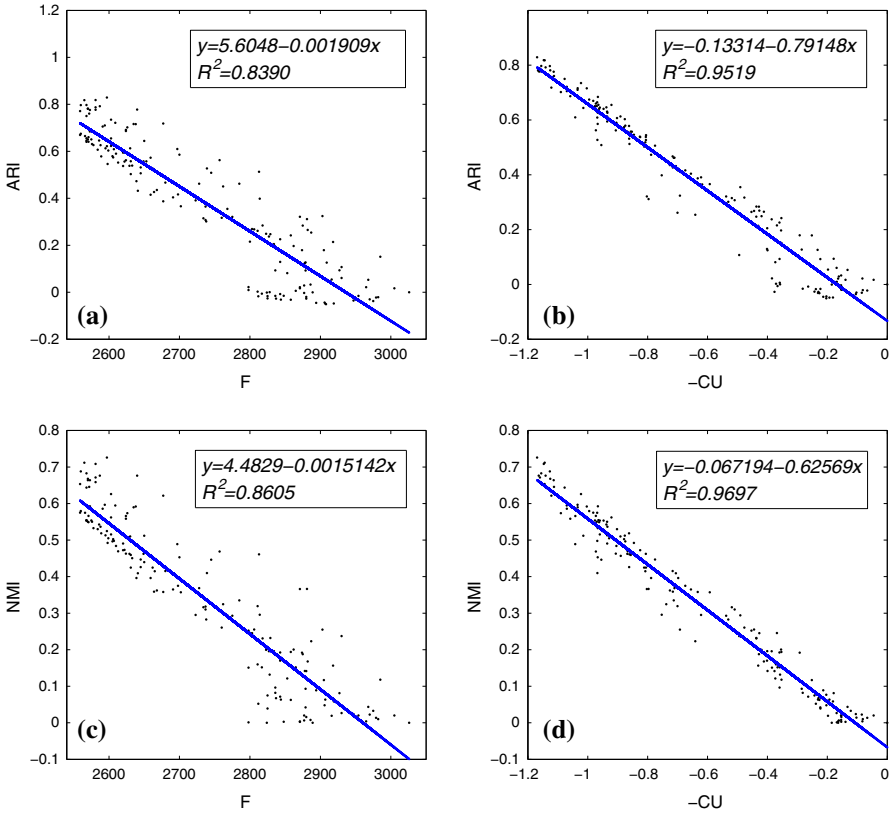

Fig. 11 On the breast cancer data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

finial solutions from different initial solutions in the process of solving an optimization problem. Therefore, on each of the data sets, we randomly select 100 initial solutions and carry out 100 runs of optimizing the functions F and CU, respectively, by the AO method. In each run, the same initial solution is used in optimizing both the functions. Table 4 shows the average and maximum ARI and NMI values of the clustering results produced by different objective functions on each of the 12 data sets. According to the table, we see that the function CU are better than the function F in being as the objective functions to cluster most of the data sets, except the zoo and credit approval data sets. On the zoo data set, the function F is better than the function CU according to the index values from Table 4. On the credit approval data set, the function F is better in the average ARI and NMI values but less in the maximum ARI and NMI values than the function CU. These indicate that any validity function is defined based on certain assumptions which impossibly conform to characteristics of all the data sets.

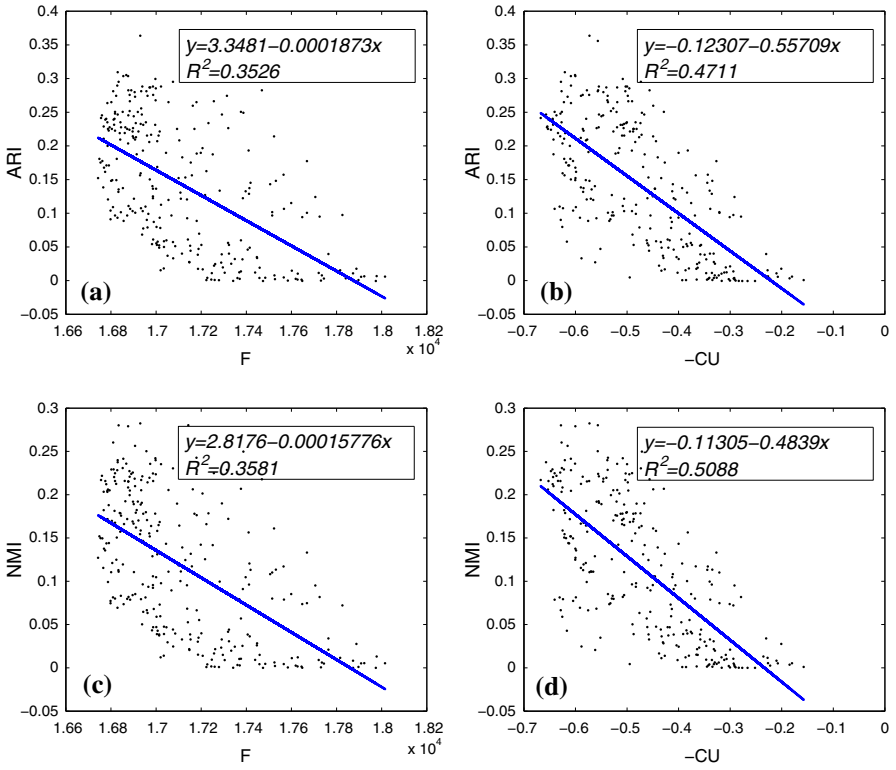

Fig. 12 On the letters data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

6.2 The effectiveness of the normalized validity functions

In the subsection, we test the effectiveness of the normalized function NG in evaluating the clustering results from different data sets. The test is carried out on the 12 data sets (shown in Table 2). On each of the data sets, we first use the k-modes algorithm to randomly produce the 100 clustering results, and compute the values of the clustering results for ARI, NMI and G. Furthermore, we use the following two methods to compare the effectiveness of the clustering results from different data sets. The one is to compute the average one of the 100 clustering results on each of the 12 data sets and compare their effectiveness, which is denoted as AVG. Table 5 shows some external and internal validity function values produced by the AVG method on the 12 data sets. The other is to select the one with the minimum G value of the 100 clustering results on each of the 12 data sets and compare their effectiveness, which is denoted as MIN. Table 6 shows some external and internal validity function values produced by the MIN method on the 12 data sets. In the following, we make use of the linear regression and the coefficient of determination R^2 to analyze the correlations between

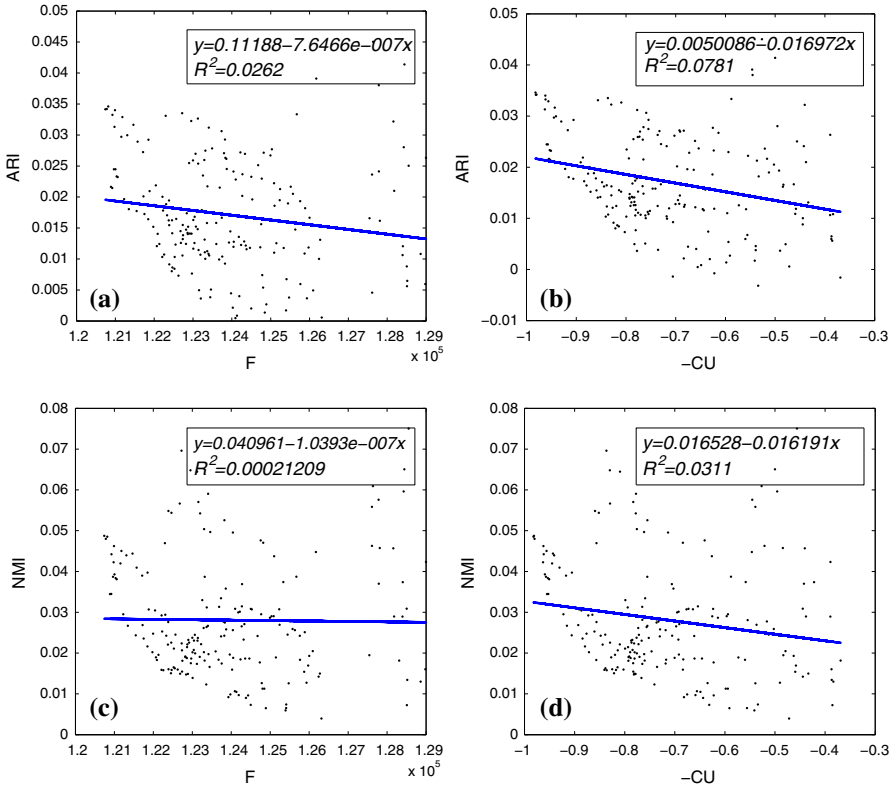

Fig. 13 On the DNA data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

the external (ARI and NMI) and internal (the original and normalized functions G) measures based on the contents of Tables 4 and 5. For a normalization method, if the normalized function NG enhances the correlation between the external measures and the original function G, it is thought to be effective. Table 7 shows the lower and upper bounds of the function G on the 12 data sets computed by our proposed method. In the experiment, we not only test the changes of the correlations while the normalized function NG is used, but also compare the effectiveness of the normalized function NG with a traditional normalized function TNG often used in bioinformatics. The function TNG is described as

$$TNG(W) = \frac{G(W)}{\sum_{W \in \Omega} G(W)/N},$$

where Ω is a set of the N clustering results on a data set.

Figure 15 shows the comparisons of the correlations between the external functions (ARI and NMI) and the internal functions (G, TNG and NG) on the 12 data sets, according to the content of Table 5. In Fig. 15b, e, we see that all the function TNG

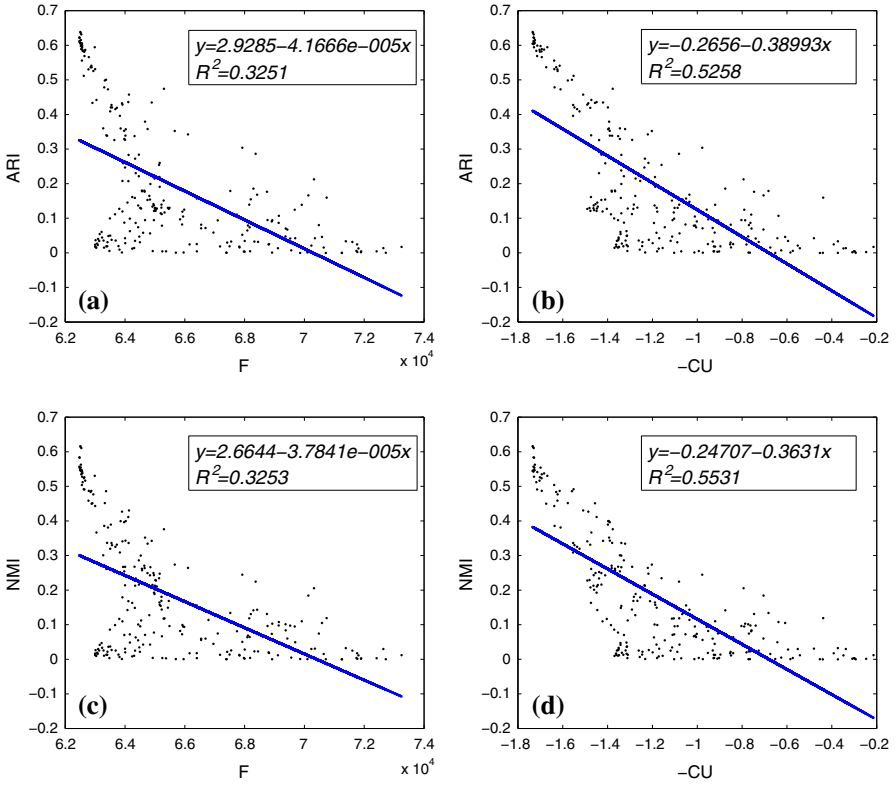

Fig. 14 On the mushroom data, **a** the k-modes objective function F values against the ARI values, **b** the negative category utility function $-CU$ values against the ARI values, **c** the k-modes objective function F values against the NMI values, **d** the negative category utility function $-CU$ values against the NMI values

values are equal to 1. This indicates that the function TNG can not compare the effectiveness of the average clustering results from the different data sets. Therefore, we also compare the correlations between the external functions (ARI and NMI) and the internal functions (G, TNG and NG) on the 12 data sets, according to the content of Table 6. The comparisons are shown in Fig. 16. According to the linear regression fitting lines in Figs. 15 and 16 (except Fig. 15b, e) , we see that the values of the function G , TNG and NG are inversely proportional to those of ARI and NMI. Furthermore, we observe that the R^2 values of NG and TNG are larger than those of G. This illustrates that the normalized functions NG and TNG enhance the correlations between the external measures and the function G. We also see that the R^2 values of NG are largest among those of G, TNG and NG. This shows that the normalized function NG performs the original function G and the normalized function TNG in comparing the effectivenss of the clustering results from the different data sets.

Table 4 The comparisons of optimizing the functions F and CU on the 12 data sets

Data set\ index	Internal function	Average ARI	Maximum ARI	Average NMI	Maximum NMI
Soybean	F	0.7088	1.0000	0.8188	1.0000
	CU	0.7479	1.0000	0.8552	1.0000
Lung cancer	F	0.0864	0.2537	0.1831	0.3942
	CU	0.1024	0.2912	0.2050	0.4088
Zoo	F	0.6312	0.9045	0.7509	0.8800
	CU	0.5973	0.8674	0.7491	0.8735
Hepatitis	F	0.1161	0.2250	0.1217	0.1744
	CU	0.1772	0.2846	0.1851	0.2226
Heart disease	F	0.2781	0.3869	0.2987	0.3181
	CU	0.3779	0.4382	0.2987	0.3456
Voting	F	0.5220	0.5368	0.4519	0.4866
	CU	0.5673	0.5779	0.4842	0.4947
Credit approval	F	0.2102	0.4432	0.1858	0.3607
	CU	0.1777	0.5036	0.1585	0.4259
Dermatology	F	0.4505	0.6945	0.5703	0.7800
	CU	0.6690	0.9531	0.8128	0.9461
Breast cancer	F	0.5137	0.7967	0.4667	0.6881
	CU	0.7952	0.7983	0.7220	0.7249
Letters(E,F)	F	0.1624	0.3907	0.1320	0.3050
	CU	0.3723	0.6869	0.3182	0.6152
DNA	F	0.0172	0.0485	0.0339	0.0899
	CU	0.4993	0.6431	0.4720	0.5895
Mushroom	F	0.2526	0.6144	0.2449	0.5613
	CU	0.4029	0.6229	0.3963	0.5837

Table 5 The validity function values produced by the AVG method on the 12 data sets

Data set\index	ARI	NMI	G	TNG	NG
Soybean	0.7086	0.8184	957.4000	1.0000	0.1686
Lung cancer	0.1089	0.2126	1.2340e+03	1.0000	0.3437
Zoo	0.6318	0.7510	927.6500	1.0000	0.2439
Hepatitis	0.1110	0.1254	2.2300e+03	1.0000	0.4313
Heart disease	0.2776	0.2286	1729.8000	1.0000	0.4121
Voting	0.5366	0.4497	4.6953e+03	1.0000	0.1817
Credit approval	0.2232	0.1915	4.4246e+03	1.0000	0.4206
Dermatology	0.4507	0.5705	7.9727e+03	1.0000	0.3111
Breast cancer	0.5137	0.4667	4.9093e+03	1.0000	0.4124
Letters (E, F)	0.2467	0.2264	2.2240e+04	1.0000	0.5344
DNA	0.0170	0.0324	1.6529e+05	1.0000	0.6007
Mushroom	0.2526	0.2449	1.2991e+05	1.0000	0.2880

Table 6 The validity function values produced by the MIN method on the 12 data sets

Data set\index	ARI	NMI	G	TNG	NG
Soybean	1.0000	1.0000	938.9400	0.9807	0.0317
Lung cancer	0.1885	0.3544	1.2236e+03	0.9915	0.2233
Zoo	0.6440	0.8071	904.8600	0.9754	0.1600
Hepatitis	0.2285	0.1760	2.2188e+03	0.9950	0.3297
Heart disease	0.3788	0.2954	1.7151e+03	0.9915	0.3153
Voting	0.5181	0.4555	4.6905e+03	0.9990	0.1755
Credit approval	0.4432	0.3607	4.4129e+03	0.9974	0.3850
Dermatology	0.6945	0.7800	7.8548e+03	0.9852	0.1984
Breast cancer	0.7712	0.6534	4.8009e+03	0.9779	0.1938
Letters (E, F)	0.1533	0.1253	2.2097e+04	0.9936	0.3597
DNA	0.0353	0.0521	1.6498e+05	0.9981	0.5043
Mushroom	0.6059	0.5465	1.2860e+05	0.9899	0.1255

Table 7 The bounds of the function G on the 12 data sets

Data set\Bounds	Lower bound G	Upper bound G
Soybean	934.6659	1.0695e+03
Lung cancer	1.2042e+03	1.2910e+03
Zoo	861.3671	1.1332e+03
Hepatitis	2.1823e+03	2.2929e+03
Heart disease	1.6672e+03	1.8191e+03
Voting	4.5555e+03	5.3247e+03
Credit approval	4.2872e+03	4.6138e+03
Dermatology	7.6473e+03	8.6933e+03
Breast cancer	4.7048e+03	5.2007e+03
Letters (E, F)	2.1803e+04	2.2621e+04
DNA	1.6338e+05	1.6656e+05
Mushroom	1.2759e+05	1.3563e+05

6.3 The effect of the parameter T on the generalized function

In the subsection, we test the effect of the parameter T on the generalized function F_g. The test is carried out on the soybean, zoo and voting data sets. We first employ the k-modes algorithm to randomly produce the 100 different clustering results on each given data set and compute the ARI and NMI values of each clustering result. Furthermore, we select several values of the parameter T. For each value of T, we compute the function F_g value of each clustering result, and analyze the correlations between its values and the values of the ARI and NMI. Finally, we analyze how the parameter T effects on the evaluation quality of the generalized function F_g by observing the effect of the parameter T on the correlations. According to Figs. 17, 18 and 19, we see that if the parameter T value is less than a certain value, the changes of the R^2 values

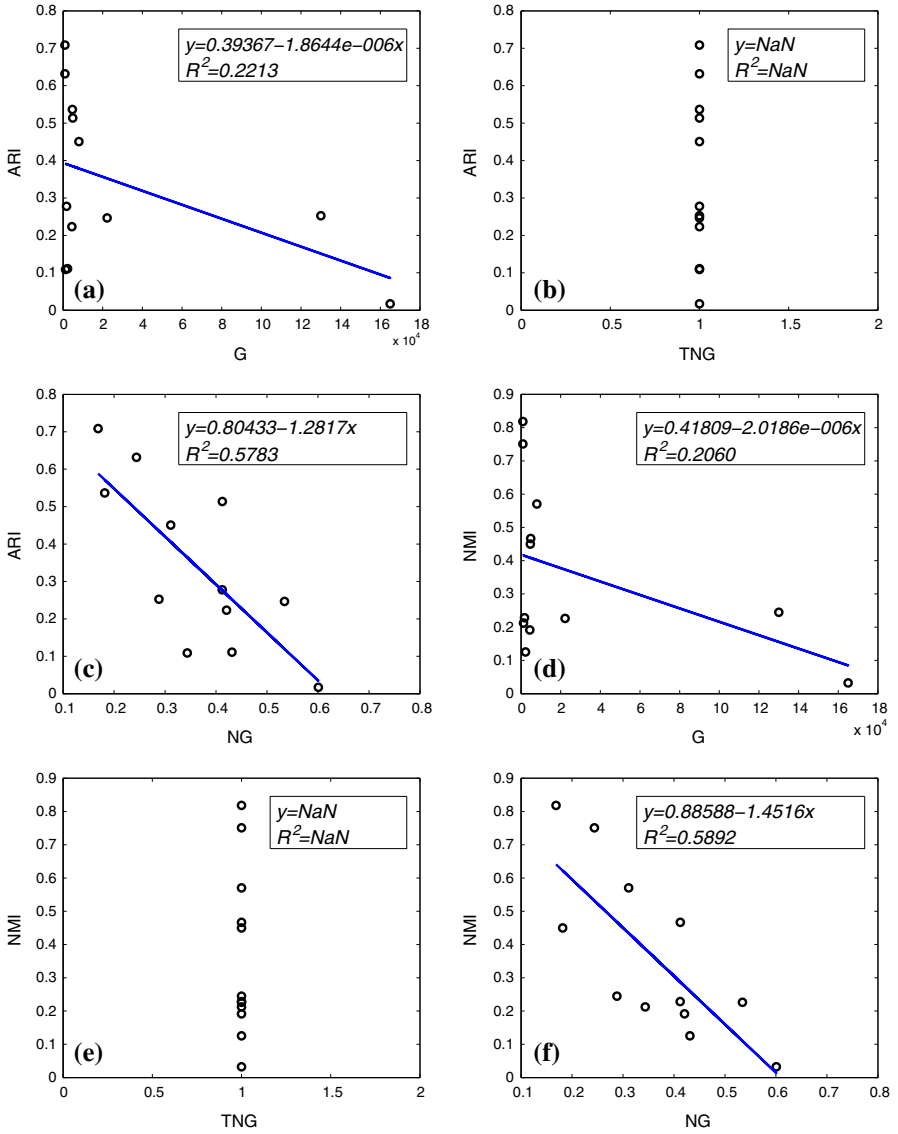

Fig. 15 Based on the AVG method, **a** the G values against the ARI values, **b** the TNG values against the ARI values, **c** the NG values against the ARI values, **d** the G values against the NMI values, **e** the TNG values against the NMI values, **f** the NG values against the NMI values

are not obvious as the T values increase. This illustrates that the generalized function F_g are robust in evaluating the clustering results when the T value is small. However, while the T value continue to grow, the R^2 values drop sharply. This indicates that the representability of each categorical value to clusters can not be recognized when the T value is very large. At this point, the generalized function F_g can not effectively evaluate clustering results. From these figures, we also see that the effects of the

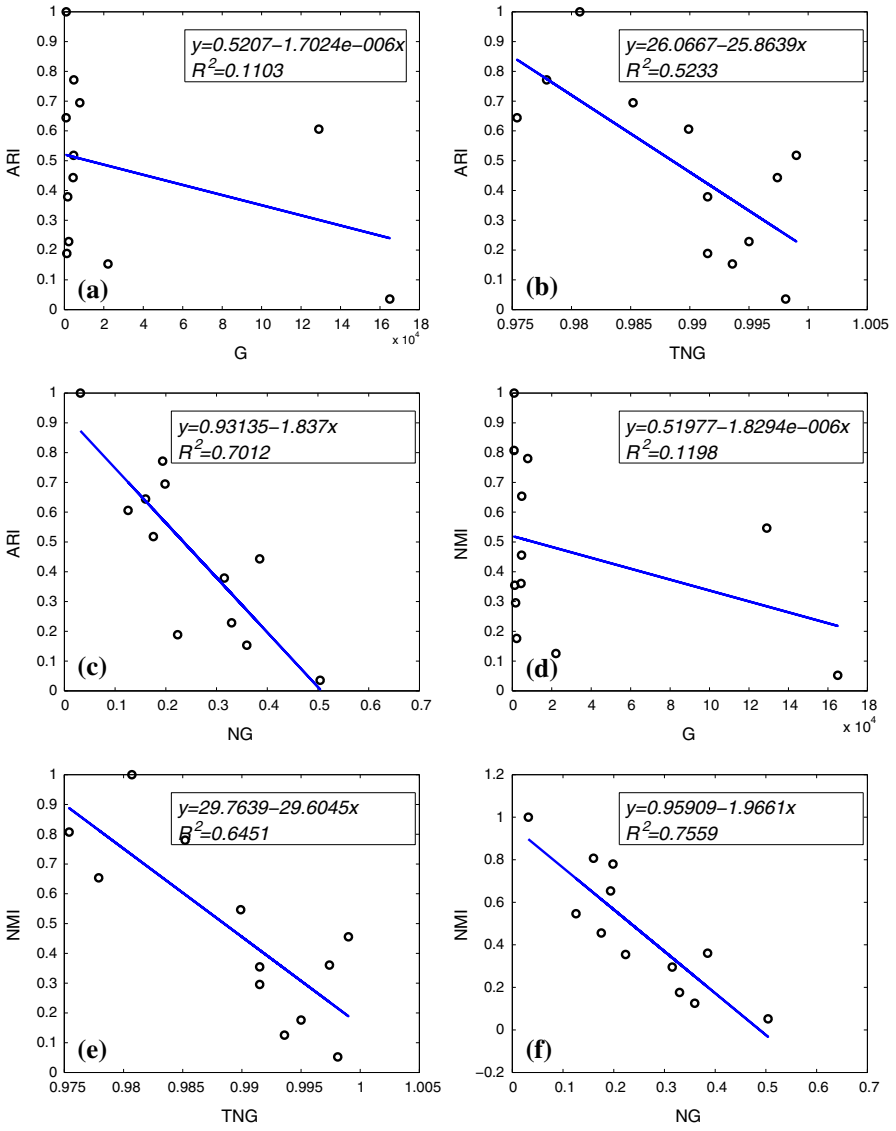

Fig. 16 Based on the MIN method, **a** the G values against the ARI values, **b** the TNG values against the ARI values, **c** the NG values against the ARI values, **d** the G values against the NMI values, **e** the TNG values against the NMI values, **f** the NG values against the NMI values

parameter T on the generalized function F_g are different on different data sets. This tells us that it is difficult to directly compute the best value of the parameter T for all the data sets, since its effect on the generalized function F_g depends on the characteristics of a data set itself. Besides, we observe that the R^2 values of the generalized function F_g with $T > 0$ are larger than those of the generalized function F_g with $T = 0$. The phenomenons are consistent with our conclusion in Sect. 3.

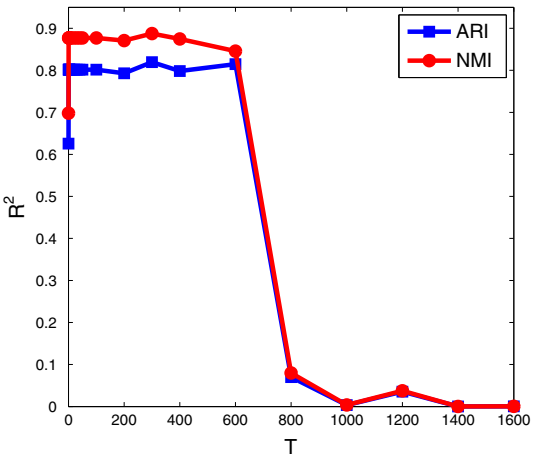

Fig. 17 The effect of the parameter T on the generalized function F_g on the soybean data set

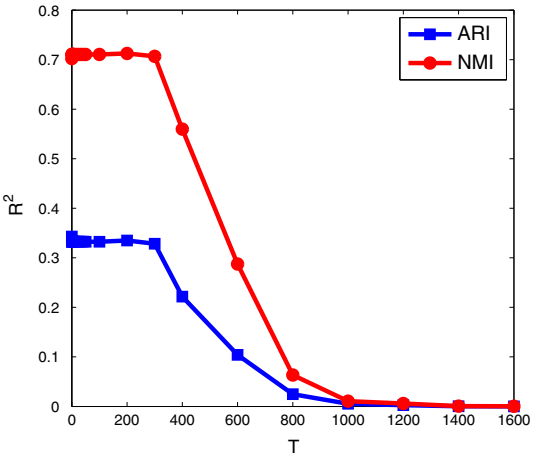

Fig. 18 The effect of the parameter T on the generalized function F_g on the zoo data set

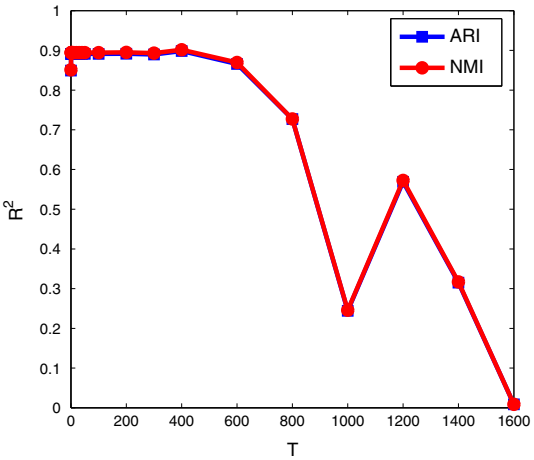

Fig. 19 The effect of the parameter T on the generalized function F_g on the voting data set

7 Conclusions

Cluster validation is a critical problem in the clustering area. In this paper, we have presented a generalized validity function for categorical data. Based on it, we have discussed the relation of the three widely-used cluster validity functions: the k-modes objective function, the category utility function and the information entropy function. The analysis results tell us that the category utility function is equivalent to the information entropy function in evaluating the clustering results, and the obtained optimal solution of the k-modes objective function on a data set is the upper bound of the optimal solution of the category utility function. Furthermore, we have discussed the relation of the within-cluster and between-cluster information in using these validity functions to evaluate the clustering results. It has been shown that these validity functions can effectively evaluate the clustering results by using the within-cluster information only. Finally, we have highlighted the importance of normalization when applying an internal validity function to compare the clustering results of different data sets. Along this line, we have provided a theoretic analysis for the bounds of the three validity functions. The experimental studies have illustrated that the normalized validity functions have better performance than the original functions when comparing clusterings of different data sets.

Acknowledgments The authors are very grateful to the editors and reviewers for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China (Nos. 71031006, 61305073, 61432011), the National Key Basic Research and Development Program of China (973) (No. 2013CB329404), the Foundation of Doctoral Program Research of Ministry of Education of China (No. 20131401120001).

References

Aggarwal CC, Magdalena C, Yu PS (2002) Finding localized associations in market basket data. IEEE Trans Knowl Data Eng 14(1):51–62

Andritsos P, Tsaparas P, Miller RJ, Sevcik KC (2004) Limbo: scalable clustering of categorical data. In: Proceedings of the ninth international conference on extending database technology

Bai L, Liang JY, Dang CY, Cao FY (2011) A novel attribute weighting algorithm for clustering high-dimensional categorical data. Pattern Recognit 44(12):2843–2861

Bai L, Liang JY, Dang CY (2013) The impact of cluster representatives on the convergence of the k-modes type clustering. IEEE Trans Pattern Anal Mach Intell 35(6):1509–1522

Barbara D, Jajodia S (2002) Applications of data mining in computer security. Kluwer, Dordrecht

Barbara D, Li Y, Couto J (2002) Coolcat: an entropy-based algorithm for categorical clustering. In: Proceedings of the eleventh international conference on information and knowledge management, pp 582–589

Baxevanis A, Ouellette F (2001) Bioinformatics: a practical guide to the analysis of genes and proteins, 2nd edn. Wiley, New York

Berry MJA, Linoff G (1996) Data mining techniques for marketing. Sales and customer support. John Wiley and Sons, New York

Chen HL, Chuang KT, Chen MS (2008) On data labeling for clustering categorical data. IEEE Trans Knowl Data Eng 20(11):1458–1472

Chen K, Liu L (2005) The "best k" for entropy-based categorical clustering. In: Proceedings of international conference on scientific and statistical database management (SSDBM), pp 253C–262C

Chen K, Liu L (2009) He-tree: a framework for detecting changes in clustering structure for categorical data streams. VLDB J 18(5):1241–1260

Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York

Dunn JC (1973) A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. J Cybern 3:32–57
Fisher DH (1987) Knowledge acquisition via incremental conceptual clustering. Mach Learn 2(2):139–172
Gluck MA, Corter JE (1985) Information uncertainty and the utility. In: Proceedings of the seventh annual conference of cognitive science society, pp 283–287
Gowda KC, Diday E (1991) Symbolic clustering using a new dissimilarity measure. Pattern Recognit 24(6):567–578
Halkidi M, Vazirgiannis M (2001) Clustering validity assessment: finding the optimal partitioning of a data set. In: Proceedings of IEEE international conference on data mining (ICDM), pp 187–194
Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. J Intell Inf Syst 17(2–3): 107–145
He Z, Deng S, Xu X (2005) Improving *k*-modes algorithm considering frequencies of attribute values in mode. In: Proceedings of computational intelligence and security, pp 157–162
Huang ZX (1997) A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Proceedings of SIGMOD workshop research issues on data mining and knowledge discovery, pp 1–8
Huang ZX, Ng MK (1999) A fuzzy *k*-modes algorithm for clustering categorical data. IEEE Trans Fuzzy Syst 7(4):446–452
Huang ZX, Ng MK, Rong H, Li Z (2005) Automated variable weighting in *k*-means type clustering. IEEE Trans Fuzzy Syst 27(5):657–668
Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice Hall, Englewood Cliffs
Li T, Ma S, Ogihara M (2004) Entropy-based criterion in categorical clustering. In: Proceedings of international conference on machine learning (ICML), pp 536–543
Liang JY, Chin KS, Dang CY, Yam RCM (2002) A new method for measuring uncertainty and fuzziness in rough set theory. Int J Gen Syst 31(4):331–342
Liu Y, Li Z, Xiong H, Gao X, Wu J (2010) Understanding of internal clustering validation measures. In: The 10th IEEE international conference on data mining (ICDM), pp 911–916
Liu Y, Li Z, Xiong H, Gao X, Wu J, Wu S (2013) Understanding and enhancement of internal clustering validation measure. IEEE Trans Syst Man Cybern B Cybern (TSMCB) 43(3):982–994
Luo P, Xiong H, Zhan GX, Wu JJ, Shi ZZ (2009) Information-theoretic distance measures for clustering validation: generalization and normalization. IEEE Trans Knowl Data Eng 21(9):1949–1962
MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. University of California Press, Berkeley, pp 281–297
Mirkin B (2001) Reinterpreting the category utility function. Mach Learn 45(3):219–228
Ng MK, Li MJ, Huang ZX, He ZY (2007) On the impact of dissimilarity measure in *k*-modes clustering algorithm. IEEE Trans Pattern Anal Mach Intell 29(3):503–507
San O, Huynh V, Nakamori Y (2004) An alternative extension of the *k*-means algorithm for clustering categorical data. Pattern Recognit 14(2):241–247
Steinbach M, Karypis G, Kumar V (2000) A comparison of document clustering techniques. In: Proceedings of workshop text mining, 6th ACMSIGKDD international conference on knowledge discovery and data mining, pp 20–23
UCI (2012) UCI machine learning repository. http://www.ics.uci.edu/mlearn/MLRepository.html
Wrigley N (1985) Categorical data analysis for geographers and environmental scientists. Longman, London
Wu J, Yuan H, Chen G (2010) Validation of overlapping clustering: a random clustering perspective. Inf Sci 180(22):4353–4369
Xiong H, Wu J, Chen J (2009) K-means clustering versus validation measures: a data distribution perspective. IEEE Trans Syst Man Cybern B Cybern 39(2):318–331
Yang YM (2004) An evaluation of statistical approaches to text categorization. J Inf Retr 1(1–2):67–88
Yu J (2005) General c-means clustering model. IEEE Trans Pattern Anal Mach Intell 27(8):1197–1211
Zhao Y, Karypis G (2004) Criterion functions for document clustering: experiments and analysis. Mach Learn 55(3):311–331