DOI: 10.11992/tis.201706029
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20171109.1250.006.html

聚类有效性评价新指标

谢娟英,周颖,王明钊,姜炜亮
(陕西师范大学 计算算计科学学院,陕西 西安 710062)

摘 要: 聚类有效性评价指标分为外部评价指标和内部评价指标两大类。现有外部评价指标没有考虑聚类结果类偏斜现象;现有内部评价指标的聚类有效性检验效果难以得到最佳类簇数。针对现有内外部聚类评价指标的缺陷,提出同时考虑正负类信息的分别基于相依表和样本对的外部评价指标,用于评价任意分布数据集的聚类结果;提出采用方差度量类内紧密度和类间分离度,以类间分离度与类内紧密度之比作为度量指标的内部评价指标。UCI 数据集和人工模拟数据集实验测试表明,提出的新内部评价指标能有效发现数据集的真实类簇数;提出的基于相依表和样本对的外部评价指标,可有效评价存在类偏斜与噪音数据的聚类结果。

关键词: 聚类;聚类有效性;评价指标;外部指标;内部指标;F-measure;Adjusted Rand Index;STDI;S2;PS2
中图分类号: TP108 **文献标志码:** A **文章编号:** 1673−4785(2017)06−0873−10

中文引用格式: 谢娟英, 周颖, 王明钊, 等. 聚类有效性评价新指标[J]. 智能系统学报, 2017, 12(6): 873–882.
英文引用格式: XIE Juanying, ZHOU Ying, WANG Mingzhao, et al. New criteria for evaluating the validity of clustering[J]. CAAI transactions on intelligent systems, 2017, 12(6): 873–882.

New criteria for evaluating the validity of clustering

XIE Juanying, ZHOU Ying, WANG Mingzhao, JIANG Weiliang
(School of Computer Science, Shaanxi Normal University, Xi'an 710062, China)

Abstract: There are two kinds of criteria for evaluating the clustering ability of a clustering algorithm, internal and external. The current external evaluation indexes fails to consider the skewed clustering result; it is difficult to get optimum cluster numbers from the clustering validity inspection results from the internal evaluation indexes. Considering the defects in the present internal and external clustering evaluation indices, we propose two external evaluation indexes, which consider both positive and negative information and which are respectively based on the contingency table and sample pairs for the evaluation of clustering results from a dataset with arbitrary distribution. The variance is proposed to measure the tightness of a cluster and the separability between clusters, and the ratio of these parameters is used as an internal evaluation index for the measurement index. Experiments on the datesets from UCI (University of California in Iven) machine learning repository and artificially simulated datasets show that the proposed new internal index can be used to effectively find the truenumber of clusters in a dataset. The proposed external indexes based on the contingency table and sample pairs are a very effective external evaluation indexes and can be used to evaluate the clustering results from existing types of skewed and noisy data.

Keywords: clustering; validity of clustering; evaluation index; external criteria; internal criteria; F-measure; Adjusted Rand Index; STDI; S2; PS2

随着人工智能技术如火如荼地发展,机器学习在各行业得到了空前的重视和应用,并取得了前所未有的成功[1-5]。聚类分析作为无监督学习方法,是各行业数据分析的主要工具之一,其旨在发现数据

收稿日期: 2017−06−08. **网络出版日期:** 2017−11−09.
基金项目: 国家自然科学基金项目 (61673251);陕西省科技攻关项目 (2013K12-03-24);陕西师范大学研究生创新基金项目 (2015CXS028, 2016CSY009);中央高校基本科研业务费重点项目 (GK201701006).
通信作者: 谢娟英. E-mail: xiejuany@snnu.edu.cn.

集样本的潜在分布模式与内在结构，发现数据集样本中所隐藏的知识。聚类分析使得同类簇的样本尽可能相似，不同类簇的样本尽可能不相似[6-7]。聚类评价指标是度量聚类结果有效性的客观指标，也是衡量聚类算法性能的客观依据，设计一个全面的聚类结果评价指标是一个困难而复杂的问题[8-13]。

根据是否利用数据集样本真实类标信息（真实的样本分布信息），聚类有效性评价指标分为外部评价指标和内部评价指标。外部评价指标通过比较聚类结果与真实分布的匹配程度，对聚类结果进行评价。现有外部评价指标分为基于相依表的，基于样本对的和基于信息熵的指标[8, 13-14]。F-measure[17-18]是最先提出的外部评价指标，是针对两类问题的评价指标，是精度和召回率的调和平均，后来被推广到多类问题。常用的外部评价指标还有 Jaccard 系数、Rand index 参数、ARI (adjusted rand index) 参数、标准化互信息 NMI (normalized mutual information) 和调整互信息 AMI (adjusted mutual information)，以及 B3(bcubed index) 等[8, 17-19]。不同外部评价指标侧重点不同，Amigó等[20]提出 4 个形式化约束 (cluster homogeneity, cluster completeness, rag bag 和 clusters size vs. quantity) 对现有外部评价指标进行比较。Vinh 等[21]指出 ARI 指标是目前最好的聚类评价指标。聚类结果类偏斜是现实世界数据，特别是生物医学数据聚类分析中的普遍现象[22-23]。尽管已经出现针对不平衡数据和不同类簇密度的聚类评价指标研究[8, 24]，但还没有考虑聚类结果偏斜的外部评价指标。鉴于此，本文利用聚类结果的相依表和样本对信息，同时考虑聚类结果的正负类信息，提出分别基于相依表和基于样本对的外部评价指标 S2(harmonic mean of sensitivity and specificity) 和 PS2(harmonic mean of sensitivity and specificity based on pairwise)，以期有效评价偏斜聚类结果。

内部评价指标没有使用原始数据分布的先验信息，常通过评价聚类结果优劣来发现数据集的内部结构和分布状态，是发现数据集最佳类簇数的常用办法[25]。内部指标有基于统计信息和基于样本几何结构的指标。IGP 指标[26](in-group proportion) 是基于统计信息的指标，通过度量在某一类簇中，距离某个样本最近的样本是否和该样本在同一类簇，来评价聚类结果的优劣。常用的基于数据集样本几何结构的内部指标有 DB 指标 (davies-bouldin)[27-28]、XB 指标 (xie-beni)[29]、Sil 指标 (silhouettes)[30]、BWP 指标 (between-within proportion)[31]等。这些聚类有效性评价内部指标自身的缺陷，使得其对于类簇结构难以判别，聚类有效性检验效果不理想，很难得到正确的聚类结果和发现最佳类簇数。针对现有内部评价指标的上述问题，本文利用方差的性质，定义类内距离和类间距离，以表达类簇间的分离性与类簇内的紧促性，提出基于类间分离性与类内紧密性之比的新内部评价指标 STDI(standard deviation based index)，以期发现数据集的真实类簇分布结构。

UCI 机器学习数据库真实数据集和人工模拟的带有刁难性的及带有噪音与类偏斜的人工模拟数据集实验测试表明，提出的内部评价新指标 STDI 能发现更合理的数据集类簇数；提出的分别基于相依表和样本对的外部评价指标 S2 和 PS2 可以有效评价有类偏斜现象的聚类结果。

1　外部指标

聚类分析中可能遇到如表 1 所示的极端情况。此时，若用 F-measure 指标评价表 1 所示极端聚类结果的有效性，将失去意义。因为，此时的 F-measure 指标值是 0.67，但实际聚类结果毫无意义。导致这种现象的原因是：F-measure 是精度和召回率的调和平均。对于两类问题，F-measure 只强调了聚类算法对正类的聚类效果，而未考虑聚类算法对负类的聚类效果。

表 1　极端聚类结果示例

Table 1　Rare case of clustering

聚类前/聚类后	真实分布相依表		聚类算法得到的相依表	
	聚类后正类	聚类后负类	聚类后正类	聚类后负类
聚类前正类	50	0	50	0
聚类前负类	0	50	50	0

为了避免此类问题，本文提出一种基于相依表的、同时考虑正负类聚类结果的评价指标 S2。S2 指标调和了聚类算法对于正负类的聚类效果，是灵敏度和特异度的调和平均。如同 F-measure 可推广于多类问题一样，S2 同样适用于作为多类问题的聚类评价指标。

设聚类结果类簇数为 K，原始类簇数为 C，则聚类结果相依表是表 2 所示的 $C\times K$ 矩阵，$\boldsymbol{U}$ 是真实分布，$\boldsymbol{V}$ 是聚类算法所得聚类结果，则任意类簇 c 的 TP_c、FN_c、FP_c、TN_c 分别定义如式 (1) 所示。其中，$\boldsymbol{I}$ 为原始类标信息，$\boldsymbol{L}$ 为聚类所得类标信息，n 为样本数。以类簇 c 为正类的 sensitivity 和 specificity 定义如式 (2) 所示。则新聚类指标 S2 如式 (3) 定义。当类簇数 K=2 时，式 (3) 的 S2 指标退化为式

(4), 其中的 sensitivity 和 specificity 同 F-measure 指标在两类问题中的定义一致。由此可见, 我们定义的新指标 S2 适用于任意类的聚类问题。

表 2　聚类结果相依表

Table 2　The contingency table of a clustering

U/V	V_1	V_2	…	V_c	V_K	SUM
U_1	n_{11}	n_{12}	…	n_{1c}	n_{1K}	$n_{1\cdot}$
U_2	n_{21}	n_{22}	…	n_{2c}	n_{2K}	$n_{2\cdot}$
⋮	⋮	⋮		⋮	⋮	⋮
U_c	n_{c1}	n_{c2}	…	n_{cc}	n_{cK}	$n_{c\cdot}$
⋮	⋮	⋮		⋮	⋮	⋮
U_C	n_{C1}	n_{C2}	…	n_{Cc}	n_{CK}	$n_{C\cdot}$
SUM	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot c}$	$n_{\cdot K}$	n

$$\begin{cases} \mathrm{TP}_c = |\{i \mid \boldsymbol{l}(\boldsymbol{x}_i) = \boldsymbol{L}(\boldsymbol{x}_i) = c, 1 \leqslant i \leqslant n\}| = n_{cc} \\ \mathrm{FN}_c = |\{i \mid \{\boldsymbol{l}(\boldsymbol{x}_i) = c\} \wedge \{\boldsymbol{L}(\boldsymbol{x}_i) \neq c\}, 1 \leqslant i \leqslant n\}| = n_{c\cdot} - n_{cc} \\ \mathrm{FP}_c = |\{i \mid \{\boldsymbol{l}(\boldsymbol{x}_i) \neq c\} \wedge \{\boldsymbol{L}(\boldsymbol{x}_i) = c\}, 1 \leqslant i \leqslant n\}| = n_{\cdot c} - n_{cc} \\ \mathrm{TN}_c = |\{i \mid \boldsymbol{l}(\boldsymbol{x}_i) \neq c, \boldsymbol{L}(\boldsymbol{x}_i) \neq c, 1 \leqslant i \leqslant n\}| = n - n_{c\cdot} - n_{\cdot c} + n_{cc} \end{cases} \tag{1}$$

$$\begin{aligned} \mathrm{sensitivity}_c &= \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c} = \frac{n_{cc}}{n_{c\cdot}} \\ \mathrm{specificity}_c &= \frac{\mathrm{TN}_c}{\mathrm{TN}_c + \mathrm{FP}_c} = \frac{n - n_{c\cdot} - n_{\cdot c} + n_{cc}}{n - n_{c\cdot}} \end{aligned} \tag{2}$$

$$S2 = \frac{1}{\min\{C, K\}} \sum_{c=1}^{\min\{C,K\}} \frac{2 \times \mathrm{sensitivity}_c \times \mathrm{specificity}_c}{\mathrm{sensitivity}_c + \mathrm{specificity}_c} \tag{3}$$

$$S2 = \frac{2 \times \mathrm{sensitivity} \times \mathrm{specificity}}{\mathrm{sensitivity} + \mathrm{specificity}} \tag{4}$$

外部评价指标中的 Rand index、Adjusted rand index、Jaccard 系数, AMI 等均是基于样本对的聚类评价指标。因此, 本文类似地提出基于样本对的聚类结果外部评价指标 PS2, 调和聚类结果的正类识别率和负类识别率, 以评价聚类结果的有效性。

任意两样本点 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$, 若 $\boldsymbol{l}(\boldsymbol{x}_i) = \boldsymbol{l}(\boldsymbol{x}_j)$, 且 $\boldsymbol{L}(\boldsymbol{x}_i) = \boldsymbol{L}(\boldsymbol{x}_j)$, 即聚类前后属于同一类, 则称为正事件 T; 反之, 如果 $\boldsymbol{l}(\boldsymbol{x}_i) = \boldsymbol{l}(\boldsymbol{x}_j)$, 但 $\boldsymbol{L}(\boldsymbol{x}_i) \neq \boldsymbol{L}(\boldsymbol{x}_j)$, 即聚类前属于同类簇, 但聚类后不属于同一类, 称之为负事件 F。依据正负事件, 可得表 3 所示混淆矩阵。其中, TP、FN、FP 和 TN 分别表示聚类前后都在同一类簇的样本对数; 聚类前在同一类簇, 聚类后不在同一类簇的样本对数; 聚类前不在同一类簇, 聚类后在于同一类簇的样本对数; 和聚类前后都不在同一类簇的样本对数。其形式化定义如式 (5) 所示。由定义可知, TP 和 TN 统计了聚类所得划分与原始分布的一致性, FN 和 FP 统计了聚类所得划分与原始分布的差异性。设 N 表示规模为 n 的数据集的所有样本对数, 则 $N = \binom{n}{2} = \frac{n(n-1)}{2}$, 即, N=TP+FN+FP+TN。

TP、FN、FP 和 TN 也可根据表 2 所示的相依表计算得到。计算公式如式 (6) 所示。基于样本对的 sensitivity, specificity 定义如式 (7) 所示, 则基于样本对的新聚类评价指标 PS2 定义为式 (8)。

表 3　聚类结果混淆矩阵

Table 3　Confusion matrix of a clustering

聚类前/聚类后	T'	F'
T	TP	FN
F	FP	TN

$$\begin{cases} \mathrm{TP} = \left|\{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{l}(\boldsymbol{x}_i) = \boldsymbol{l}(\boldsymbol{x}_j),\ \boldsymbol{L}(\boldsymbol{x}_i) = \boldsymbol{L}(\boldsymbol{x}_j)\}\right| \\ \mathrm{FN} = \left|\{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{l}(\boldsymbol{x}_i) = \boldsymbol{l}(\boldsymbol{x}_j),\ \boldsymbol{L}(\boldsymbol{x}_i) \neq \boldsymbol{L}(\boldsymbol{x}_j)\}\right| \\ \mathrm{FP} = \left|\{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{l}(\boldsymbol{x}_i) \neq \boldsymbol{l}(\boldsymbol{x}_j),\ \boldsymbol{L}(\boldsymbol{x}_i) = \boldsymbol{L}(\boldsymbol{x}_j)\}\right| \\ \mathrm{TN} = \left|\{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{l}(\boldsymbol{x}_i) \neq \boldsymbol{l}(\boldsymbol{x}_j),\ \boldsymbol{L}(\boldsymbol{x}_i) \neq \boldsymbol{L}(\boldsymbol{x}_j)\}\right| \end{cases} \tag{5}$$

$$\begin{cases} \mathrm{TP} = \sum_{i=1}^{C} \sum_{j=1}^{K} \binom{n_{ij}}{2} \\ \mathrm{FN} = \sum_{i=1}^{C} \binom{n_{i\cdot}}{2} - \mathrm{TP} \\ \mathrm{FP} = \sum_{j=1}^{K} \binom{n_{\cdot j}}{2} - \mathrm{TP} \\ \mathrm{TN} = N - (\mathrm{TP} + \mathrm{FN} + \mathrm{FP}) \end{cases} \tag{6}$$

$$\begin{cases} \mathrm{sensitivity} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \\ \mathrm{specificity} = \dfrac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \end{cases} \tag{7}$$

$$\begin{cases} \mathrm{PS2} = \dfrac{2 \times \mathrm{sensitivity} \times \mathrm{specificity}}{\mathrm{sensitivity} + \mathrm{specificity}} = \\ \dfrac{2 \times \mathrm{TP} \times \mathrm{TN}}{\mathrm{TP}(\mathrm{FP} + \mathrm{TN}) + \mathrm{TN}(\mathrm{TP} + \mathrm{FN})} \end{cases} \tag{8}$$

2　内部指标

方差作为一种度量样本分布情况的概率统计量, 通常用来描述样本的离散程度[32]。样本方差越小, 样本分布越密集, 反之则越分散。方差的性质可以用于计算类内距离和类间距离, 同一类簇中样本分布越密集, 方差越小, 因此将同一类簇中样本的方差作为类内距离, 度量类簇内部的紧促性。

基于“类内尽可能紧密, 类间尽可能分离”原则, 利用方差思想定义度量类内距离和类间距离测度, 类间距离越大越好, 类内距离越小越好, 提出将类间距离与类内距离之比作为聚类效果的内部评价指标 STDI(standard deviation based index), 如式 (9) 所示。从式 (9)STDI 的定义可知, 其值越大, 表明聚类结果越好。

$$\mathrm{STDI} = \frac{\frac{1}{K}\left(\sum_{k=1}^{K} \|\boldsymbol{c}_k - \bar{\boldsymbol{x}}\|^2\right)}{\sum_{k=1}^{K} \frac{1}{n_k}\left(\sum_{i=1}^{n_k} \|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2\right)} \tag{9}$$

式中: $\boldsymbol{c}_k$ 是类簇 k 的质心, $\bar{\boldsymbol{x}}$是所有样本的质心, $\boldsymbol{x}_i$ 是

类簇 k 的第 i 个样本，n_k 是类簇 k 的样本数，K 是数据集的类簇数。STDI 指标的分子表示各类簇间方差，分母表示各类簇方差之和。显然簇内方差越小，则分母越小，表示类簇内部分布越紧密，簇间方差越大，则分子越大，表示各类簇的分离性越好。因此，STDI 的值越大越好。

3　实验分析

本节将分别测试提出的内部指标和外部指标的性能。因为篇幅所限，内部指标只使用图 1 所示的具有挑战性的人工模拟数据集进行测试，该数据集经常被识别为 3 个类簇。外部评价指标将使用来自 UCI 机器学习数据库[33]的真实数据集和人工模拟数据集两大类数据进行测试。其中的人工模拟数据包括：类簇样本分布不平衡的偏斜数据，以及类簇样本分布平衡但各类簇间存在部分交叠的数据。这样设计人工模拟数据集的目的在于：检测提出的外部指标 S2 与 PS2 对带有噪音以及类别分布不平衡数据聚类结果的判断能力。测试外部指标的人工模拟数据集如图 2 所示，表 4 是图 2 各数据集的详细信息，测试外部指标的 UCI 机器学习数据库的真实数据集如表 5 所示。

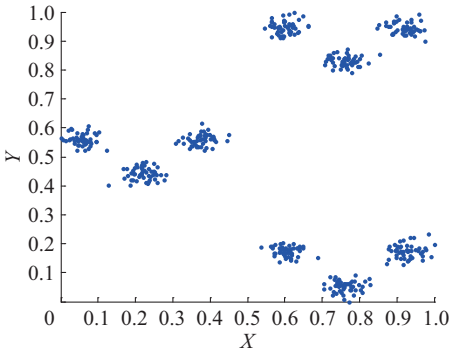
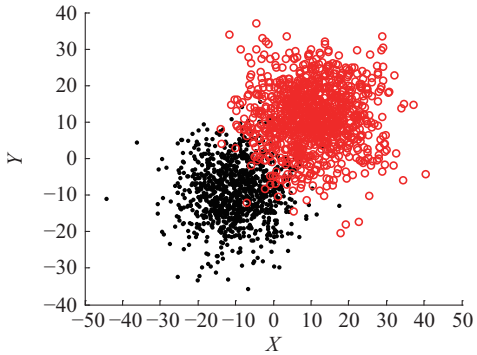
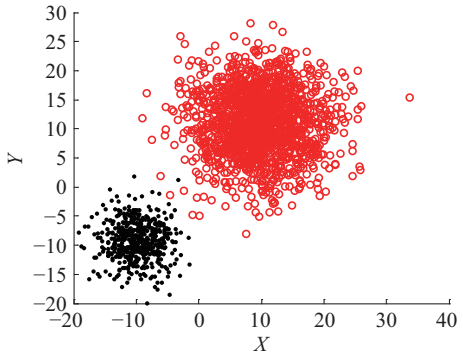


图 1　测试内部指标 STDI 的人工数据集原始分布

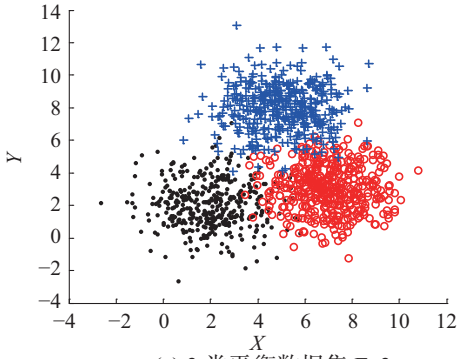
Fig. 1　The synthetic data set to test the new internal criteria STDI



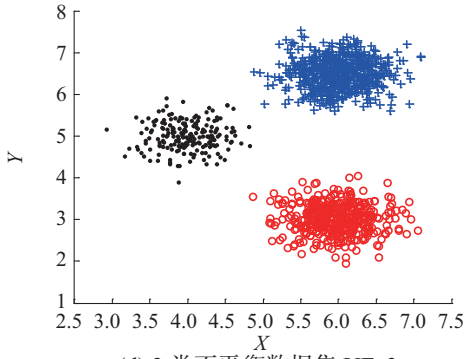
(a) 2 类平衡数据集 Ec2



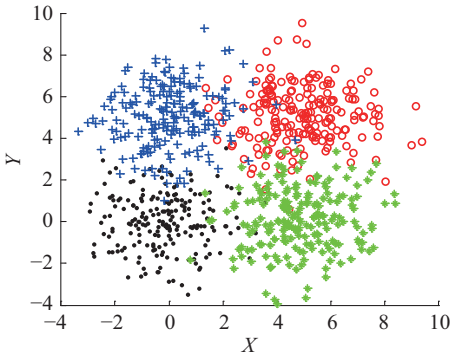
(b) 2 类不平衡数据集 UEc2



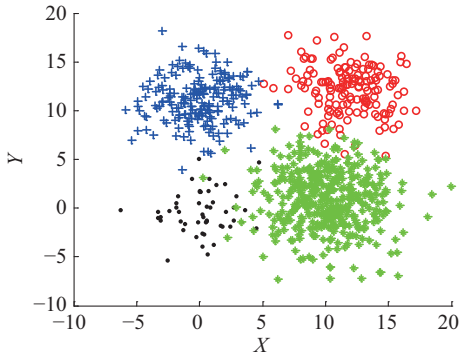
(c) 3 类平衡数据集 Ec3



(d) 3 类不平衡数据集 UEc3



(e) 4 类平衡数据集 Ec4



(f) 4 类不平衡数据集 UEc4

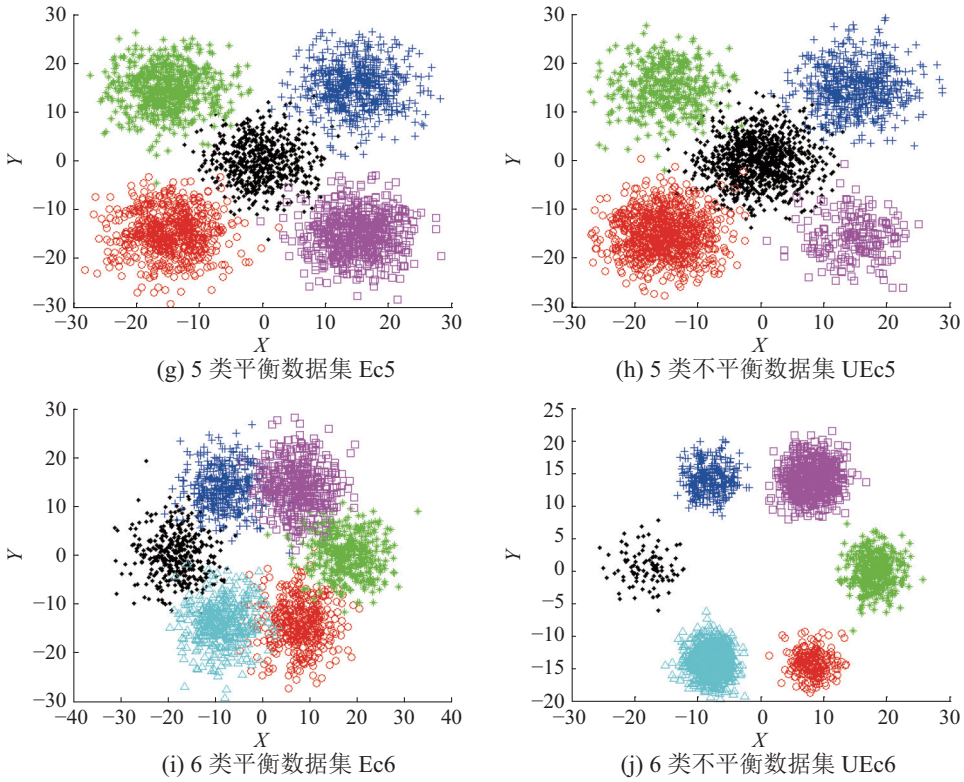


图 2　测试外部指标 S2 和 PS2 的人工数据集原始分布

Fig. 2　The synthetic data sets to test the new external criteria S2 and PS2

表 4　测试新外部指标 S2 和 PS2 的人工模拟数据集信息

Table 4　The detail information of synthetic data sets to test the proposed external criteria S2 and PS2

数据集	样本数	类簇数	各类簇样本数
Ec2	2 000	2	1 000 1 000
Ec3	1 200	3	400 400 400
Ec4	800	4	200 200 200 200
Ec5	3 000	5	600 600 600 600 600
Ec6	2 400	6	400 400 400 400 400 400
UEc2	2 000	2	500 1 500
UEc3	1 200	3	200 400 600
UEc4	800	4	50 150 200 400
UEc5	3 000	5	1 000 800 600 1 400 200
UEc6	2 400	6	100 200 300 400 600 800

表 5　测试新外部指标 S2 和 PS2 的 UCI 数据集

Table 5　The data sets from UCI machine learning repository to test the proposed external criteria S2 and PS2

数据集	样本数	类簇数	各类簇样本数
Iris	150	3	50 50 50
Seeds	210	3	70 70 70
Segmentation	210	7	30 30 30 30 30 30 30
Soybean	47	4	10 10 10 17
wine	178	3	59 71 48
wdbc	569	2	357 212
Bupa	345	2	145 200
pima-indians-diabetes	768	2	500 268
Balance_scale	625	3	49 288 288
New_thyroid	215	3	150 35 30
Ionosphere	351	2	38 313
Haberman	306	2	225 81

3.1　内部指标有效性测试实验

内部指标不需要任何先验知识, 通过评价聚类结果, 发现数据集样本的潜在分布与内在结构, 常用于发现数据集的类簇数。因此, 我们以能否准确发现数据集的真实类簇数来测试提出的内部指标 STDI 指标的有效性, 并与现有内部指标 DB、XB、IGP、Sil 和 BWP 的性能进行比较。图 3 给出了各内部指标对图 1 所示人工模拟数据集的实验结果。这里的聚类算法使用的是 SD 算法[35]。

从图 3 各指标的实验结果可以看出, 只有图 3(a) 展示的 STDI 指标的实验结果可以发现图 1 所示人工数据集的真实类簇数 9, 其余 5 个指标均在类簇数为 3 时最佳, 即其余指标发现的该数据集类簇数是 3。因此, 只有用本文提出内部聚类指标 STDI 可以得到该人工模拟数据集的正确类簇数。

分析原因是：本文提出的 STDI 指标采用各类簇质心方差度量类间分离程度，用各类簇样本方差度量类内紧密程度，当类簇数为 9 时，各类簇质心方差较大，而簇内样本方差较小，因此得到最佳聚类结果，发现数据集的正确类簇数。由此可见，本文提出的 STDI 指标是非常有效的一种聚类评价指标。

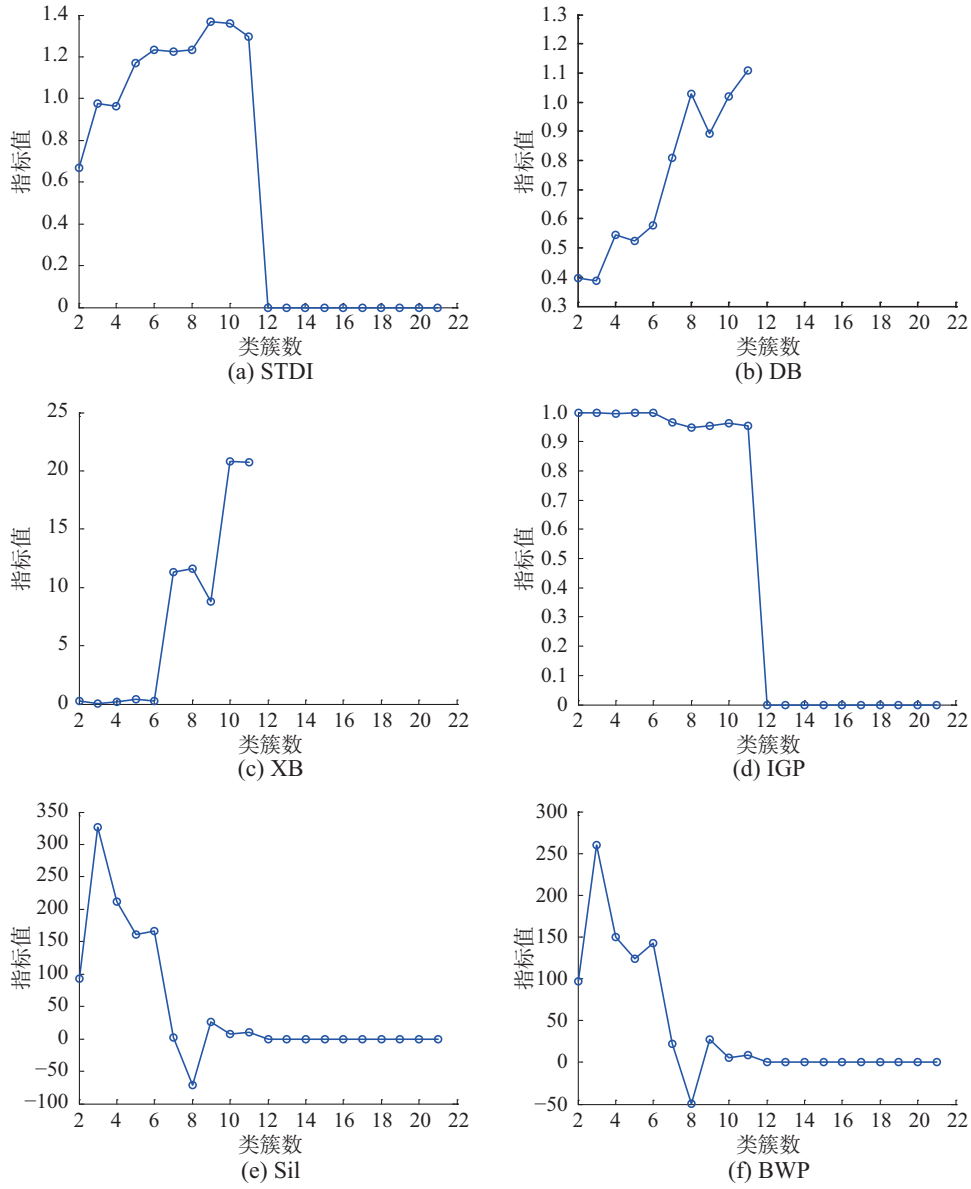


图 3　各内部指标在人工数据集的测试结果

Fig. 3　The results on synthetic data set of internal criteria

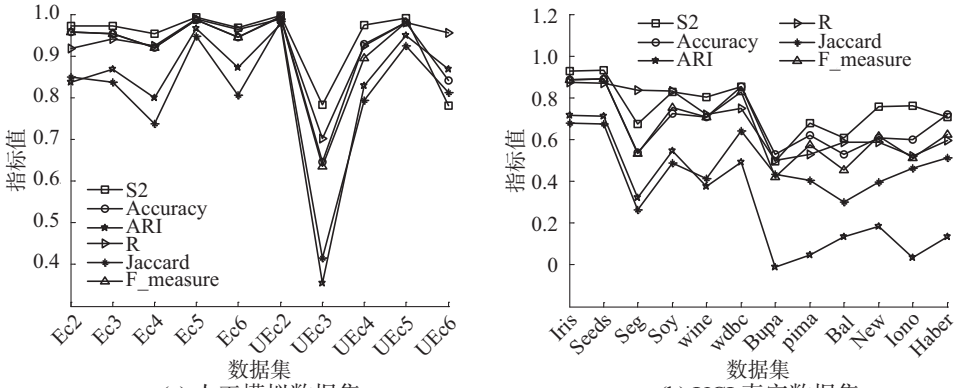
3.2　外部指标有效性测试实验

本小节对提出的 2 种聚类有效性评价外部指标 S2 和 PS2 进行测试，聚类算法选取快速 K-medoids 算法[35]。为了充分说明提出的外部评价指标 S2 和 PS2 的有效性，特别设计了带有噪音，类簇分布平衡和不平衡的人工模拟数据集，并选择了来自 UCI 机器学习数据库的样本数、类簇数和各类簇样本规模各异的真实数据集来进行测试，同时将提出的 S2 和 PS2 指标与聚类准确率 Accuracy，以及经典外部评价指标 F-measure、Rand index、Jaccard 系数和 ARI 的指标值进行比较。

图 2 和表 4 所示人工模拟数据集的类簇数从 2~6，类簇数相同的人工模拟数据集包括两类：类簇样本数均衡，但簇间样本重叠的情况；类簇样本数不平衡，即存在类簇偏斜，簇间样本重叠或很少量重叠的情况。这样的人工模拟数据集将测试提出的外部评价指标 S2 和 PS2 对存在类偏斜或样本重叠分布的数据聚类结果的评价情况。表 5 来自 UCI 机器学习数据库的 12 个真实数据集的样本数，类簇数和类簇样本分布也各不相同。这些真实数据集将进一步检测提出的外部评价指标 S2 和 PS2 的有效性。

为了清楚展示 S2 和 PS2 指标的性能，分别将 S2 和 PS2 的实验测试结果与聚类准确率 Accuracy，

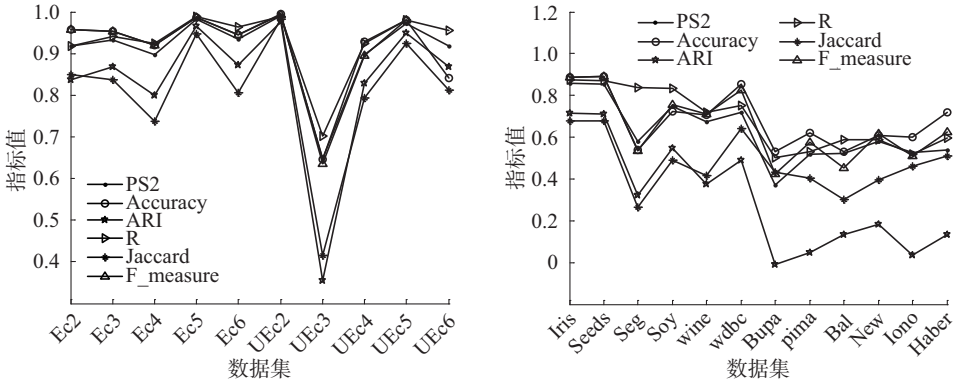
经典外部评价指标 F-measure、Rand index、Jaccard 系数和 ARI 指数进行比较, 并将 S2 和 PS2 指标与聚类准确率独立比较。图 4 展示了 S2 指标在人工模拟数据集和真实数据集的测试结果与其他指标的比较。图 5 给出了 PS2 指标的实验测试结果与其他指标的比较。S2 与 PS2 的性能比较如图 6 所示, 图 6 同时展示了聚类准确率指标。图 4 和图 5 中的 R 是 Rand index 的简写。



(a) 人工模拟数据集　　(b) UCI 真实数据集

图 4　S2 指标与其他指标的测试结果比较

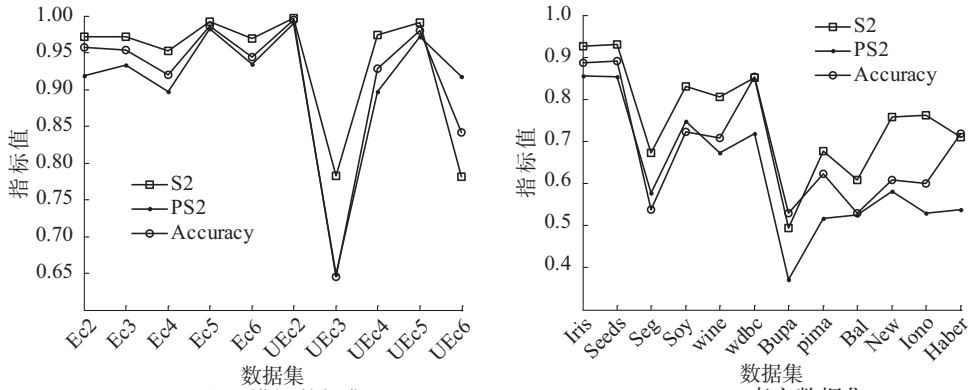
Fig. 4　The comparison of S2 with other criteria



(a) 人工模拟数据集　　(b) UCI 真实数据集

图 5　PS2 指标的测试结果与其他指标的比较

Fig. 5　The comparison of PS2 with other criteria



(a) 人工模拟数据集　　(b) UCI 真实数据集

图 6　S2 与 PS2 指标与聚类准确率比较

Fig. 6　The comparison of S2 and PS2 and clustering accuracy

图 4(a) 人工模拟数据集的实验结果揭示, 除了含有 6 个不平衡类簇的人工模拟数据集外, 本文提出的同时考虑正负类信息的聚类有效性评价指标 S2 与其他指标相比具有最高值, 且与其他指标在各数据集测试的指标值走势一致。因此, 可以说提出的 S2 指标可以有效评价存在类偏斜分布的聚类结果。图 4(b) 所示的 UCI 机器学习数据库真实数据集的实验测试结果显示, 提出的外部评价指标

S2 在 12 个真实数据集的指标值只有在 Segmentation 和 Bupa 两个数据集的测试指标值不是最高，在其余 10 个真实数据集的测试结果值均高于聚类准确率 Accuracy，以及经典外部指标 Rand index 指数，ARI，Jaccard 系数和 F-measure。另外，提出的 S2 指标在各真实数据集的测试值与 Accuracy，Jaccard，ARI 和 F-measure 各指标值的走势基本一致，但与 Rand index 指标不太一致。图 4(a) 和 (b) 的实验结果共同揭示，提出的 S2 指标的测试值与聚类准确率 Accuracy，外部指标 F-measure，Rand index 指数，ARI 和 Jaccard 系数在各数据集的基本走势大体一致。当前最优的外部评价指标 ARI 在各指标值中位居后两位，特别是在真实数据集，ARI 特别突出的位于后两位。这更进一步说明了提出的同时考虑正负类信息的外部评价指标 S2 的有效性。

图 5(a) 人工模拟数据集的实验结果显示，除了含有 6 个不平衡类簇的人工模拟数据集，提出的基于样本对信息，同时考虑正负类信息的外部评价指标 PS2 在其他人工模拟数据集的指标值基本与聚类准确率重合，或略低于聚类准确率，但走势一致。图 5(b) 真实数据集实验结果显示，提出的 PS2 指标低于或等于聚类准确率，聚类准确率或 Rand index 指数在真实数据集的测试结果高于等于提出的 PS2 指标。当前最佳聚类评价指标 ARI 在带有噪音和类簇分布不平衡的人工模拟数据集，以及样本规模，类簇数和各类簇样本规模变化各异的真实数据集的测试结果与其他指标相比，取值较低，在 6 个比较指标中居后两位。

图 6(a) 人工模拟数据集实验结果显示，除了在含有 6 个不平衡类簇的人工模拟数据集的 S2 指标低于 PS2 指标和聚类准确率外，在其余人工模拟数据集上，S2 指标的指标值均高于 PS2 指标，聚类准确率居中。图 6(b) 真实数据集实验结果显示，在真实数据集的 S2 指标明显高于 PS2 指标值。真实数据集的聚类准确率 Accuracy 除了在 Bupa 数据集高于 S2 和 PS2 指标，在 Segmentation 数据集低于 S2 和 PS2 指标外，在其余数据集的聚类准确率均低于等于 S2 指标，但高于 PS2 指标。聚类分析的目的是发现数据集的正确类簇分布。图 6(a) ~ (b) 的实验结果揭示，提出的分别基于相依表和样本对，且同时考虑正负类信息的外部评价指标 S2 和 PS2 均能正确评价聚类结果的有效性，其走势与聚类准确率大体一致。其中，S2 指标的走势更趋近于聚类准确率。

4　结束语

聚类作为无监督学习，是大数据集背景下知识发现的重要方法之一。聚类学习结果的有效性评价是聚类分析不可或缺的重要组成部分。现有聚类评价指标的外部评价指标侧重于正类，对聚类结果类偏斜问题缺少考虑，为此，提出了分别基于相依表和样本对的，同时考虑正负类信息的外部评价新指标 S2 和 PS2。另外，针对现有内部评价指标在发现数据集最佳类簇数方面的局限，提出了基于方差的类内紧密度和类间分离性度量，定义了以类间分离性与类内紧密度之比为度量指标的内部评价新指标 STDI。UCI 机器学习数据库真实数据集和带有刁难性的人工模拟数据集实验测试表明，提出的新内部指标 STDI 能有效发现数据集的真实类簇数；提出的外部指标 S2 和 PS2 是非常有效的聚类有效性外部评价指标，可有效评价存在类偏斜与噪音数据的聚类结果。

参考文献：

[1] ESTEVA A, KUPREL B, NOVOA RA, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115–118.

[2] FARINA D, VUJAKLIJA I, SARTORI M, et al. Man/machine interface based on the discharge timings of spinal motor neurons after targeted muscle reinnervation[J]. Nature biomedical engineering, 2017, 1: 25.

[3] GULSHAN V, PENG L, CORAM M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs[J]. JAMA, 2016, 316(22): 2402–2410.

[4] LONG E, LIN H, LIU Z, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts[J]. Nature biomedical engineering, 2017, 1: 0024.

[5] ORRINGER DA, PANDIAN B, NIKNAFS Y S, et al. Rapid intraoperative histology of unprocessed surgical specimens via fibre-laser-based stimulated Raman scattering microscopy[J]. Nature biomedical engineering, 2017, 1: 0027.

[6] HAN J, PEI J, KAMBER M. Data mining: concepts and techniques[M]. Singapore: Elsevier, 2011.

[7] JAIN AK, DUBES RC. Algorithms for clustering data [M]. Prentice-Hall, 1988.

[8] DE SOUTO MCP, COELHO ALV, FACELI K, et al. A comparison of external clustering evaluation indices in the context of imbalanced data sets[C]//2012 Brazilian Symposium on Neural Networks (SBRN). [S.l.], 2012: 49-54.

[9] HUANG S, CHRNG Y, LANG D, et al. A formal algorithm for verifying the validity of clustering results based on model checking[J]. PloS one, 2014, 9(3): e90109.

[10] RENDÓN E, ABUNDEZ I, ARIZMENDI A, et al. Intern-

al versus external cluster validation indexes[J]. International journal of computers and communications, 2011, 5(1): 27–34.

[11] ROSALES-MENDÉZ H, RAMÍREZ-CRUZ Y. CICE-BCubed: A new evaluation measure for overlapping clustering algorithms[C]//Iberoamerican Congress on Pattern Recognition. Berlin: Springer Berlin Heidelberg, 2013: 157-164.

[12] SAID AB, HADJIDJ R, FOUFOU S. Cluster validity index based on jeffrey divergence[J]. Pattern analysis and applications, 2017, 20(1): 21–31.

[13] XIONG H, WU J, CHEN J. K-means clustering versus validation measures: a data-distribution perspective[J]. IEEE transactions on systems, man, and cybernetics, part b (cybernetics), 2009, 39(2): 318–331.

[14] POWERS D M W. Evaluation: from precision, recall and F-factor to ROC, informedness, markedness and correlation[J]. Journal of machine learning technologies, 2011, 2: 2229–3981.

[15] LARSEN B, AONE C. Fast and effective text mining using linear-time document clustering[C]//Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA: ACM, 1999: 16-22.

[16] ZU EISSEN, B S S M, WIßBROCK F. On cluster validity and the information need of users[C]//Conference on Artificial Intelligence and Applications, Benalmádena, Spain, 2003. Calgary, Canada: ACTA Press, 2003: 216-221.

[17] 谢娟英. 无监督学习方法及其应用[M]. 北京: 电子工业出版社, 2016.
XIE Juanying, Unsupervised learning methods and applications[M]. Beijing: Publishing House of Electronics Industry, 2016.

[18] XIE J Y, GAO H C, XIE W X, et al. Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors[J]. Information sciences, 2016, 354: 19–40.

[19] 谢娟英, 高红超, 谢维信. K 近邻优化的密度峰值快速搜索聚类算法[J]. 中国科学: 信息科学, 2016, 46(2): 258–280.
XIE Juanying, GAO Hongchao, XIE Weixin. K-nearest neighbors optimized clustering algorithm by fast search and finding the density peaks of a dataset[J]. Scientia sinica informationis, 2016, 46(2): 258–280.

[20] AMIGÓ E, GONZALO J, ARTILES J, et al. A comparison of extrinsic clustering evaluation metrics based on formal constraints[J]. Information retrieval, 2009, 12(4): 461–486.

[21] VINH NX, EPPS J, BAILEY J. Information theoretic measures for clusterings comparison: is a correction for chance necessary [C]//Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Canada, 2009. New York, USA: ACM, 2009: 1073-1080.

[22] D'HAESELEER P. How does gene expression clustering work[J]. Nature biotechnology, 2005, 23(12): 1499.

[23] QUACKENBUSH J. Computational analysis of microarray data[J]. Nature reviews genetics, 2001, 2(6): 418–427.

[24] CHOU CH, SU MC, LAI E. A new cluster validity measure for clusters with different densities[C]//IASTED International Conference on Intelligent Systems and Control. Calgary, Canada: ACTA Press, 2003: 276-281.

[25] 谢娟英, 周颖. 一种新聚类评价指标[J]. 陕西师范大学学报: 自然科学版, 2015, 43(6): 1–8.
XIE Juanying, ZHOU Ying. A new criterion for clustering algorithm[J]. Journal of Shaanxi normal university: natural science edition, 2015, 43(6): 1–8.

[26] KAPP AV, TIBSHIRANI R. Are clusters found in one dataset present in another dataset[J]. Biostatistics, 2007, 8(1): 9–31.

[27] DAVIES DL, BOULDIN DW. A cluster separation measure[J]. IEEE transactions on pattern analysis and machine intelligence, 1979(2): 224–227.

[28] HASHIMOTO W, NAKAMURA T, MIYAMOTO S. Comparison and evaluation of different cluster validity measures including their kernelization[J]. Journal of advanced computational intelligence and intelligent informatics, 2009, 13(3): 204–209.

[29] XIE XL, BENI G. A validity measure for fuzzy clustering[J]. IEEE transactions on pattern analysis and machine intelligence, 1991, 13(8): 841–847.

[30] ROUSSEEUW PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of computational and applied mathematics, 1987, 20: 53–65.

[31] 周世兵, 徐振源, 唐旭清. 一种基于近邻传播算法的最佳聚类数确定方法[J]. 控制与决策, 2011, 26(8): 1147–1152.
ZHOU Shibing, XU Zhenyuan, TANG Xuqing. Method for determining optimal number of clusters based on affinity propagation clustering[J]. Control and decision, 2011, 26(8): 1147–1152.

[32] 盛骤, 谢式千. 概率论与数理统计及其应用[M]. 北京: 高等教育出版社, 2004.
SHENG Zhou, XIE Shiqian. Probability and mathematical statistics and its application[M]. Beijing: Higher education press, 2004.

[33] LICHMAN M, UCI Machine learning repository[EB/OL]. 2013, University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml.

[34] 谢娟英, 高瑞. 方差优化初始中心的 K-medoids 聚类算法[J]. 计算机科学与探索, 2015, 9(8): 973–984.
XIE Juanying, GAO Rui. K-medoids clustering algorithms with optimized initial seeds by variance[J]. Journal of fron-

tiers of computer science and technology, 2015, 9(8): 973–984.

[35] PARK HS, JUN CH. A simple and fast algorithm for K-medoids clustering[J]. Expert systems with applications, 2009, 36(2): 3336–3341.

作者简介：

谢娟英，女，1971 年生，副教授，博士，主要研究方向为机器学习、数据挖掘和生物医学大数据分析。国际期刊 HISS 副编委。发表学术论文 60 余篇，单篇 google scholar 他引次数百余次，SCI 源刊数据库单篇他引次数 40 余次。出版专著 2 部。

周颖，女，1992 年生，硕士研究生，主要研究方向为数据挖掘。

王明钊，男，1990 年生，硕士研究生，主要研究方向为数据挖掘。

第二届智能计算与信号处理国际学术会议 (ICSP 2018)
2018 2nd International Conference on Intelligent Computing and Signal Processing

第二届智能计算与信号处理国际学术会议 (ICSP 2018) 定于 2018 年 3 月 23 日至 25 日在中国武汉隆重举行。会议主要围绕智能计算与信号处理等研究领域展开讨论。旨在为智能计算与信号处理的专家学者及企业发展人提供一个分享研究成果、讨论存在的问题与挑战、探索前沿科技的国际性合作交流平台。欢迎海内外学者投稿和参会。

论文评审及出版

1、论文必须是英文稿件，且论文应具有学术或实用价值，未在国内外学术期刊或会议发表过。发表论文的作者需提交全文进行同行评审，只做报告不发表论文的作者只需提交摘要。

2、作者可通过 CrossCheck, Turnitin 或其他查询体统自费查重，否则由文章重复率引起的被拒搞将由作者自行承担责任。涉嫌抄袭的论文将不被出版，且公布在会议主页。

3、论文需按照会议官网的模板排版，不得少于 4 页。

4、本次论文直接由出版社安排审稿，一旦被录用，均可被发表和检索。

征文主题

(1) 智能计算

(2) 信号处理

(3) 自动化软件工程

(4) 生物信息学与科学计算

(5) 其它相关领域

大会网站以及组委会联系方式

(1) 大会网站: http://www.icicsp.org/

(2) 投稿邮箱: ICICSP@yeah.net

(3) 会务组联系电话（徐老师）：

Tel: +86- 18702044440 (cellphone), +86-020- 29035993 (office phone)

(4) 会务组即时通讯：(QQ) 1571351296

(5)AEIC 理工科学术交流群：219312476

(6)AEIC 官网：http://www.keoaeic.org