DOI:10.11992/tis.201611007
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.TP.20170227.2211.022.html

最近最远得分的聚类性能评价指标

冯柳伟[1,2],常冬霞[1,2],邓勇[3],赵耀[1,2]
(1. 北京交通大学 信息科学研究所,北京 100044;2. 北京交通大学 计算机与信息科学学院,北京 100044; 3.中国科学院 软件研究所,北京 100190)

摘 要:聚类算法是数据分析中广泛使用的方法之一,而类别数往往是决定聚类算法性能的关键。目前,大部分聚类算法需要预先给定类别数,在很多情况下,很难根据数据集的先验知识获得有效的类别数。因此,为了获得数据集的类别数,本文基于最近邻一致性和最远邻相异性的准则,提出了一种最近最远得分评价指标,并在此基础上提出了一种自动确定类别数的聚类算法。实验结果证明了所提评价指标在确定类别数时的有效性和可行性。

关键词:最近邻一致性;最远邻相异性;K-means 聚类算法;评分机制;评价指标;层次聚类
中图分类号: TP391 **文献标志码**:A **文章编号**:1673-4785(2017)01-0067-08

中文引用格式:冯柳伟,常冬霞,邓勇,等.最近最远得分的聚类性能评价指标[J]. 智能系统学报, 2017, 12(1): 67-74.
英文引用格式:**FENG Liuwei, CHANG Dongxia, DENG Yong, et al. A clustering evaluation index based on the nearest and furthest score [J]. CAAI Transactions on Intelligent Systems**, 2017, 12(1): 67-74.

A clustering evaluation index based on the nearest and furthest score

FENG Liuwei [1,2,3], CHANG Dongxia[1,2,3], DENG Yong[4], ZHAO Yao[1,2,3]
(1. Institute of Information Science, Beijing Jiaotong University Beijing 100044, China; 2. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China; 3. Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The clustering algorithm is one of the widely-used methods in data analysis. However, the number of clusters is essential to determine the performance of the clustering algorithm. At present, the number of clusters usually need to be specified in advance. In most cases, it is difficult to obtain the valid cluster number according to a priori knowledge of the dataset. To obtain the number of clusters automatically, a Nearest and Furthest Score (NFS) index was proposed based on the principles of the nearest neighbor consistency and the furthest neighbor difference. Moreover, an Automatic Clustering NFS (ACNFS) algorithm was also proposed, which can determine the number of clusters automatically. The experimental results prove the index is reasonable and practicable to determine the cluster number.

Keywords: the nearest neighbor consistency; the furthest neighbor difference; K-means clustering algorithm; scoring mechanism; evaluation index; hierarchical clustering

聚类算法作为数据分析中广泛使用的主要方法之一,已经广泛应用于模式识别、机器学习、图像处理和数据挖掘等方面[1-4]。简单来说,聚类就是根据数据的特征将数据划分为几类,使得同一类别数据间的相似度尽可能大,而不同类别数据间的相似度则尽可能小。目前,常用聚类算法可以分为划分法、层次法、基于密度的方法、基于网格的方法和基于模型的方法。事实上,很多聚类算法往往需要预先知道聚类问题的类别数。然而,在很多实际情况下,很难根据数据特征获得有效的类别数。因此,为了获得有效的类别数,很多学者基于聚类的

收稿日期:2016-11-07. 网络出版日期:2017-02-27.
基金项目:国家自然科学基金"重点"项目(61532005).
通信作者:常冬霞. E-mail:dxchang@ bjtu.edu.cn.

不同性质分别提出了一系列评价聚类结果的评价指标。对给定范围的类别数依次对数据集进行聚类,并采用评价指标对每次的聚类结果进行评价,然后选择一个使评价指标最优的类别数。目前,常用有效性评价指标大致可以分为 3 种类型,分别是基于数据集模糊划分的指标、基于数据集样本几何结构的指标和基于数据集统计信息的指标。其中,1991 年,Xie 等[5]利用模糊聚类的目标函数,同时考虑数据集本身的结构和模糊隶属度的性质,提出了 Xie-Beni 指标。之后,很多学者基于数据集模糊划分提出了一系列改善的评价指标[6-9],但这些指标不适合对硬聚类算法的聚类结果进行评价。另外一类是基于数据集样本几何结构的评价指标[10-16]。1974 年,Caliński 和 Harabasz 提出了基于全部样本的类内离差矩阵和类间离差矩阵测度的 Caliński-Harabasz(CH)指标[12]。1979 年,Davies 和 Bouldin 提出了基于样本的类内散度与各聚类中心间距离测度的 Davies-Bouldin(DB)指标[13],以及随后提出的基于最大化类内相似度和最小化类间相似度目标的 Weighted inter-intra(Wint)指标[14]、基于样本类内离差矩阵的 Krzanowski-Lai(KL)指标[15]、周世兵等提出的基于样本间的最小类间距离与类内距离的 Between-Within Proportion(BWP)指标[16]。但是这些评价指标均具有一定的局限性,对数据结构无法完全分离的数据集进行评价得到的结果并不理想。2007 年,Kapp 等基于数据集统计的思想,使用类内数据点的 in-group 比例来评价聚类结果,提出了 In-Group Proportion(IGP)评价指标[17]。该评价指标使用样本与其最近邻样本划分到同一类的比例来衡量聚类结果的质量。但是由于 IGP 只关注最近邻一致性,使得 IGP 指标值会随着聚类数的增加而减少,导致在很多实际情况下,利用 IGP 指标得到的类别数往往比实际的类别数小。针对这种情况,本文基于最近邻一致性和最远邻相异性的原则,提出了一种最近最远得分指标(NFS),并基于此指标,提出了一种基于 NFS 指标自动确定类别数的聚类算法。实验结果验证了本文所提的评价指标的有效性和可行性。

1 已有评价指标

众所周知,很多聚类算法需要用户根据先验知识给出算法所需要的类别数。但是,在很多实际应用中很难获得有效的先验知识,因此,确定聚类问题的类别数成为了聚类分析的一个研究的热点。目前传统的确定类别数的方法是根据评价指标来确定类别数。至今提出的评价指标包括 CH 指标、BWP 指标和 IGP 指标等。

1.1 Calinski-Harabasz(CH)指标

CH 指标是 Caliński 和 Harabasz 提出的确定最佳聚类数的评价指标[12]。该指标是一种基于样本的类内距离和类间离差矩阵的测度,其判断函数为

$$\mathrm{CH}(k)=\frac{\mathrm{BGSS}}{K-1}\Big/\frac{\mathrm{WGSS}}{n-K} \tag{1}$$

式中:n 为数据集样本数,K 为类别数。且

$$\mathrm{WGSS}=\frac{1}{2}[(n_1-1)\bar{d}_1^2+\cdots+(n_K-1)\bar{d}_K^2]$$

$$\mathrm{BGSS}=\frac{1}{2}[(K-1)\bar{d}^2+(n-K)A_K]$$

式中:$\bar{d}_j^2$ 是第 j 类中样本间的平均距离,$j=1,2,\cdots,k$;$\bar{d}^2$ 是所有样本间的平均距离。且 A_K 为

$$A_K=\frac{1}{n-K}\sum_{i=1}^{K}(n_i-1)(\bar{d}^2-\bar{d}_i^2)$$

CH 指标值越大表示聚类结果的类内距离越小而类间距离越大,聚类结果性能越好。但是随着类别数搜索范围的变化,CH 指标得到的最佳聚类数会发生变化,并且随着搜索范围增大,CH 指标得到的最佳聚类数有逐渐增大的趋势[18]。

1.2 BWP 指标

BWP 指标是周世兵等人提出的一种基于样本的几何结构设计的确定聚类类别数的评价指标[16],该指标利用聚类结果的类内紧密性和类间分离性来衡量聚类结果。指标的最大值对应的类数作为聚类数。该指标的判断函数为

$$\mathrm{BWP}(K)=\frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n_j}\mathrm{bwp}(i,j) \tag{2}$$

式中:K 是类别数,$\mathrm{bwp}(i,j)$ 为

$$\mathrm{bwp}(i,j)=\frac{b(i,j)-w(i,j)}{b(i,j)+w(i,j)}$$

式中:$b(i,j)$ 是第 j 类中的第 i 个样本到其他每类中样本平均距离的最小值,称为最小类间距;$w(i,j)$ 是第 j 类中的第 i 个样本的类内距离。且

$$b(i,j)=\min_{1\le m\le k,m\ne j}\left(\frac{1}{n_m}\sum_{p=1}^{n_m}\|\boldsymbol{x}_p^{(m)}-\boldsymbol{x}_i^{(j)}\|\right)$$

$$w(i,j)=\frac{1}{n_j-1}\sum_{q=1,q\ne i}^{n_j}\|\boldsymbol{x}_q^{(j)}-\boldsymbol{x}_i^{(j)}\|^2$$

BWP 指标值越大则表示聚类结果的类内越紧密而类间越远离,聚类结果性能越好。但是该评价指标不适合非完全分离的数据集。

1.3 IGP 指标

IGP 指标是 Kapp 提出的评价指标[17]。此指标的设计思想是:当对新的样本进行分类时,新样本

应该被划分到与其最相似的样本所在类别。因此该指标使用样本与其最近邻样本划分到同一类的比例来衡量聚类结果的质量。该指标的评价函数为

$$\mathrm{IGP}(K)=\frac{1}{K}\sum_{u=1}^{K}\mathrm{igp}(u,X) \tag{3}$$

式中:K 是类别数;$\mathrm{igp}(u,X)$ 表示数据集 X 中的第 u 类的指标值,且

$$\mathrm{igp}(u,X)=\frac{\{j\mid \mathrm{Class}_X(j)=\mathrm{Class}_X(j^N)=u\}}{\{j\mid \mathrm{Class}_X(j)=u\}}$$

式中:j^N 是距离样本 j 最近的样本,$\mathrm{Class}_X(j)$ 表示数据集 X 中的第 j 个样本所属的类别。

IGP 指标的值越大表示样本和其最近邻划分到同一类的概率越高,聚类结果越好。但是 IGP 指标只关注了最近邻一致性,使得 IGP 指标值不适合判断非完全分离的数据集。

2　最近最远得分评价指标

为了能准确地得到聚类问题的类别数,本文在基于最近邻一致性和最远邻相异性的原则上,提出了最近最远得分(nearest and furthest score,NFS)评价指标。

2.1　相关概念定义

$\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$ 是一个 n 维矢量空间的有限子集,K 是类别数,$\boldsymbol{C}=\{c_1,c_2,\cdots,c_K\}$ 是聚类算法所得类别中心集合。

定义 1　最近得分。定义 $\mathrm{ns}(i)$ 是第 i 个样本的最近得分,第 j 个样本是距离其最近的样本,若样本 i 与样本 j 属于同一类别,则第 i 个样本的最近得分值为 1;否则其最近得分值为-1,即

$$\mathrm{ns}(i)=\begin{cases}1, & \mathrm{sc}(i)=\mathrm{sc}(j)\\ -1, & \mathrm{sc}(i)\neq\mathrm{sc}(j)\end{cases} \tag{4}$$

式中:$\mathrm{sc}(i)$ 代表第 i 个样本所属的类别,$\mathrm{sc}(j)$ 代表距离样本 i 最近的样本 j 所属的类别。

定义 2　最远得分。定义 $\mathrm{fs}(i)$ 是第 i 个样本的最远得分,第 l 个样本是距离其最远的样本,若样本 i 与样本 l 属于不同类别,则第 i 个样本的最远得分值为 1,否则其最远得分值-1,即

$$\mathrm{fs}(i)=\begin{cases}-1, & \mathrm{sc}(i)=\mathrm{sc}(l)\\ 1, & \mathrm{sc}(i)\neq\mathrm{sc}(l)\end{cases} \tag{5}$$

式中:$\mathrm{sc}(i)$ 代表第 i 个样本所属的类别;$\mathrm{sc}(l)$ 代表距离样本 i 最远的样本 l 所属的类别。

定义 3　样本得分。定义 $s(i)$ 是第 i 个样本的得分值,则第 i 个样本的最近得分和最远得分的平均值为第 i 个样本的得分,即

$$s(i)=\frac{\mathrm{ns}(i)+\mathrm{fs}(i)}{2} \tag{6}$$

定义 4　每类的得分。定义 $\mathrm{cs}(j)$ 为第 j 类的得分,其定义为属于第 j 类的所有样本的得分的平均值,即

$$\mathrm{cs}(j)=\frac{\sum_{i=1}^{n_j}s(i)}{n_j} \tag{7}$$

式中 n_j 为第 j 类中的样本总数。

定义 5　NFS。定义 $\mathrm{nfs}(K)$ 为在类别数为 K 下聚类结果的最近最远得分,定义为所有类得分的平均值,即

$$\mathrm{nfs}(K)=\frac{1}{K}\sum_{j=1}^{K}\mathrm{cs}(j) \tag{8}$$

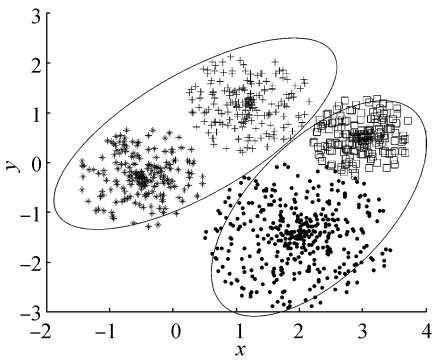
式中 K 是类别数。

2.2　NFS 指标的分析

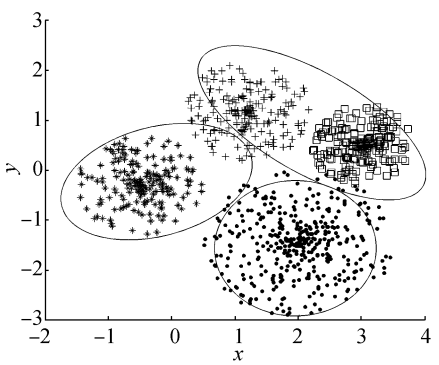
NFS 指标的设计原则是每个样本应该和距离其最近的样本划分到同一类别中,而与距离其最远的样本划分到不同类别中。因此,每个样本拥有两个影响评分的因子,分别是最近得分因子和最远得分因子。对于最近得分因子,如果某个样本与距离其最近的样本划分到同一类中,则得 1 分,表示对此聚类结果的支持,而如果划分到不同类中,则得-1 分,表示对此聚类结果的反对。最远得分因子的规定也是如此。

随着类别数的增加,样本和其最近邻样本被划分到不同类别中的概率将会增加,从而导致了最近得分累积和的减少;然而随着类别数的增加,样本和其最远邻样本被划分到同一类的概率将会减少,从而导致了最远得分累积和的增加。因此,在评价聚类结果时仅采用最近得分或最远得分将很难得到正确的类别数,需要综合利用这两个得分才能获得好的聚类结果。为了说明这个问题,我们将利用图 1 所示的数据集在不同类别数下的聚类结果来说明同时考虑最近得分和最远得分的必要性。观察图 1 所示数据集,可知此数据集的最佳类别数为 4。如果只考虑最近邻得分时,如图 1(a)所示,当 $K=2$ 时,所有样本和其最近邻样本都在同一类中,而如图 1(b)和图 1(c)所示,当 $K=3$ 或 $K=4$ 时,样本和其最近邻划分到不同类的比率就会增大,聚类结果的最近得分值会下降,因此 $K=2$ 时,使最近得分值达到最大,故采用最近得分准则所得到的最佳类别数为 2。而如果只考虑最远得分时,如图 1(a)所示,当 $K=2$ 时,样本和其最远邻样本被划分到同一类的比率很大,而如图 1(b)和图 1(c)所示,当 $K=3$ 或 $K=4$ 时,样本与其最远邻样本划分到同一类的比率下降,而且当 $K=3$ 时,样本和其最远邻样本分别划分到不同类中,此时聚类结果的最远得分值达到

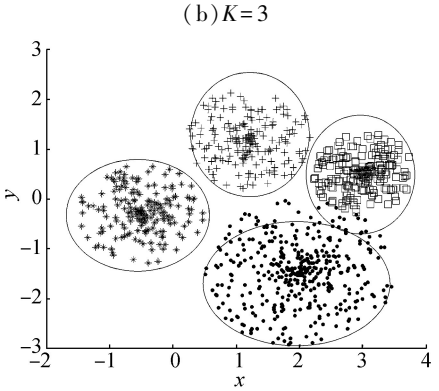
最大,因此,采用最远得分准则所得到的最佳类别数为 3。为了选择出正确的类别数,评价指标应该综合考虑最近得分和最远得分这两个因素。基于以上的理论,在 NFS 评价指标中每个样本的得分设计为最近得分和最远得分的均值。



(a) $K=2$



(b) $K=3$



(c) $K=4$

图 1 不同聚类数下的聚类

Fig.1 Clustering under different cluster number

NFS 指标值衡量的是样本对聚类结果的满意度,NFS 指标值越大聚类结果就越好。因此,依据 NFS 选取最佳类别数的公式为

$$K_{\text{opt}} = \max_{2\leqslant K\leqslant K_{\max}} \{\text{nfs}(K)\} \tag{9}$$

3 基于 NFS 的自动聚类算法

3.1 ACNFS 算法

基于 NFS 的自动聚类算法(automatic clustering algorithm based on the NFS,ACNFS)主要包括 3 个主要步骤。第 1 步确定类别数的搜索范围,第 2 步计算在不同类别数下的 NFS 评价指标值,第 3 步是根据评价指标得到使聚类结果达到最好的类别数。该算法具体步骤如下。

1)使用基于最短距离的分层聚类算法[19]并设定合适的截止阈值得到类别数,把该类别数作为搜索上限 $K_{\max}$,设置搜索下限 $K_{\min}=2$,得到聚类数的搜索范围 $[K_{\min},K_{\max}]$。

2)对于搜索范围 $[K_{\min},K_{\max}]$ 中不同的聚类数 K 分别运行以下步骤:

① 利用 K-means[20]算法对数据集进行聚类;

②根据式(4)~(8)计算聚类结果的 NFS 指标值;

③根据式(9)得到最佳聚类数 K_{opt}。

ACNFS 算法流程如图 2 所示。

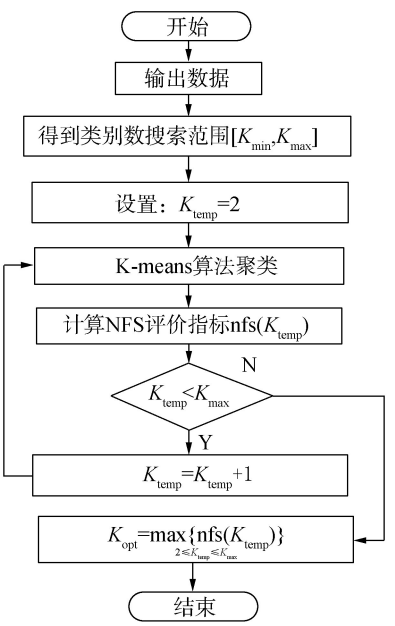


图 2 ACNFS 算法流程图

Fig.2 The flow chart of ACNFS algorithm

3.2 ACNFS 时间复杂度分析

假定数据集有 n 个样本,则 ACNFS 算法的时间复杂度分析如下。

1)ACNFS 算法首先采用基于最短距离的分层聚类算法,该算法每次把距离最近的两类合并成一类,其时间复杂度为 $O(tn^2)$,其中 t 为迭代次数且 $t\ll n$。

2)算法第二步中使用 K-means 算法进行聚类,其时间复杂度为 $O(nKl)$,其中 l 为迭代次数,K 为类别数,且 $l\ll n$,$K\ll n$。

3)采用 K-menas 聚类之后需要采用 NFS 评价指标对聚类结果进行评估。我们需要计算样本间的距离从而计算每个样本的最近得分与最远得分,因此计算 NFS 指标值的时间复杂度为 $O(n^2)$。

为了获得类别数,需要重复 K-means 和计算 NFS 指标值 $K_{\max}-K_{\min}+1$ 次,因此 2)和 3)的总的时间复杂度为

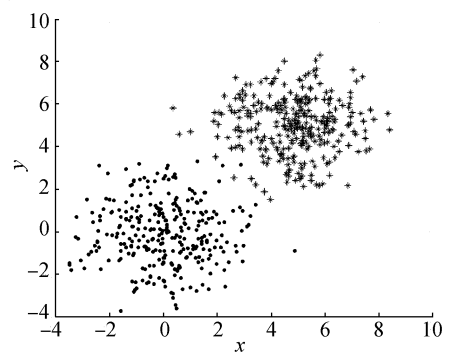
$O((n^2+nkl)\times(K_{max}-K_{min}+1))$,且$(K_{max}-K_{min}+1)\ll n$。

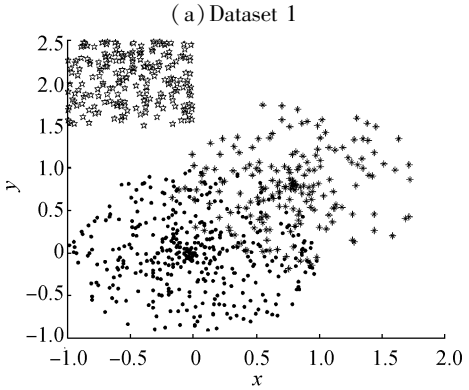
因此,ACNFS 算法总的时间复杂度即为$O(tn^2+(n^2+nkl)\times(K_{max}-K_{min}))$。

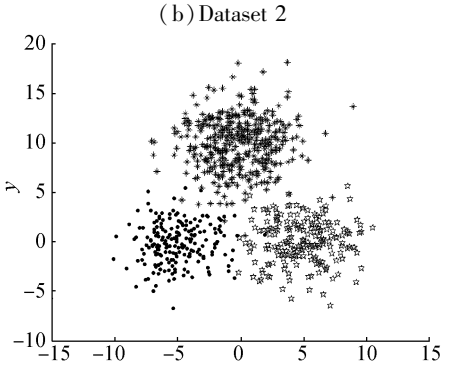
4　实验

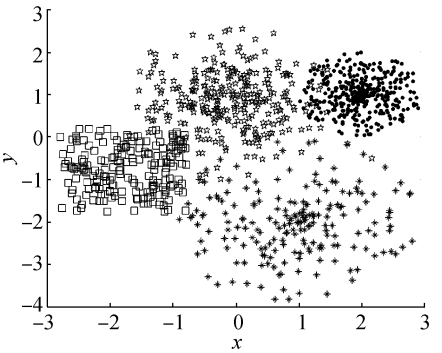
为了验证本文所提 NFS 指标和 ACNFS 算法的有效性,我们进行了仿真实验。在实验中我们将采用 6 个人工数据集和 4 个 UCI 真实数据,对 CH 指标、BWP 指标、IGP 指标和 NFS 指标确定类别数的性能进行比较。实验证明,基于 NFS 指标所提出的 ACNFS 算法可以有效地确定数据集的类别数,实现类别中心和类别数的自动估计。

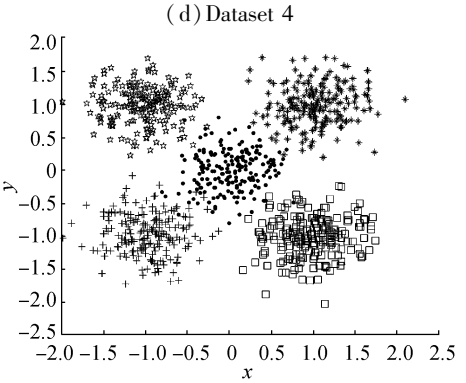
在仿真实验中,我们所采用的人工数据集如图 3 所示,而真实数据集则为 UCI 数据库中的 IRIS, Balance Scale, Wine 以及 Soybean-small 数据集。为了方便,我们将上述所有数据集的特征总结到表 1 中,其中 n 表示数据集中样本点的个数,K 表示数据集中样本点的类别数,d 表示数据集中样本点的特征维数,最后一列表示各类别中样本点的个数。


(a) Dataset 1


(b) Dataset 2


(c) Dataset 3


(d) Dataset 4


(e) Dataset 5

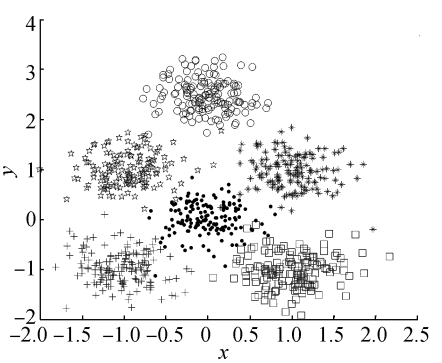

(f) Dataset 6

图 3　人工数据集

Fig.3　Artificial dataset

表 1　实验中所使用的 10 组数据集特征

Table 1　The characters of the ten data sets used in our experiments

数据集	n	K	d	每类中样本点的个数
Dataset 1	800	2	2	400,400
Dataset 2	800	3	2	400,200,200
Dataset 3	800	3	2	400,200,200
Dataset 4	1 100	4	2	200,400,200
Dataset 5	1 000	5	2	400,200,200,300
Dataset 6	900	6	2	200,200,200,200,200
Iris	150	3	4	150,150,150,150,150,150
Balance Scale	630	3	4	50,50,50
Wine	178	3	13	54,288,288
Soybean-small	47	4	35	59,71,48

首先,分别采用 CH 指标、BWP 指标、IGP 指标和 NFS 指标估计人工数据集 Dataset 1~Dataset 6 的类别数,实验结果如表 2~7 所示。

表 2 中给出了对数据集 Dataset 1 采用不同评价指标得到的类别数和各类别数出现的百分比。第 2 列到第 8 列的数据是百分数值,而百分数值代表算法在运行 N 次,得到相应的类别数的次数与 N 的比值。实验中取 $N=20$。表 2 中的最后 1 列表示通过投票准则获得的数据集的类别数。从表 2 可以看出,对人工数据集 Dataset 1 采用这 4 种评价指标均能得到正确的类别数,而且评价结果稳定,每次都能得到正确类别数。

表 2 Dataset 1 的类别数

Table 2 The cluster number of Dataset 1

评价指标	2	3	4	5	6	7	8	最终类别数
CH	100	0	0	0	0	0	0	2
BWP	100	0	0	0	0	0	0	2
IGP	100	0	0	0	0	0	0	2
NFS	100	0	0	0	0	0	0	2

表 3 给出了对数据集 Dataset 2 采用不同评价指标得到的类别数和各类别数出现的百分比。从 Dataset 2 的散点分布图中可以看出,有两类数据无法完全分离。因此,只有采用 NFS 指标得到的类别数是正确的类别数,而且评价结果稳定,每次都能得到正确的类别数;而采用其他的评价指标却无法得到正确的类别数。

表 3 Dataset 2 的类别数

Table 3 The cluster number of Dataset 2

评价指标	2	3	4	5	6	7	8	最终类别数
CH	100	0	0	0	0	0	0	2
BWP	100	0	0	0	0	0	0	2
IGP	100	0	0	0	0	0	0	2
NFS	0	100	0	0	0	0	0	3

表 4 给出了对数据集 Dataset 3 采用不同评价指标得到的类别数和各类别数出现的百分比。由于在数据集 Dataset 3 中,不同类之间是完全可分的,因此采用这 4 个评价指标,均可以得到正确的类别数,而且评价结果稳定。

表 4 Dataset 3 的类别数

Table 4 The cluster number of Dataset 3

评价指标	2	3	4	5	6	7	8	最终类别数
CH	0	100	0	0	0	0	0	3
BWP	0	100	0	0	0	0	0	3
IGP	0	100	0	0	0	0	0	3
NFS	0	100	0	0	0	0	0	3

表 5 给出了对数据集 Dataset 4 采用不同评价指标得到的类别数和各类别数出现的百分比。采用 NFS 指标、CH 指标和 BWP 指标均可以得到正确的类别数,而采用 IGP 指标却无法得到正确的类别数。从具体情况分析,NFS 指标和 CH 指标性能较好,评价结果稳定,每次都可以得到正确的类别数;而 BWP 指标差一点,正确率只有 71.3%。

表 5 Dataset 4 的类别数

Table 5 The cluster number of Dataset 4

评价指标	2	3	4	5	6	7	8	最终类别数
CH	0	0	100	0	0	0	0	4
BWP	0	0	71.3	20	8.7	0	0	4
IGP	95	5	0	0	0	0	0	2
NFS	0	0	100	0	0	0	0	4

表 6 给出了对数据集 Dataset 5 采用不同评价指标得到的类别数和各类别数出现的百分比。采用 NFS 指标、CH 指标和 BWP 指标均可以得到正确的类别数,而采用 IGP 指标却无法得到正确的类别数。

表 6 Dataset 5 的类别数

Table 6 The cluster number of Dataset 5

评价指标	2	3	4	5	6	7	8	最终类别数
CH	0	0	0	100	0	0	0	5
BWP	0	0	0	100	0	0	0	5
IGP	75	0	25	0	0	0	0	2
NFS	0	0	0	100	0	0	0	5

表 7 给出了对数据集 Dataset 6 采用不同评价指标得到的类别数和各类别数出现的百分比。采用四种评价指标均能得到该数据集的正确类别数。其中 NFS 指标、BWP 指标和 CH 指标的性能较好,评价结果稳定;而 IGP 指标性能差一点,正确率为 90%。

表 7 Dataset 6 的类别数

Table 7 The cluster number of Dataset 6

评价指标	2	3	4	5	6	7	8	最终类别数
CH	0	0	0	0	100	0	0	6
BWP	0	0	0	0	100	0	0	6
IGP	10	0	0	0	90	0	0	6
NFS	0	0	0	0	100	0	0	6

由表 2~7 所示的实验结果可如,对于可分性较好的人工数据集,CH 指标、BWP 指标和 NFS 指标均能获得正确的类别数。下面将采用 UCI 中的真实数据集 IRIS、Balance Scale、Wine 以及

Soybean-small来验证 CH 指标、BWP 指标、IGP 指标和 NFS 指标在确定类别数时的性能。

表 8 给出了数据集 Wine 在采用不同评价指标时,在不同的类别数下的指标值,其中带下划线的数据是该指标下的最大值。NFS 指标和 BWP 指标在类别数 $K=3$ 时取最大值,而其他指标在类别数 $K=2$ 时取最大值,但是由于数据集 Wine 的真实类别数为 3,因此采用 NFS 指标和 BWP 指标可以得到正确的类别数,而采用其他评价指标则无法得到正确的类别数。

表 8　Wine 的指标值

Table 8　The index value for Wine

类别数	CH	BWP	IGP	NFS
2	7 521 600	0.325 07	0.962 17	0.867 06
3	4 926 850	0.334 04	0.940 62	0.900 9
4	3 356 440	0.302 3	0.789 52	0.774 77
5	2 578 840	0.269 58	0.71	0.697 46
6	2 147 415	0.232 55	0.673 66	0.692 12
7	1 823 905	0.203 82	0.649 83	0.666 35
8	1 608 870	0.187 71	0.648 15	0.633 2
9	1 438 765	0.184 71	0.612 81	0.616 55
10	1 291 850	0.178 71	0.594 33	0.576 6
最终类别数	2	3	2	3

表 9 给出了 4 组真实数据集分别在采用不同评价指标下得到的类别数,这里依然是运行多次实验通过投票准则确定最终的类别数,括号中的数据表示类别数出现的百分比。参考表 1 中各数据集的真实类别数,可以得到如下结论:采用 NFS 指标可以得到所有真实数据集的正确的类别数,其中对于 Balance Scale 和 Wine 数据集,评价结果稳定,效果较好,而对于 IRIS 和 Soybean-small 数据集,评价结果差一点,只有 60% 和 45% 的正确率;然而采用 BWP 指标只可以得到数据集 Wine 的正确类别数,而且评价结果稳定;但是采用 CH 指标和 IGP 指标则无法得到数据集的正确类别数。

表 9 真实数据集的类别数

Table 9　The cluster number of the real datasets

数据集	CH	BWP	IGP	NFS
IRIS	2(100)	2(100)	2(100)	3(60)
Balance Scale	2(100)	8(70)	2(100)	3(100)
Wine	2(100)	3(100)	2(100)	3(100)
Soybean-small	2(100)	3(32.9)	3(40)	4(45)

5　结束语

众所周知,很多聚类算法需要根据先验知识给出算法所需要的类别数。但是,在很多实际应用中很难获得有效的先验知识,因此确定聚类问题的类别数成为了一个研究的热点。本文首先基于最近邻一致性和最远邻相异性的原则,提出了一种最近最远得分评价指标(NFS),并在此基础上提出了一种基于 NFS 自动聚类算法,实现了对类别数和类别中心的自动估计。与已经提出的评价指标相比,NFS 指标是基于数据集统计信息的指标,而且 NFS 指标考虑了最近样本和最远样本两个方面,通过评分机制还保证了每个样本都对评价指标产生影响。从而使 NFS 指标在 IRIS 等数据集中呈现较好的结果。但是 NFS 指标并不是最完美的,因此还需要继续进行相关研究。

参考文献:

[1]刘恋, 常冬霞, 邓勇. 动态小生境人工鱼群算法的图像分割[J]. 智能系统学报, 2015, 10(5): 669-674.
LIU Lian, CHANG Dongxia, DENG Yong. An image segmentation method based on dynamic niche artificial fish-swarm algorithm [J]. CAAI transactions on intelligent systems, 2015, 10(5): 669-674.

[2]NIKOLAOU T G, KOLOKOTSA D S, STAVRAKAKIS G S, et al. On the application of clustering techniques for office buildings' energy and thermal comfort classification [J]. IEEE transactions on smart grid, 2012, 3(4): 2196-2210.

[3] CHANG Hong, YEUNG D Y. Robust path-based spectral clustering with application to image segmentation [C]// Proceedings of the Tenth IEEE International Conference on Computer Vision. Beijing, China, 2005, 1: 278-285.

[4] SHI Jianbo, MALIK J. Normalized cuts and image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2000, 22(8): 888-905.

[5]XIE X L, BENI G. A validity measure for Fuzzy clustering [J]. IEEE transactions on pattern analysis and machine intelligence, 1991, 13(8): 841-847.

[6]PAL N R, BEZDEK J C. On cluster validity for the fuzzy c-means model [J]. IEEE transactions on fuzzy systems, 1995, 3(3): 370-379.

[7]郑宏亮, 徐本强, 赵晓慧, 等. 新的模糊聚类有效性指标[J]. 计算机应用, 2014, 34(8): 2166-2169.
ZHENG Hongliang, XU Benqiang, ZHAO Xiaohui, et al. Novel validity index for fuzzy clustering [J]. Journal of computer applications, 2014, 34(8): 2166-2169.

[8]岳士弘, 黄媞, 王鹏龙. 基于矩阵特征值分析的模糊聚

类有效性指标[J]. 天津大学学报: 自然科学与工程技术版, 2014, 47(8): 689-696.
YUE Shihong, HUANG Ti, WANG Penglong. Matrix eigenvalue analysis - based clustering validity index [J]. Journal of Tianjin university: science and technology, 2014, 47(8): 689-696.

[9]卿铭, 孙晓梅. 一种新的聚类有效性函数: 模糊划分的模糊熵[J]. 智能系统学报, 2015, 10(1): 75-80.
QING Mei, SUN Xiaomei. A new clustering effectiveness function: fuzzy entropy of fuzzy partition [J]. CAAI transactions on intelligent systems, 2015, 10(1): 75-80.

[10]王开军, 李健, 张军英, 等. 聚类分析中类数估计方法的实验比较[J]. 计算机工程, 2008, 34(9): 198-199, 202.
WANG Kaijun, LI Jian, ZHANG Junying, et al. Experimental comparison of clusters number estimation for cluster analysis[J]. Computer engineering, 2008, 34(9): 198-199, 202.

[11]王勇, 唐靖, 饶勤菲, 等. 高效率的 K-means 最佳聚类数确定算法[J]. 计算机应用, 2014, 34(5): 1331-1335.
WANG Yong, TANG Jing, RAO Qinfei, et al. High efficient K-means algorithm for determining optimal number of clusters[J]. Journal of computer applications, 2014, 34(5): 1331-1335.

[12]CALIŃSKI T, HARABASZ J. A dendrite method for cluster analysis [J]. Communications in statistics, 1974, 3(1): 1-27.

[13] DAVIES D L, BOULDIN D W. A cluster separation measure[J]. IEEE transactions on pattern analysis and machine intelligence, 1979, PAMI-1(2): 224-227.

[14]DIMITRIADOU E, DOLNIČAR S, WEINGESSEL A. An examination of indexes for determining the number of clusters in binary data sets[J]. Psychometrika, 2002, 67(1): 137-159.

[15]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.

[16]周世兵, 徐振源, 唐旭清. K-means 算法最佳聚类数确定方法[J]. 计算机应用, 2010, 30(8): 1995-1998.
ZHOU Shibing, XU Zhenyuan, TANG Xuqing. Method for determining optimal number of clusters in K-means clustering algorithm[J]. Journal of computer applications, 2010, 30(8): 1995-1998.

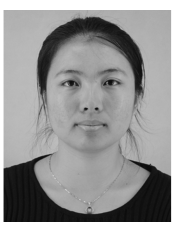
[17] KAPP A V, TIBSHIRANI R. Are clusters found in one dataset present in another dataset[J]. Biostatistics, 2007, 8(1): 9-31.

[18]周世兵. 聚类分析中的最佳聚类数确定方法研究及应用[D]. 无锡: 江南大学, 2011.
ZHOU Shibing. Research and application on determining optimal number of cluster in cluster analysis[D]. Wuxi: Jiangnan University, 2011.

[19] Gower J C, Ross G J S. Minimum spanning trees and single linkage cluster analysis[J]. Journal of the royal statistical society, 1969, 18(1):54-64.

[20] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, USA, 1967: 281-297.

作者简介:



冯柳伟, 女, 1992 年生, 硕士研究生, 研究方向为聚类算法。

常冬霞, 女, 1977 年生, 副教授, 硕士生导师, 主要研究方向为进化计算、非监督分类算法、图像分割以及图像分类。发表学术论文 10 余篇, 其中 SCI 检索 5 篇, EI 检索 2 篇。

邓勇, 男, 1974 年生, 副研究员, 博士, 主要研究方向为智能信息处理、数据库系统技术及应用等。主持和参与国家“863”计划 1 项, 北京市自然科学基金项目 1 项。发表学术论文 20 余篇, 其中收录 10 余篇。