噪音特征对聚类内部有效性的影响

杨 虎[1] 付 宇[2] 范 丹[1]
(中央财经大学信息学院 北京 100081)[1] (中国人民大学统计学院 北京 100872)[2]

摘 要 聚类内部有效性指标是在未知样本真实分类情况下用于评价聚类结果优劣、寻找最佳聚类个数的指标,是聚类分析研究中的重要内容。虽然已有大量的研究分析了聚类内部有效性指标的性能,且有研究结论表明某些内部有效性指标的性能良好,能够辅助聚类算法找到最佳聚类个数,但这些研究未考虑真实数据中的噪音特征对内部有效性指标的影响,研究结论可能会误导内部有效性指标的选取和应用。为此,选取了10种常用的内部有效性指标来研究噪音特征对内部有效性特征选择和聚类结果的影响。结果表明,数据中的噪音特征会影响内部有效性指标的性能,除KL指标、CH指标和CCC指标对噪音特征的反应相对不敏感外,其他内部有效性指标均对噪音特征敏感,且聚类结果的准确性会随着噪音的增强而降低。

关键词 内部有效性,噪音特征,聚类个数,聚类准确度

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.07.004

Influence of Noisy Features on Internal Validation of Clustering

YANG Hu[1] FU Yu[2] FAN Dan[1]
(School of Information, Central University of Finance and Economics, Beijing 100081, China)[1]
(School of Statistics, Renmin University of China, Beijing 100872, China)[2]

Abstract Internal validation measures of clustering are extremly essential in clustering analysis, and they are used to evaluate the effect of clustering results and are indicators to find the optimal cluster number when the true situation of sample is unknown. Although a large number of studies focus on the performance of internal validation measures of clustering and have found that some measures perform better than others, they ignore the influence of noisy features existing in real data. Therefore, it may mislead the selection and application of internal validation measures of clustering. This study selected 10 clustering validation measures to determine the number of clusters of simulation datasets and real datasets, so as to analyze the influence of noisy features on internal validation choosing and clustering results. Results indicate that noisy features among dataset have impact on all internal validation indices of clustering but KL, CH and CCC, and accuracy of the clustering results will decrease along with the increase of noise.

Keywords Internal validation, Noisy features, Number of clusters, Clustering accuracy

1 引言

聚类分析是一种无监督学习算法,它借助样本之间的相似度把没有标记的样本分配到所属类别中,使得相同类别的样本之间的相似程度高,不同类别的样本之间的相似程度低。聚类分析已被广泛用于分析互联网、金融、卫生、健康等领域的数据并识别数据中蕴含的潜在结构信息[1-3]。由于聚类分析所处理的数据一般没有标记,数据的潜在分类个数未知,因此不存在最优的聚类算法能够正确地划分数据,并且在没有经验信息的情况下还面临着寻找最佳聚类个数的难题[4]。

启发式搜索是确定最佳聚类个数最常用的办法,从最小聚类个数开始搜索,直到最大聚类个数,取其中使得评价指标最小或最大的聚类个数作为最佳聚类个数,并输出对应的聚类结果。在评价每一种聚类个数对应的聚类划分结果与数据潜在结构的匹配程度时,需要借助聚类有效性指标。常见的聚类有效性指标主要有内部有效性指标、外部有效性指标和相对有效性指标3种[5-6]。外在聚类有效性指标由于依赖于样本的真实分类,因此很难用于确定最佳聚类个数,但可用于

收稿日期:2017-06-25 返修日期:2017-09-27 本文受国家自然科学基金青年科学基金项目(71701223)资助。

杨 虎(1983-),男,博士,副教授,主要研究领域为大数据分析与统计计算、聚类分析算法,E-mail:hu.yang@cufe.edu.cn(通信作者);付 宇(1995-),男,硕士生,主要研究领域为聚类分析,E-mail:386340673@qq.com;范 丹(1982-),女,博士,讲师,主要研究领域为自适应建模、时间序列分析,E-mail:fandan@cufe.edu.cn。

评价聚类算法的准确度。相对有效性指标在聚类之前需要确定一个决策目标,然后基于该决策目标评价不同参数集下的聚类算法结果,从而确定最优聚类划分结果[6-8]。与外部有效性指标相似,相对有效性指标也依赖于经验知识。内部有效性指标利用聚类紧致性和分离性来判断聚类结果与数据的潜在分类之间的差异性,不依赖于经验知识,且与聚类算法的原理一致,是确定最佳聚类个数较为常用的指标[9-10]。

虽然内部有效性指标在评价聚类划分结果和确定聚类个数方面不依赖于先验知识,只依赖于数据结构,但由于其构建过程缺乏理论支撑,导致不存在一种通用的、普适的内部有效性指标能够评价大多数数据的聚类结果。因此,面对不同的数据和应用场景,产生了与之对应的内部有效性指标,研究不同的内部有效性指标的聚类检验效率自然成为了学术界的研究热点。为了选取一种较为通用的聚类有效性指标,Liu 等设置了 5 种聚类仿真实验场景,利用 k-means 算法对仿真数据进行聚类,然后利用 11 种常用的内部有效性指标确定聚类个数并评测算法的聚类效果,结果表明 SDbw 是 11 种内部有效性指标中效果最好的[11]。此后,Giancarlo 等考虑到聚类的稳定性以及聚类算法在基因数据中的应用情况,也通过仿真实验检测了 30 种内部有效性指标对聚类结果的评测效果,并根据实验结果选出了性能最优的 6 个内部有效性指标[12]。考虑到实际应用中的数据可能会包含来自样本的噪音,从而影响内部有效性指标的检验效率,Gurrutxaga 等[13]设计了聚类结果相似性度量方法,用于评测内部有效性指标之间的检验效果。此外,针对不同的聚类算法,设计不同的评测指标也是聚类有效性的研究重点,例如设计基于原型的聚类有效性评测指标[8]。

在大数据背景下,聚类分析还面临着来自聚类特征的干扰,即数据中有可能包含一些与聚类分析结果不相关或影响聚类分析结果的特征,这些特征被称为噪音特征(Noisy Features),简称噪音[14-15]。从内部有效性指标的构造情况来看,其大部分依赖于样本之间的相似度或密度度量,部分还依赖于样本数据。在这种情况下,噪音特征显然会影响相似度或密度度量,从而影响内在的有效性指标。常见的内部有效性指标有:Dunn 指标(基于聚类之间的距离和聚类的直径构造的内在聚类有效性指标[16])、Calinski-Harabasz(CH)指标、(基于全部样本的聚类内离差矩阵和聚类间离差矩阵构造的内在聚类有效性指标[17])、McClain 指标(基于聚类间平均距离和聚类内平均距离构造的内在聚类有效性指标[18])、Davies-Bouldin (DB)指标(基于样本的聚类内离散测度和各聚类中心间距的测度构造的内在聚类有效性指标[19])、CCC 指标(基于样本平方和和交叉乘积矩阵构造的三次方内在聚类有效性指标[20])、Silhouette 指标(基于样本之间的不相似度构造的内在聚类有效性指标[21])、Krzanowski-Lai(KL)指标(基于所有样本的类内离差矩阵构造的内在聚类有效性指标[22])、Xie 和 Beni(XB)(基于类内距离和类间距离的比值型构造的模糊聚类有效性指标[23])等。这些指标都属于基于聚类、离差平方和构造的内在聚类有效性指标。此外,还有基于样本密度构造的指标,如 SD 指标(基于聚类的类内样本分离程度和类间样本分离性构造的内在聚类有效性指标[24])、SD-bw 指标(考虑了样本聚类密度的基于 SD 的拓展指标[25])。

由上述内容可知,大部分内部有效性指标的构造过程都忽略了噪音特征的影响。在未知真实聚类个数的情况下,从诸多内部有效性指标中选取一种评价指标来确定聚类个数和评价聚类结果是困难的。虽然已有许多学者尝试对各种内部有效性指标的性能进行对比分析,以找到最佳的内部有效性指标,但值得注意的是,这些研究都忽略了噪音特征在真实数据中存在的事实。在噪音特征的影响下,内部有效性指标将会受到影响。为了研究噪音特征对聚类有效性的影响,本文通过模拟仿真实验生成带有一定比例噪音的仿真数据,通过内部有效性指标确定数据的最佳聚类个数并评估聚类准确性,从而说明噪音特征对聚类内部有效性的影响。为了进一步说明噪音特征对内部有效性指标的影响,本文还对真实数据进行聚类分析,以说明数据中可能存在一些区分度不高的特征会影响聚类分析的结果。

2　相关研究

内部有效性指标主要用于检验聚类结果与数据中潜在的结构是否一致,聚类算法则用于识别数据中潜在的结构特征。基于这一目的,内部有效性指标的检验通常由聚类紧致性和聚类分离性两部分测度组成[6,26]。聚类紧致性主要用于检验同一个聚类中样本之间的相似程度,使相似或距离靠近的样本尽可能地被分在同一组;而聚类分离性则检验不同聚类中样本的异质程度,使得不相似或距离不靠近的样本尽可能地被分在不同的组。这两种测度都可以通过样本之间的相似度或距离来测量。例如,同一个聚类中的样本通过与聚类中心的距离来识别样本之间的紧致性,而通过两个聚类中心之间的距离来测量样本之间的分离性。通过变换测量方式,也可以构造其他指标来测度样本之间的紧致性和分离性。例如:KL 指标是由类内离散矩阵的迹来构造的测度指标,可以用于测量聚类中样本到聚类中心的距离,从而判断聚类结果的效果;SD 指标不仅考虑了聚类中心的紧致性,还考虑了聚类中心的离散特性,从而利用距离和密度测度来构造聚类结果的测量指标。

设聚类紧致性测度为 $intra\_distance(\ )$,聚类分离性测度为 $inter\_distance(\ )$,则内部有效性指标可被概括为[27]:

$$\text{argmax}\ (f_C(intra\_distance, inter\_distance))$$

或

$$\text{argmin}\ (f_C(intra\_distance, inter\_distance))$$

其中,$f_C(\ )$是在给定聚类类别 C 的条件下关于 $intra\_distance$ 和 $inter\_distance$ 的函数。为了获取最佳的聚类个数,确定最佳聚类划分,通过测量两种聚类的个数或两种聚类结

果之间的差异性，可以判断哪种聚类结果更优。常见的内在有效性指标通过两种聚类情况的对比来检验聚类的有效性。

在介绍内部有效性指标之前，首先介绍一些基本的符号和定义。$\boldsymbol{X}$ 表示包括 n 个观测样本的集合 $\boldsymbol{X}=(x_1.,x_2.,\cdots,x_n.)^{\mathrm{T}}$，其中 $x_i.$ 表示第 i 个样本，它包含 p 个观测值 $(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$，简化为 x_i。数据集合 X 也可以用列向量来表示，即 $\boldsymbol{X}=(x._1,x._2,\cdots,x._p)$，其中 $x._j$ 表示第 j 个观测变量，它包含 n 个观测值 $(x_{1j},x_{2j},\cdots,x_{nj})^{\mathrm{T}}$。假设聚类个数为 K，第 k 个聚类 C_k 的中心为：$c_k=(c_{k1},c_{k2},\cdots,c_{kp})^{\mathrm{T}}$。定义二范式 $\|\boldsymbol{y}\|=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}$ 表示向量的内积。$\mathrm{tr}(\boldsymbol{A})$ 用于求矩阵 $\boldsymbol{A}$ 的迹。下面介绍几种常用的内部有效性指标。

1）KL 指标。它是由 Krzanowski 等于 1998 年提出的[22]。在给定聚类数 q 的情况下，将 KL 指标定义如下：

$$KL(q)=|DIFF_q/DIFF_{q+1}|$$

其中，$DIFF_q=(q-1)^{\frac{2}{p}}\mathrm{tr}(W_{q-1})-(q)^{\frac{2}{p}}\mathrm{tr}(W_q)$，$W_q$ 为类内离散矩阵且 $W_q=\sum_{k=1}^{q}\sum_{i\in c_k}(x_i-c_k)(x_i-c_k)^{\mathrm{T}}$。由 KL 指标的构造形式可知，它通过测量在不同聚类个数 q 给定的情况下，样本到聚类中心之间的距离，对比两种聚类结果的差异性，来检验给定聚类个数 q 的聚类有效性。当 KL 取值最大时，取得最优聚类个数 q。

2）CH 指标。它是由 Calinski 等于 1974 年提出的[17]，与 KL 指标不同，CH 指标不仅考虑了类内的离散情况，还考虑了聚类之间的离散情况。将 CH 指标定义如下：

$$CH(q)=\frac{\mathrm{tr}(B_q)}{q-1}\cdot\frac{n-q}{\mathrm{tr}(W_q)}$$

其中，$B_q=\sum_{k=1}^{q}n_k(c_k-\bar{x})(c_k-\bar{x})^{\mathrm{T}}$ 是类间离散矩阵，W_q 与 KL 指标中的定义相同。CH 指标与 KL 指标类似，它的优势在于不仅考虑了聚类内的样本之间的紧致性，同时也考虑了不同聚类之间的分离性。当 CH 取值最大时，取得最优聚类个数 q。

3）Mcclain 指标。它是由 McClain 等于 1975 年提出的[18]，该指标由两部分组成，定义如下：

$$Mcclain(q)=\frac{\bar{S}_w}{\bar{S}_b}=\frac{S_w/N_w}{S_b/N_b}$$

其中，$\bar{S}_w$ 为平均类内距离，$\bar{S}_b$ 为平均类间距离；$S_w=\sum_{k=1}^{q}\sum_{i,j\in C_k}d(x_i,x_j)$ 为类内距离之和，$S_b=\sum_{k=1}^{q}\sum_{l=k+1}^{q}d(x_i,x_j)$ 为类间距离之和；$N_w=\sum_{k=1}^{q}\frac{n_k(n_k-1)}{2}$ 为属于同一个类的两两样本对的数量，$N_b=N_t-N_w=\frac{n(n-1)}{2}-\sum_{k=1}^{q}\frac{n_k(n_k-1)}{2}$ 为属于不同类别的两两样本对的数量。由 Mcclain 指标的构造原理可知，它是聚类紧致性和聚类分离性测度的比。当其类内距离之和较小、类间聚类之和尽可能大，即 Mcclain 取值最小时，取得最优聚类个数 q。

4）DB 指标。它是由 Davies 等于 1979 年提出的[19]，定义如下：

$$DB(q)=\frac{1}{q}\sum_{k=1}^{q}\max_{k\neq l}\left(\frac{\delta_k+\delta_l}{d_{kl}}\right)$$

其中，$k,l=1,2,\cdots,q$ 为聚类数；$d_{kl}=\sqrt[v]{\sum_{j=1}^{p}|c_{kj}-c_{lj}|^v}$ 为聚类中心 c_k 和 c_l 之间的距离，当 $v=2$ 时，d_{kl} 为欧几里得距离；$\delta_k=\sqrt[u]{\frac{1}{n_k}\sum_{i\in c_k}\sum_{j=1}^{p}|x_{ij}-c_{kj}|^u}$ 表示类内离散程度，当 $u=2$ 时 δ_k 表示类 c_k 的内部对象与类中心之间距离的标准差。与 Mcclain 指标类似，DB 指标也同时测度了聚类紧致性和聚类分离性之比，不同点是该指标取最坏情况之和来测量聚类的有效性。当 DB 取值最小时，取得最优聚类个数 q。

5）Silhouette 指标。它是由 Rousseeuw 于 1987 年提出的[21]，定义如下：

$$Sihouette(q)=\frac{1}{n}\sum_{i=1}^{n}S(i)$$

其中，$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$。Sihouette 指标的取值范围为 $[-1,1]$。$a(i)=\frac{1}{n_r-1}\sum_{j\in c_{ri}}d_{ij}$ 表示第 i 个对象与类 C_r 中其他对象之间的平均不相似度；$b(i)=\min_{s\neq r}\{d_{ic_s}\}$，$d_{ic_s}=\frac{1}{n_s}\sum_{j\in c_s}d_{ij}$ 表示第 i 个对象与类 C_s 中其他对象间的平均不相似度。Silhouette 指标通过样本到不同聚类之间的平均距离来测量聚类的有效性。当 Silhouette 取值最大时，取得最优聚类个数 q。

6）Dunn 指标。它是利用聚类间最小距离与聚类内最大距离的比值构造的[16]，定义如下：

$$Dunn(q)=\min_{1<i\leqslant j\leqslant q}d(C_i,C_j)/\max_{1<k\leqslant q}diam(C_k)$$

其中，$d(C_i,C_j)=\min_{x\in c_i,y\in c_j}d(x,y)$ 表示类 C_i 和 C_j 之间的不相似度；$diam(C)=\max_{x,y\in c}d(x,y)$ 表示类的直径，用于度量某个聚类的离散程度。如果一个数据集包含紧致且分离良好的聚类，那么类的直径应该足够小，而类间的距离应该足够大。因此，当 Dunn 取值最大时，取得最优聚类个数 q。

7）CCC 指标（Cubic Clustering Criterion）。它是 SAS 统计软件包中提供的检测统计量[20]，定义如下：

$$CCC=\ln\left[\frac{1-E(R^2)}{1-R^2}\right]\cdot\frac{1}{(0.01+E(R^2))^{1.2}}\sqrt{\frac{np^*}{2}}$$

其中，$R^2=1-\mathrm{tr}(X^{\mathrm{T}}X-\bar{X}^{\mathrm{T}}Z^{\mathrm{T}}Z\bar{X})/\mathrm{tr}(X^{\mathrm{T}}X)$，$E(R^2)=1-\left[\frac{\sum_{j=1}^{p^*}\frac{1}{n+u_j}+\sum_{j=p^*+1}^{p}\frac{u_j^2}{n+u_j}}{\sum_{j=1}^{p}u_j^2}\right]\left[\frac{(n-q)^2}{n}\right]\left[1+\frac{4}{n}\right]$，$\bar{\boldsymbol{X}}=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}$，$\boldsymbol{Z}$ 是 $n\times p$ 阶标识矩阵。当第 i 个观测对象属于第 k 个簇时，$Z_{ik}=1$，反之 $Z_{ik}=0$；$u_j=\frac{s_j}{c}$，s_j 为 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/(n-1)$ 的第 j 个特征值的平方根，$c=\left(\frac{v^*}{q}\right)^{\frac{1}{p^*}}$，$v^*=\prod_{j=1}^{p^*}s_j$，$p^*$ 是比 q 小的最大整数，用于确保 u_{p^*} 比 1 大。当 CCC 的取值最大时，获得最优聚类个数 q。

8）SD 指标。它是基于类的平均紧致程度和类间总离散程度构造的[24]，定义如下：

$$SD(q)=a*Scat(q)+Dis(q)$$

其中，$Scat(q)=\frac{\frac{1}{q}\sum_{k=1}^{q}\|\delta^{(k)}\|}{\|\delta\|}$，$\delta$ 是由数据集中每个变量方差组成的向量，$\delta=(VAR(\boldsymbol{V}_1),VAR(\boldsymbol{V}_2),\cdots,VAR(\boldsymbol{V}_p))$，$\delta^{(k)}$ 是每个聚类 C_k 的方差向量，$Scat(q)$ 用于测量类内的平均紧致程度。$Dis(q)=\frac{D_{\max}}{D_{\min}}\sum_{k=1}^{q}(\sum_{z=1}^{q}\|c_k-c_z\|)^{-1}(k,z\in\{1,2,\cdots,q\}$，$D_{\max}=\max(\|c_k-c_z\|)$ 是各聚类中心间的最大距离，$D_{\min}=\min(\|c_k-c_z\|)$ 是各聚类中心间的最小距离，$Dis(q)$ 测度聚类之间总的离散程度。$\alpha=Dis(q_{\max})$ 是权重，用于调节紧致程度的占比。当 SD 的取值最小时，取得最优聚类个数 q。

9）SDbw 指标。它是基于类间紧致性和分离性构造的[25]，定义如下：

$$SDbw(q)=Scat(q)+Density.bw(q)$$

其中，$Scat(q)$ 与 SD 中的定义一致，$Density.bw(q)$ 用于测量聚类之间的密度 $Density.bw(q)=\frac{1}{q*(q-1)}\sum_{i=1}^{q}(\sum_{j=1,i\neq j}^{q}S_{ij})$，$S_{ij}=density(u_{ij})/\max(density(c_i),density(c_j))$，$density(u_{ij})=\sum_{l=1}^{n_{ij}}f(x_l,u_{ij})$，$Stedv=\frac{1}{q}\sqrt{\sum_{k=1}^{q}\|\sigma^{(k)}\|}$，$u_{ij}$ 是由 c_i 和 c_j 定义的线段中点，n_{ij} 是属于类 C_i 和 C_j 的元组数。如果 $d(x,u_{ij})>Stedv$，则 $f(x_i,u_{ij})=0$，否则 $f(x_i,u_{ij})=1$。当 SDbw 的取值最小时，取得最优聚类个数 q。

10）XB 指标。Xie 等提出了比值型模糊聚类有效性指标 XB，它用各样本到聚类中心的距离之和来度量聚类内的紧致性，用所有聚类中心之间距离的最小值来度量聚类间的分离性[23]。XB 指标的构造如下：

$$XB(q)=\frac{\sum_{i=1}^{q}\sum_{x\in c_i}u_{ij}^{m}d^2(x,c_i)}{n\cdot\min_{i,j\neq i}d^2(c_i,c_j)}$$

其中，$d^2(c_i,c_j)$ 表示类间分离度，$d^2(x,c_i)/n$ 表示类间紧致性。当 XB 的取值最小时，取得最优聚类个数 q。

此外，还有很多评价聚类结果有效性的内部有效性指标，如 Friedman 指标（基于聚类类内离差矩阵和聚类之间离差矩阵构造的内在聚类有效性指标[28]）、Scott 指标（基于样本数据的总离差平方和及聚类之间的离差平方和提出的内在聚类有效性指标[29]）、C 指标（基于所有样本点之间的距离提出的内在聚类有效性指标[30]）、Ptbiserial 指标（基于原始相异矩阵和对应的 0-1 矩阵构造的点二系列系数内在聚类有效性指标[31]）。

通过分析内部有效性指标的形式可知，内部有效性指标不仅依赖于样本之间的相似度或距离测度，还依赖于给定的数据。数据中与聚类分析结果不相关或影响聚类分析结果的噪音特征会影响内部有效性指标，从而影响最优聚类个数 K 的确定，进而影响聚类结果的准确性。

3　研究方案

由于 k-means 聚类算法的过程简单且便于理解[33]，本文选取 k-means 聚类算法来对仿真数据和真实数据进行聚类分析，进而研究噪音特征是否会影响聚类内部有效性指标。考虑到 k-means 算法需要确定未知聚类个数 K，本文通过内部有效性指标来辅助 k-means 聚类算法确定最佳聚类个数，从而评测噪音特征是否会影响聚类内部有效性。

3.1　k-means 聚类算法

在给定聚类个数 K 的情况下，聚类算法的目的是把 n 个样本划分（或投影）到 K 个聚类中[32]，使得聚类之间的差异尽可能大、聚内内的差异尽可能小。假设第 k 个聚类的中心为 $c_k=(c_{k1},c_{k2},\cdots,c_{kp})$，它包含 n_k 个样本，并用 $d(x_k,x_l)$ 来度量两个样本点之间的不相似程度，且 $d(x_k,x_l)\in[0,+\infty)$，那么可将 k-means 聚类算法形式化如下：

$$\arg\min_{\boldsymbol{C}}\sum_{k=1}^{K}\sum_{x_i\in c_k}d(x_i,c_k)$$

该算法的计算过程如算法 1 所示。

算法 1　k-means 聚类算法

输入：观测样本数 n，聚类个数 K

输出：样本聚类

Step 1　从 n 个观测对象中任意选择 K 个样本作为初始聚类中心；

Step 2　计算每个样本与聚类中心之间的距离，并根据最小距离的原则重新划分样本；

Step 3　重新计算每个聚类的中心；

Step 4　当聚类结果收敛时终止算法，否则转 Step 2。

由算法的结构可知，k-means 聚类算法通常包括两个步骤：1）在假定聚类中心的情况下分配样本，使得损失函数最小；2）在第 1）步完成的情况下，给定聚类类别，重新计算聚类中心。重复上述两个步骤，直到聚类中心的变化幅度小于给定阈值[33]。

3.2　聚类个数的确定

由于 k-means 算法的聚类个数未知，需要借助内部有效性指标来确定聚类个数。识别最优聚类个数的目的是最优化内部有效性指标 $f_C(\ )$。优化过程主要包含 4 个步骤：1）在给定聚类个数 K 的取值范围 $[K_{\min},K_{\max}]$ 中挑选一个聚类个数；2）利用聚类算法分析观测样本集合；3）判断循环是否结束；4）选取使得内部有效性指标 $f_C(\ )$ 最大或最小时对应的聚类数 K 为最佳聚类数，并输出聚类分析结果。在缺乏经验知识和对数据的了解的情况下，K 的下限 $K_{\min}$ 取最小值 2，上限可以取样本量的开方，即 $K_{\max}\leqslant\sqrt{n}$，其中 n 为观测样本的数目。求解最佳聚类个数的算法如算法 2 所示。

算法 2　最佳聚类个数求解算法

输入：观测样本个数 n，聚类个数范围

输出：最佳聚类个数及聚类划分

Step 1　依次从 $K_{\min}$ 到 $K_{\max}$ 中选取一个数作为聚类个数 K；

Step 2　给定聚类个数 K 的情况下，调用算法 1，得到 n 个样本的 K 个划分；

Step 3 计算内部有效性指标，如果聚类个数 K 等于 K_{max} 则停止，否则转 Step 1；

Step 4 从 $f_{K_{min}}$，…，$f_{K_{max}}$ 中选取最优的内部有效性指标对应的聚类个数作为最佳聚类数并返回结果。

3.3 实验数据

本文的实验数据包括仿真数据和真实数据两部分。

3.3.1 仿真数据

仿真数据借助 R 软件生成。仿真数据集包括 n 个观测样本，属于 K 个聚类。每个数据集的特征由两部分组成，即 p_s 个信号特征和 p_n 个噪音特征(噪音)。每个聚类的有效信息独立并由 p_s 个信号变量随机产生；而 p_n 个噪音产生的数据则为干扰信息。仿真数据的产生根据文献[11]中的 5 种场景设置，包括：聚类之间完全分离、聚类之间完全分离但包括干扰数据、聚类密度不相等、聚类大小不同，以及聚类中包含子类(大的聚类中包含小的子类)。

本文在仿真实验数据中适当加入噪音特征，以测量噪音特征是否会影响聚类内部的有效性。噪音特征的设置情况包括 4 种：1)无噪音特征；2)6 个噪音特征；3)12 个噪音特征；4)24 个噪音特征。仿真数据的噪音特征的设置情况如表 1 所列。

表 1 仿真数据的设置

Table 1 Setting of simulated data

场景	信号数	噪音数	样本数	聚类数	离群点数
聚类之间完全分离	6	0	400	4	0
	6	6	400	4	0
	6	12	400	4	0
	6	24	400	4	0
聚类之间完全分离但包括干扰数据	6	0	430	4	30
	6	6	430	4	30
	6	12	430	4	30
	6	24	430	4	30
聚类密度不相等	6	0	500	3	0
	6	6	500	3	0
	6	12	500	3	0
	6	24	500	3	0
聚类大小不同	6	0	500	3	0
	6	6	500	3	0
	6	12	500	3	0
	6	24	500	3	0
聚类中包含子类	6	0	500	5	0
	6	6	500	5	0
	6	12	500	5	0
	6	24	500	5	0

注：信号数表示除噪音特征外的其他特征的个数，噪音数表示噪音特征的个数

3.3.2 真实数据

本文还选取了 UCI 数据库[1)]中 3 个常用的数据集来评测噪音特征对聚类内部有效性的影响。3 个数据集分别为：鸢尾花数据集(450 个样本、4 个特征、3 个聚类)，红酒数据集(178 个样本、13 个特征、3 个聚类)，种子数据集(210 个样本、7 个特征、3 个聚类)。

1) https://archive.ics.uci.edu/ml/datasets.html

4 实验结果

4.1 实验平台

本文采用 R3.3.0 软件作为实验平台，操作系统是 Win10 企业版，Intel(R) Core i7-3520M CPU@ 2.90 GHz(双核) 64 位，8 GB 内存，120 GB 固态硬盘。采用 R package‘mvtnorm’产生多元正态分布数据，一元正态分布和均匀分布由 R 内置函数产生。

4.2 评价指标

由于本文通过内部有效性指标辅助聚类算法来确定最优聚类个数 K，因此聚类的结果依赖于内部有效性指标和聚类算法。为了评价噪音对聚类内在有效性的影响，本文采用最优聚类个数的平均数来衡量最优聚类个数 K 的准确程度，并给出相应的内部有效性指标的取值。同时，采用 Jaccard 指数[34]、F-score[35] 和 Purity(聚类纯度)[36] 3 种外部有效性指标来进一步确定聚类算法的准确程度。3 种聚类外在有效性指标的取值范围为 0～1，数值越接近 1，说明聚类准确度越高，反之则聚类准确度越低。

假定 a 代表任意两个样本(样本对)属于同一个估计聚类且属于同一个真实聚类的样本数量；b 代表任意两个样本属于同一个估计聚类但不属于同一个真实聚类的样本量；c 代表任意两个样本同时属于一个真实聚类但不属于同一个估计聚类的样本量；d 代表任意两个样本不属于同一个估计聚类也不属于同一个真实聚类的样本量。Jaccard 指数的定义如下：

$$Jaccard=a/(a+b+c)$$

假定样本正类判对样本数为 TP，正类判错样本数为 FP，负类判对样本数为 TN，负类判错样本数为 FN，定义精度 $precision=TP/(TP+FP)$，召回率 $recall=TP/(TP+FN)$ 表示聚类错误率，则将 F-score 定义如下：

$$F\text{-}score=2\cdot precision\cdot recall/(precision+recall)$$

假定 w_i 是第 i 个样本的真实类别，c_i 是第 i 个样本的聚类类别，K 是数据的真实分类个数，Q 是给定的聚类个数，则将纯度 Purity 定义如下：

$$Purity=\frac{1}{n}\sum_{k=1}^{K}\max_{j}|w_k\cap c_j|$$

其中，$\max_{j}|\omega_k\cap c_j|$ 表示第 j 个聚类类别中属于第 k 个真实聚类的最大样本个数，j 的取值范围为 1～Q。

4.3 仿真实验结果

利用 k-means 聚类算法分析仿真数据得到的结果。由于 k-means 聚类算法需要给定聚类个数 K，因此使用 10 种内部有效性指标来辅助 k-means 聚类算法确定聚类个数，并通过 Jaccard，F-score 和 Purity 来评价给定聚类个数条件下聚类结果的准确性。为了得到稳定的结果，在每种仿真实验场景下产生 100 个数据集，再使用聚类算法和内部有效性指标对数

据进行聚类分析,然后选取最优聚类个数,并计算聚类准确度指标 Jaccard,F-score 和 Purity。将 100 次结果的平均值进行整理汇总,如表 2－表 6 所列,所有指标在表中均用小写字母表示。

表 2　在聚类之间完全分离情况下的实验结果(包含 4 个聚类和 6 个信号特征)

Table 2　Experimental results when clusters are separated well(inculding 4 clusters and 6 signal features)

评价指标	噪音特征数	聚类内在有效性指标									
		kl	*ch*	*mcclain*	*db*	*sihouette*	*dunn*	*ccc*	*sd*	*sdbw*	*xb*
最优聚类数均值	无噪音特征	4.82	4.57	3.80	3.75	4.24	3.13	4.79	3.46	12.00	3.98
	6 个噪音特征	4.80	4.58	2.52	3.42	4.09	3.10	4.72	3.42	12.00	3.92
	12 个噪音特征	4.70	4.57	2.19	3.33	4.02	3.11	4.70	3.39	11.98	3.90
	24 个噪音特征	4.52	4.58	2.09	3.14	3.89	3.08	4.66	3.30	11.98	3.79
Jaccard	无噪音特征	0.90	0.93	0.79	0.84	0.89	0.74	0.90	0.80	0.38	0.99
	6 个噪音特征	0.90	0.93	0.54	0.79	0.88	0.73	0.91	0.78	0.37	0.97
	12 个噪音特征	0.91	0.93	0.47	0.77	0.87	0.73	0.91	0.78	0.36	0.97
	24 个噪音特征	0.91	0.93	0.44	0.73	0.85	0.73	0.92	0.77	0.36	0.93
F-score	无噪音特征	0.99	1.00	0.90	0.90	0.96	0.99	1.00	0.85	0.42	1.00
	6 个噪音特征	1.00	1.00	0.71	0.84	0.94	1.00	1.00	0.84	0.42	0.99
	12 个噪音特征	1.00	1.00	0.66	0.84	0.94	1.00	1.00	0.85	0.42	0.99
	24 个噪音特征	0.99	1.00	0.63	0.81	0.93	0.99	1.00	0.83	0.42	0.98
Purity	无噪音特征	1.00	1.00	0.86	0.89	0.96	0.78	1.00	0.85	1.00	1.00
	6 个噪音特征	1.00	1.00	0.61	0.84	0.94	0.78	1.00	0.84	1.00	0.98
	12 个噪音特征	1.00	1.00	0.54	0.83	0.93	0.78	1.00	0.84	1.00	0.98
	24 个噪音特征	0.99	1.00	0.51	0.78	0.91	0.77	1.00	0.82	1.00	0.95

表 3　在聚类之间完全分离但有干扰情况下的实验结果(包含 4 个聚类和 6 个信号特征)

Table 3　Experimental results when clusters are separated well but with some outliers(inculding 4 clusters and 6 signal features)

评价指标	噪音特征数	聚类内在有效性指标									
		kl	*ch*	*mcclain*	*db*	*sihouette*	*dunn*	*ccc*	*sd*	*sdbw*	*xb*
最优聚类数均值	无噪音特征	6.04	5.11	3.96	3.36	5.62	4.53	5.42	4.17	4.45	4.49
	6 个噪音特征	6.05	4.88	2.47	3.36	5.39	4.31	5.50	4.15	5.79	4.15
	12 个噪音特征	6.49	4.86	2.22	3.33	5.27	4.38	5.15	4.14	6.38	3.97
	24 个噪音特征	6.54	4.57	2.03	3.29	4.96	4.38	4.92	4.15	6.22	3.73
Jaccard	无噪音特征	0.79	0.92	0.62	0.68	0.92	0.74	0.92	0.83	0.86	0.89
	6 个噪音特征	0.84	0.91	0.45	0.68	0.90	0.71	0.91	0.83	0.78	0.85
	12 个噪音特征	0.84	0.91	0.41	0.67	0.90	0.72	0.91	0.83	0.75	0.83
	24 个噪音特征	0.78	0.88	0.39	0.66	0.88	0.73	0.91	0.83	0.75	0.77
F-score	无噪音特征	0.91	0.98	0.78	0.83	0.98	0.91	0.98	0.92	0.95	0.97
	6 个噪音特征	0.94	0.98	0.65	0.83	0.97	0.94	0.98	0.92	0.88	0.96
	12 个噪音特征	0.92	0.98	0.62	0.82	0.96	0.92	0.98	0.92	0.85	0.94
	24 个噪音特征	0.87	0.97	0.61	0.82	0.95	0.87	0.98	0.92	0.88	0.92
Purity	无噪音特征	0.91	0.97	0.69	0.77	0.97	0.80	0.97	0.90	0.93	0.94
	6 个噪音特征	0.94	0.96	0.53	0.77	0.95	0.78	0.97	0.90	0.94	0.91
	12 个噪音特征	0.96	0.96	0.49	0.76	0.95	0.79	0.97	0.89	0.94	0.90
	24 个噪音特征	0.93	0.94	0.47	0.75	0.93	0.80	0.96	0.90	0.94	0.85

表 4　在聚类密度不等情况下的实验结果(包含 3 个聚类和 6 个信号特征)

Table 4　Experimental results when the densities of clusters are different(inculding 3 clusters and 6 signal features)

评价指标	噪音特征数	聚类内在有效性指标									
		kl	*ch*	*mcclain*	*db*	*sihouette*	*dunn*	*ccc*	*sd*	*sdbw*	*xb*
最优聚类数均值	无噪音特征	5.20	4.95	2.82	2.96	3.07	2.43	4.61	3.07	11.97	2.94
	6 个噪音特征	4.98	3.44	2.42	2.61	2.81	2.43	3.93	2.93	12.00	2.90
	12 个噪音特征	4.35	3.44	2.28	2.49	2.65	2.43	3.83	2.79	11.99	2.86
	24 个噪音特征	3.62	3.37	2.23	2.44	2.56	2.43	3.67	2.73	11.99	2.81
Jaccard	无噪音特征	0.57	0.69	0.49	0.52	0.54	0.45	0.68	0.54	0.37	0.54
	6 个噪音特征	0.57	0.59	0.44	0.47	0.50	0.45	0.62	0.52	0.36	0.53
	12 个噪音特征	0.59	0.59	0.42	0.46	0.48	0.45	0.61	0.50	0.36	0.53
	24 个噪音特征	0.58	0.58	0.41	0.45	0.46	0.45	0.60	0.49	0.36	0.52
F-score	无噪音特征	0.95	1.00	0.87	0.90	0.91	0.95	1.00	0.91	0.68	0.98
	6 个噪音特征	0.96	1.00	0.76	0.82	0.85	0.96	1.00	0.88	0.68	0.97
	12 个噪音特征	0.99	1.00	0.71	0.79	0.82	0.99	1.00	0.86	0.69	0.95
	24 个噪音特征	0.98	0.98	0.68	0.77	0.79	0.98	0.99	0.85	0.69	0.93
Purity	无噪音特征	0.76	0.86	0.56	0.58	0.60	0.49	0.83	0.60	0.91	0.59
	6 个噪音特征	0.74	0.67	0.48	0.52	0.56	0.49	0.74	0.58	0.91	0.58
	12 个噪音特征	0.72	0.67	0.46	0.50	0.53	0.49	0.73	0.55	0.90	0.57
	24 个噪音特征	0.68	0.66	0.45	0.49	0.51	0.49	0.71	0.54	0.89	0.56

表 5 在聚类大小不等情况下的实验结果(包含 3 个聚类和 6 个信号特征)

Table 5 Experimental results when the sizes of clusters are different (inculding 3 clusters and 6 signal features)

评价指标	噪音特征数	聚类内在有效性指标									
		kl	*ch*	*mcclain*	*db*	*sihouette*	*dunn*	*ccc*	*sd*	*sdbw*	*xb*
最优聚类数均值	无噪音特征	4.35	2.75	2.15	2.30	2.30	2.30	2.90	2.40	11.95	2.35
	6 个噪音特征	5.45	2.50	2.15	2.30	2.35	2.30	2.60	2.40	12.00	2.25
	12 个噪音特征	4.40	2.50	2.10	2.25	2.25	2.60	2.60	2.40	12.00	2.20
	24 个噪音特征	4.95	2.25	2.00	2.20	2.20	2.55	2.55	2.30	11.95	2.15
Jaccard	无噪音特征	0.63	0.86	0.75	0.79	0.79	0.79	0.88	0.83	0.20	0.80
	6 个噪音特征	0.53	0.86	0.76	0.79	0.81	0.80	0.88	0.83	0.20	0.78
	12 个噪音特征	0.62	0.86	0.75	0.78	0.79	0.78	0.88	0.83	0.20	0.76
	24 个噪音特征	0.63	0.78	0.72	0.77	0.77	0.76	0.88	0.79	0.20	0.73
F-score	无噪音特征	0.70	0.94	0.91	0.93	0.93	0.70	0.95	0.94	0.29	0.93
	6 个噪音特征	0.61	0.96	0.92	0.93	0.93	0.61	0.97	0.94	0.29	0.93
	12 个噪音特征	0.70	0.96	0.92	0.93	0.93	0.70	0.97	0.94	0.28	0.90
	24 个噪音特征	0.70	0.93	0.91	0.92	0.92	0.70	0.96	0.93	0.29	0.88
Purity	无噪音特征	0.95	0.92	0.83	0.86	0.86	0.86	0.94	0.88	0.98	0.87
	6 个噪音特征	0.94	0.90	0.83	0.86	0.87	0.86	0.92	0.88	0.98	0.85
	12 个噪音特征	0.94	0.90	0.82	0.85	0.85	0.86	0.92	0.88	0.97	0.83
	24 个噪音特征	0.93	0.85	0.80	0.84	0.84	0.85	0.91	0.86	0.98	0.81

表 6 在大类中包含子类情况下的实验结果(包含 5 个聚类和 6 个信号特征)

Table 6 Experimental results when larger clusters include subclusters(inculding 5 clusters and 6 signal features)

评价指标	噪音特征数	聚类内在有效性指标									
		kl	*ch*	*mcclain*	*db*	*sihouette*	*dunn*	*ccc*	*sd*	*sdbw*	*xb*
最优聚类数均值	无噪音特征	3.92	3.37	3.04	2.72	3.26	2.66	3.91	2.98	11.18	2.88
	6 个噪音特征	3.92	3.36	2.69	2.60	3.15	2.65	3.81	2.72	11.23	2.87
	12 个噪音特征	3.89	3.37	2.45	2.51	3.01	2.57	3.75	2.66	11.28	2.75
	24 个噪音特征	3.54	3.35	2.25	2.46	2.87	2.61	3.64	2.63	11.31	2.67
Jaccard	无噪音特征	0.90	0.98	0.91	0.88	0.96	0.86	0.78	0.91	0.24	0.98
	6 个噪音特征	0.90	0.98	0.85	0.87	0.95	0.85	0.81	0.88	0.23	0.97
	12 个噪音特征	0.92	0.98	0.80	0.84	0.92	0.85	0.84	0.87	0.23	0.94
	24 个噪音特征	0.96	0.98	0.76	0.83	0.89	0.84	0.88	0.87	0.23	0.92
F-score	无噪音特征	0.99	1.00	0.93	0.83	0.98	0.99	1.00	0.88	0.94	0.97
	6 个噪音特征	0.99	1.00	0.83	0.82	0.95	0.99	1.00	0.85	0.92	0.96
	12 个噪音特征	0.99	1.00	0.76	0.82	0.92	0.99	1.00	0.84	0.91	0.94
	24 个噪音特征	1.00	0.99	0.67	0.81	0.89	1.00	1.00	0.83	0.90	0.92
Purity	无噪音特征	1.00	1.00	0.96	0.93	0.99	0.91	1.00	0.96	1.00	0.98
	6 个噪音特征	1.00	1.00	0.91	0.92	0.98	0.91	1.00	0.93	1.00	0.97
	12 个噪音特征	1.00	1.00	0.88	0.90	0.96	0.91	1.00	0.93	1.00	0.95
	24 个噪音特征	1.00	0.99	0.84	0.89	0.93	0.90	1.00	0.92	1.00	0.93

从实验结果可以看出,对于聚类之间完全分离、聚类之间完全分离但包含离群点、聚类密度不同、聚类大小不同、聚类包含子类等 5 种仿真实验场景而言,噪音特征的出现会在一定程度上影响一些内在有效性指标,导致基于这些内在有效性指标来确定聚类个数的准确性降低。具体分析后可以发现,噪音特征的增加会使得聚类之间的区分度下降,从而减少最优聚类的个数。在 5 种实验场景下,大部分指标的最优聚类个数均向更少的方向发展,只有 KL 指标和 SDbw 指标存在一点波动。由于 SDbw 的最优聚类个数较多,其波动性不明显。进一步通过 Jaccard 指标、F-score 指标、Purity 指标分析聚类的准确性可以发现,除了 KL 指标、CH 指标、CCC 指标在噪音增加的情况下仍然表现出对噪音不敏感的特性外,其余 7 种指标的聚类准确度均会在噪音增加的情况下出现不同程度的下降。研究结果表明,噪音特征会影响内部有效性指标确认最佳聚类个数的过程,导致最优聚类个数偏离正确的聚类个数,从而导致聚类的准确性下降。

4.4 基于真实数据的结果

为了测试 10 种内部有效性指标的性能,按照特征的个数,在真实数据中增加一定比例的噪音特征,增加的噪音特征数量为信号特征数量的倍数(4 种场景分别为:无噪音特征、增加 1 倍噪音、增加 2 倍噪音、增加 4 倍噪音)。通过增加噪音特征的数量来分析噪音特征对内部有效性指标的影响。噪音特征独立由正态分布产生,正态分布的期望和方差则独立由均匀分布产生,限定期望和方差的取值范围在信号特征均值和方差的最小值与最大值之间。同理,统计 10 种内部有效性指标对应的最优聚类个数,并在给定最优聚类个数的条件下计算对应的 Jaccard,F-score 和 Purity 指标,实验结果如表 7—表 9 所列。

通过分析真实数据聚类结果可知,与模拟实验结果类似,当数据中的噪音特征比例增加时,内部有效性指标在一定程度上会受到影响,聚类算法不能准确确定聚类个数,进而导致 Jaccard 指标、F-score 指标、Purity 指标等聚类准确度下降。对于鸢尾花数据,KL 指标、CH 指标和 CCC 指标在噪音特征较少时,最优聚类个数和聚类准确度相对较高;但随着噪音的增加,聚类准确度下降。对于红酒数据,10 种聚类内在有效性指标的准确度都相对较低,且聚类准确度在噪音特征增加的情况下下降。对于种子数据,KL 指标和 CH 指标在噪音较少时表现较好,聚类准确度相对较高,但也会受到噪音特征

规模的影响。可见,噪音特征规模会影响内部聚类有效性指　　标的性能,这与模拟实验结果基本一致。

表 7　鸢尾花数据集上的实验结果(450 个样本、4 个特征、3 个聚类)

Table 7　Experimental results on Iris data set(450 observations,4 features,3 clusters)

评价指标	噪音特征数	聚类内在有效性指标									
		kl	*ch*	*mcclain*	*db*	*sihouette*	*dunn*	*ccc*	*sd*	*sdbw*	*xb*
最优聚类数均值	无噪音特征	4.00	3.00	2.00	2.00	2.00	4.00	3.00	2.00	11.00	2.00
	增加 1 倍噪音	5.60	2.00	2.00	3.86	2.00	7.36	2.00	5.10	11.99	2.01
	增加 2 倍噪音	5.58	2.00	2.00	5.83	2.00	7.48	2.00	8.40	11.82	2.00
	增加 4 倍噪音	6.21	2.00	2.00	10.77	2.00	8.96	2.00	10.44	11.84	2.00
Jaccard	无噪音特征	0.60	0.70	0.57	0.57	0.57	0.60	0.70	0.57	0.30	0.57
	增加 1 倍噪音	0.39	0.57	0.57	0.50	0.57	0.34	0.57	0.40	0.19	0.56
	增加 2 倍噪音	0.38	0.57	0.57	0.41	0.57	0.32	0.57	0.26	0.17	0.54
	增加 4 倍噪音	0.35	0.57	0.57	0.20	0.57	0.24	0.57	0.18	0.15	0.52
F-score	无噪音特征	1.00	1.00	0.97	0.97	0.97	1.00	1.00	0.97	0.59	0.97
	增加 1 倍噪音	0.77	0.97	0.97	0.86	0.97	0.77	0.97	0.76	0.36	0.95
	增加 2 倍噪音	0.75	0.97	0.97	0.74	0.97	0.75	0.97	0.57	0.37	0.92
	增加 4 倍噪音	0.71	0.97	0.97	0.43	0.97	0.71	0.97	0.44	0.37	0.89
Purity	无噪音特征	0.88	0.89	0.67	0.67	0.67	0.88	0.89	0.67	0.95	0.67
	增加 1 倍噪音	0.75	0.67	0.67	0.69	0.67	0.77	0.67	0.72	0.83	0.67
	增加 2 倍噪音	0.72	0.67	0.67	0.70	0.67	0.74	0.67	0.75	0.78	0.67
	增加 4 倍噪音	0.71	0.67	0.67	0.73	0.67	0.72	0.67	0.73	0.74	0.66

表 8　红酒数据集上的实验结果(178 个样本、13 个特征、3 个聚类)

Table 8　Experimental results on Wine data set(178 observations,13 features,3 clusters)

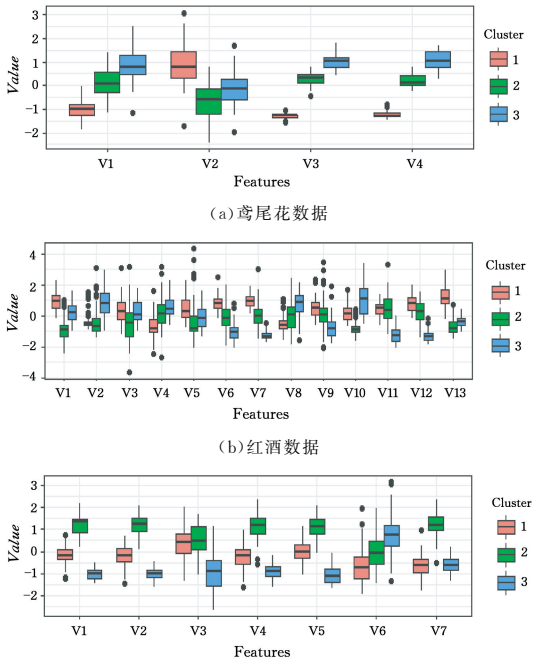
评价指标	噪音特征数	聚类内在有效性指标									
		kl	*ch*	*mcclain*	*db*	*sihouette*	*dunn*	*ccc*	*sd*	*sdbw*	*xb*
最优聚类数均值	无噪音特征	12.00	12.00	2.00	2.00	2.00	4.00	2.00	4.00	11.00	2.00
	增加 1 倍噪音	7.51	2.20	2.00	10.43	2.55	10.19	2.07	7.98	11.87	2.69
	增加 2 倍噪音	7.57	2.12	2.00	11.54	3.22	9.90	2.00	9.66	11.87	2.48
	增加 4 倍噪音	7.91	2.09	2.00	11.83	6.40	10.24	2.00	10.68	11.83	3.39
Jaccard	无噪音特征	0.16	0.31	0.31	0.10	0.23	0.12	0.31	0.11	0.10	0.27
	增加 1 倍噪音	0.18	0.18	0.47	0.47	0.47	0.35	0.47	0.35	0.19	0.46
	增加 2 倍噪音	0.20	0.37	0.37	0.12	0.36	0.15	0.37	0.14	0.12	0.32
	增加 4 倍噪音	0.21	0.42	0.41	0.15	0.42	0.15	0.41	0.19	0.12	0.40
F-score	无噪音特征	0.47	0.47	0.88	0.88	0.88	0.47	0.88	0.61	0.55	0.87
	增加 1 倍噪音	0.48	0.79	0.73	0.36	0.79	0.48	0.72	0.43	0.29	0.69
	增加 2 倍噪音	0.43	0.64	0.63	0.29	0.66	0.43	0.63	0.34	0.29	0.55
	增加 4 倍噪音	0.36	0.52	0.50	0.24	0.43	0.36	0.50	0.28	0.24	0.47
Purity	无噪音特征	0.74	0.74	0.66	0.66	0.66	0.72	0.66	0.72	0.74	0.66
	增加 1 倍噪音	0.65	0.63	0.60	0.66	0.64	0.65	0.60	0.65	0.66	0.62
	增加 2 倍噪音	0.63	0.58	0.57	0.64	0.61	0.63	0.57	0.63	0.64	0.54
	增加 4 倍噪音	0.59	0.51	0.50	0.60	0.58	0.59	0.50	0.59	0.60	0.53

表 9　种子数据集上的实验结果(210 个样本、7 个特征、3 个聚类)

Table 9　Experimental results on Seed data set(210 observations,7 features,3 clusters)

评价指标	噪音特征数	聚类内在有效性指标									
		kl	*ch*	*mcclain*	*db*	*sihouette*	*dunn*	*ccc*	*sd*	*sdbw*	*xb*
最优聚类数均值	无噪音特征	3.00	3.00	2.00	2.00	2.00	8.00	2.00	2.00	11.00	2.00
	增加 1 倍噪音	6.00	2.01	2.00	7.74	2.00	10.20	2.00	7.03	11.91	2.05
	增加 2 倍噪音	6.25	2.00	2.00	9.96	2.00	10.11	2.00	9.34	11.89	2.04
	增加 4 倍噪音	6.22	2.01	2.00	11.69	2.03	10.16	2.00	10.48	11.89	2.03
Jaccard	无噪音特征	0.68	0.68	0.52	0.52	0.52	0.33	0.52	0.52	0.22	0.52
	增加 1 倍噪音	0.35	0.51	0.51	0.30	0.51	0.20	0.51	0.28	0.17	0.50
	增加 2 倍噪音	0.32	0.50	0.50	0.21	0.50	0.18	0.50	0.19	0.15	0.50
	增加 4 倍噪音	0.30	0.49	0.49	0.14	0.49	0.16	0.49	0.15	0.14	0.48
F-score	无噪音特征	0.83	0.83	0.59	0.59	0.59	0.83	0.59	0.59	0.43	0.19
	增加 1 倍噪音	0.38	0.41	0.41	0.36	0.41	0.38	0.41	0.38	0.31	0.31
	增加 2 倍噪音	0.32	0.40	0.40	0.26	0.40	0.32	0.40	0.27	0.24	0.34
	增加 4 倍噪音	0.29	0.38	0.38	0.19	0.38	0.29	0.38	0.21	0.19	0.36
Purity	无噪音特征	0.89	0.89	0.66	0.66	0.66	0.88	0.66	0.66	0.89	0.66
	增加 1 倍噪音	0.72	0.66	0.66	0.73	0.66	0.79	0.66	0.73	0.80	0.67
	增加 2 倍噪音	0.69	0.65	0.65	0.71	0.65	0.72	0.65	0.71	0.74	0.66
	增加 4 倍噪音	0.67	0.65	0.64	0.68	0.65	0.68	0.64	0.68	0.69	0.65

为了进一步说明真实数据的特性,按聚类分组绘制各信号　　特征的箱线图,用于观察信号特征的分布情况,如图 1 所示。



(a)鸢尾花数据

(b)红酒数据

(c)种子数据

图 1 3 种真实数据集中各特征的分布情况图

Fig. 1 Boxplot of each feature grouped by clustering labels for three real data sets

从图 1 可以看出,鸢尾花数据和种子数据的特征在各类别中存在一定的区分度;而红酒数据中则存在一些区分度不高的特征,这些特征有可能是干扰聚类分析的噪音。结合表 6—表 9 也可以看出,红酒数据的聚类准确度较低,这是由于数据中的噪音特征造成的,进一步说明了噪音特征会影响聚类内部有效性指标的性能。

结束语 本文主要通过仿真实验和真实数据分析来研究噪音特征对内部有效性指标性能的影响。实验结果表明,随着噪音数量的增加,一些内部有效性指标在判定聚类个数时出现了越来越大的偏差,导致其聚类的准确性随之下降。内部有效性指标通常由聚类紧致性和聚类分离性两部分组成,而通常采用距离和密度构成的指标来测度紧致性和分离性,这导致噪音会影响内部有效性指标;特别是当内部有效性指标的构造与聚类结果和数据同时相关时,噪音对其的影响会更加明显。当评价指标采用密度来测量聚类紧致性和分离性时,由于密度指标更容易受到干扰,因此聚类个数和聚类的准确度与真实值相比均出现了较大的偏差。可见,噪音特征对内部有效性指标的影响是明显的,它通过影响聚类个数的确定过程来影响最终的聚类结果。因此,在选取和使用内部聚类有效性的过程中,需要考虑噪音特征带来的影响。

虽然本文通过仿真数据和真实数据分析论证了噪音特征会影响内部有效性指标,但要解决噪音特征带来的影响还需要做更多的研究工作。在未来的研究中,可以考虑使用加惩罚的方法来剔除噪音对聚类算法的影响,如稀疏聚类算法[37]或加约束的聚类算法[38]等。除了考虑噪音特征对内部有效性指标的影响外,还需要考虑特征的相关性、样本量等,它们也会影响内部有效性指标的选择和应用,并最终影响聚类分析算法的性能。此外,在研究的过程中,还需要梳理更多的真实数据(例如基因数据),利用内部有效性指标来判别真实数据中的聚类个数,提升聚类算法的准确度,进行更深入系统的研究,以建立一套内部有效性指标选取的准则和依据。

参 考 文 献

[1] LEE J M,SONNHAMMER E L. Genomic gene clustering analysis of pathways in eukaryotes[J]. Genome Research,2003,13(5):875-882.

[2] ZASLAVSKY L,CIUFO S,FEDOROV B,et al. Clustering analysis of proteins from microbial genomes at multiple levels of resolution[J]. Bmc Bioinformatics,2016,17(8):545-552.

[3] LI X,HIPEL K W,DANG Y. An improved grey relational analysis approach for panel data clustering[M]. Oxford:Pergamon Press,Inc. 2015.

[4] ARBELAITZ O,GURRUTXAGA I,MUGUERZA J,et al. An extensive comparative study of cluster validity indices[J]. Pattern Recognition,2013,46(1):243-256.

[5] BEN-DAVID S,LUXBURG U V,PÁL D. A Sober Look at Clustering Stability[J]. Lecture Notes in Computer Science,2006,4005:5-19.

[6] SALEM S A,NANDI A K. Development of assessment criteria for clustering algorithms[M]. Berlin:Springer-Verlag,2009.

[7] BOLSHAKOVA N,AZUAJE F,CUNNINGHAM P. A knowledge-driven approach to cluster validity assessment[J]. Bioinformatics,2005,21(10):2546-2547.

[8] YUE S,WANG J,WANG J,et al. A new validity index for evaluating the clustering results by partitional clustering algorithms[J]. Soft Computing,2016,20(3):1127-1138.

[9] CHAWLA N. Discovering Knowledge in Data:An Introduction to Data Mining[J]. Publications of the American Statistical Association,2014,100(472):1465-1465.

[10] ZHAO Y,KARYPIS G. Evaluation of hierarchical clustering algorithms for document datasets[C]// Eleventh International Conference on Information & Knowledge Management. ACM,2002:515-524.

[11] LIU Y,LI Z,XIONG H,et al. Understanding of Internal Clustering Validation Measures[C]// IEEE,International Conference on Data Mining. IEEE,2011:911-916.

[12] GIANCARLO R,UTRO F. Algorithmic paradigms for stability-based cluster validity and model selection statistical methods,with applications to microarray data analysis[J]. Theoretical Computer Science,2012,428(6):58-79.

[13] GURRUTXAGA I,MUGUERZA J,ARBELAITZ O. Towards a standard methodology to evaluate internal cluster validity indices[J]. Pattern Recognition Letters,2011,32(3):505-515.

[14] JIANG D,TANG C,ZHANG A. Cluster analysis for gene expression data:a survey[J]. IEEE Transactions on Knowledge & Data Engineering,2004,16(11):1370-1386.

[15] SMYTH C,COOMANS D,EVERINGHAM Y. Clustering noisy data in a reduced dimension space via multivariate regression trees[J]. Pattern Recognition,2006,39(3):424-431.

[16] DUNNÂ J C. Well-Separated Clusters and Optimal Fuzzy Partitions[J]. Journal of Cybernetics,1974,4(1):95-104.

[17] CALI′NSKI T,HARABASZ J. A dendrite method for cluster analysis[J]. Communications in Statistics,1974,3(1):1-27.

(下转第 52 页)

算机学报,2011,34(10):1741-1752.

[4] CARSTOIU D,LEPADATU E,GASPAR M. Hbase-non SQL Database,Performances Evaluation[J]. International Journal of Advancements in Computing Technology,2010,2(5):42-52.

[5] ZHOU X,ZHANG X,WANG Y,et al. Efficient Distributed Multi-dimensional Index for Big Data Management[M]// Web-Age Information Management. Springer Berlin Heidelberg,2013:130-141.

[6] TAKASU A. An Efficient Distributed Index for Geospatial Databases[M]// Database and Expert Systems Applications. Springer International Publishing,2015:28-42.

[7] ZHOU X,LI H,ZHANG X,et al. ABR-Tree:An Efficient Distributed Multidimensional Indexing Approach for Massive Data [C]// International Conference on Algorithms and Architectures for Parallel Processing. Springer,Cham,2015:781-790.

[8] HUANG S,WANG B,ZHU J,et al. R-HBase:A Multi-dimensional Indexing Framework for Cloud Computing Environment [C]// IEEE International Conference on Data Mining Workshop. IEEE,2015:569-574.

[9] HUANG S,WANG B,DENG S,et al. HMVR-tree:A Multi-version R-tree Based on HBase for Concurrent Access[M]// Big Data Computing and Communications. Springer International Publishing,2016.

[10] XIA Y,HUANG Z,ZHANG X,et al. Parallel Indexing for Past,Current and Future Locations of Moving Objects[C]// International Conference on Service Sicence,Technology and Engineering. DEStech Publications,2016:20-27.

[11] DU N,ZHAN J,ZHAO M,et al. Spatio-temporal data index model of moving objects on fixed networks using hbase[C]// 2015 IEEE International Conference on Computational Intelligence & Communication Technology (CICT). IEEE,2015:247-251.

[12] GEORGE L. HBase:The Definitive Guide[M]. The People's Posts and Telecommunications Press,2013. (in Chinese)
GEORGE L. Hbase 权威指南[M]. 北京:人民邮电出版社,2013.

[13] Zookeeper[EB/OL]. https://zookeeper. apache. org.

[14] WANG B T,ZHAO K L,CHANG L D,et al. Optimization Technique fot Continuous Range Query Based on Storm[J]. Computer Science and Engineering,2017,39(1):1-14. (in Chinese)
王波涛,赵凯利,常立东,等. 基于 Storm 的连续范围查询优化技术[J]. 计算机工程与科学,2017,39(1):1-14.

[15] Apache Storm[EB/OL]. http://storm. apache. org.

(上接第 30 页)

[18] MCCLAIN J O,RAO V R. CLUSTISZ:A Program to Test for the Quality of Clustering of a Set of Objects[J]. Journal of Marketing Research,1975,12(4):456-460.

[19] DAVIES D L,BOULDIN D W. A cluster separation measure [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,1979,PAMI-1(2):224-227.

[20] INCORPORATED C S I. SAS - C Socket Library for TCP-IP, Release 5. 01:SAS Technical Report C-111[R]. SAS Publishing,1992.

[21] ROUSSEEUW P. Silhouettes:A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational & Applied Mathematics,1987,20(20):53-65.

[22] KRZANOWSKI W J,LAI Y T. A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering[J]. Biometrics,1988,44(1):23-34.

[23] XIE X L,BENI G. A Validity Measure for Fuzzy Clustering[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,1991,13(13):841-847.

[24] HALKIDI M,VAZIRGIANNIS M,BATISTAKIS Y. Quality Scheme Assessment in the Clustering Process[M]// Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg,2000:265-276.

[25] HALKIDI M,VAZIRGIANNIS M. Clustering validity assessment:finding the optimal partitioning ofa data set[C]// IEEE International Conference on Data Mining. IEEE,2001:187-194.

[26] AMORIM R C D,HENNIG C. Recovering the number of clusters in data sets with noise features using feature rescaling factors[J]. Information Science,2015,324:126-145.

[27] CAMPO D N,STEGMAYER G,MILONE D H. A new index for clustering validation with overlapped clusters[J]. Expert Systems with Applications,2016,64(C):549-556.

[28] FRIEDMAN H P,RUBIN J. On Some Invariant Criteria for Grouping Data[J]. Publications of the American Statistical Association,1967,62(320):1159-1178.

[29] SCOTT A J,SYMONS M J. Clustering Methods Based on Likelihood Ratio Criteria[J]. Biometrics,1971,27(2):387-397.

[30] HUBERT L J,LEVIN J R. A general statistical framework for assessing categorical clustering in free recall[J]. Psychological Bulletin,1975,83(6):1072-1080.

[31] MILLIGAN G W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms[J]. Psychometrika,1980,45(3):325-342.

[32] JAIN A K,MURTY M N,FLYNN P J. Data clustering:a review[J]. Acm Computing Surveys,1999,31(3):264-323.

[33] XU R,WUNSCH I D. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks,2005,16(3):645-678.

[34] LAROSE D T. Introduction to Data Mining[M]. Boston:China Machine Press,2010.

[35] SALTON G,HARMAN D. Information retrieval[M]. Chichester:John Wiley and Sons Ltd. ,2003.

[36] MANNING C D,RAGHAVAN P,SCHÜTZE H. An Introduction to Information Retrieval[J]. Journal of the American Society for Information Science & Technology,2008,61(4):852-853.

[37] WITTEN D M,TIBSHIRANI R. A framework for feature selection in clustering[J]. Publications of the American Statistical Association,2010,105(490):713-726.

[38] SUN W,WANG J,FANG Y. Regularized k-means clustering of high-dimensional data and its asymptotic consistency[J]. Electronic Journal of Statistics,2012,6(2):148-167.