文章编号: 1001 - 9081(2010) 08 - 1995 - 04

K-means 算法最佳聚类数确定方法

周世兵[1],徐振源[1,2],唐旭清[2]
(1. 江南大学 信息工程学院,江苏 无锡 214122; 2. 江南大学 理学院,江苏 无锡 214122)
(worldguard@ sina. com)

摘 要: K-means 聚类算法是以确定的类数 k 为前提对数据集进行聚类的,通常聚类数事先无法确定。从样本几何结构的角度设计了一种新的聚类有效性指标,在此基础上提出了一种新的确定 K-means 算法最佳聚类数的方法。理论研究和实验结果验证了以上算法方案的有效性和良好性能。

关键词: K-means 聚类; 聚类数; 聚类有效性指标; 聚类分析
中图分类号: TP18 **文献标志码:** A

Method for determining optimal number of clusters in K-means clustering algorithm

ZHOU Shi-bing[1], XU Zhen-yuan[1,2], TANG Xu-qing[2]
(1. *College of Information Technology, Jiangnan University, Wuxi Jiangsu* 214122, *China*;
2. *College of Science, Jiangnan University, Wuxi Jiangsu* 214122, *China*)

Abstract: K-means clustering algorithm clusters datasets according to the certain clustering number k. However, k cannot be confirmed beforehand. A new clustering validity index was designed from the standpoint of sample geometry. Based on the index, a new method for determining the optimal clustering number in K-means clustering algorithm was proposed. Theoretical research and experimental results demonstrate the validity and good performance of the above-mentioned algorithm.

Key words: K-means clustering; number of clusters; clustering validity index; cluster analysis

0 引言

K-means 聚类算法是聚类分析中使用最为广泛的算法之一。该算法对大型数据集的处理效率较高,特别是当样本分布呈现类内团聚状时,可以达到很好的聚类结果。K-means 算法是以确定的类数 k 和选定的初始聚类中心为前提,使各样本到其判属类别中心距离(平方) 之和最小的最佳聚类。在实际中,k 值是难以准确界定的,目前已经提出了一些检验聚类有效性的函数指标,主要有 Calinski-Harabasz(CH) 指标[1]、Davies-Bouldin(DB) 指标[1]、Krzanowski-Lai(KL) 指标[1]、Weighted inter-intra(Wint) 指标[1]、In-Group Proportion(IGP) 指标[1]等。人们使用这些聚类有效性指标计算合适的聚类数 k,即最佳聚类数 k_{opt}。但是,由于这些有效性指标自身的缺陷,对于聚类结构难以判别的情况,它们的聚类有效性检验效果不够理想,很难得到正确的最佳聚类数。针对这种情况,本文基于样本的几何结构,设计了一种新的有效性指标,在此基础上,提出了一种确定样本最佳聚类数的方法,用来评估 K-means 算法的聚类结果和确定样本的最佳聚类数。理论研究和实验结果表明,与其他指标和方法相比,本文提出的新指标和方法具有更好的性能和可行性。

1 K-means 聚类算法

1.1 K-means 算法介绍

该算法取定 k 类和选取 k 个初始聚类中心,按最小距离原则将各样本分配到 k 类中的某一类,之后不断地计算类心和调整各样本的类别,最终使各样本到其判属类别中心的距离平方之和最小。算法步骤[6-7] 如下:

1) 针对 n 个样本,任选 k 个样本作为初始聚类中心(z_1, z_2, …, z_k);

2) 对每个样本 x_i 找到离它最近的聚类中心 z_v,并将其分配到 z_v 所标明的类 u_v;

3) 采取平均的方法计算重新分类后的各类心;

4) 计算 $D = \sum_{i=1}^{n} \left[\min_{r=1,\cdots,k} d(x_i, z_r)^2\right]$;

5) 如果 D 值收敛,则 return(z_1, z_2, …, z_k, U) 并终止本算法,否则转至 2)。

1.2 K-means 算法优缺点

对于大数据集,K-means 算法是相对可伸缩的和高效率的,因为它的时间复杂度是 $O(nkt)$,其中 n 是样本数,k 是聚类数,t 是迭代次数。通常 $k \ll t$ 且 $t \ll n$。用 K-means 算法聚类时,对于类内紧密、类间远离的聚类结构,它的聚类效果较好。该算法的缺点是必须事先给定聚类数 k,不准确的 k 值会导致聚类质量下降。另外对于比较复杂的聚类结构,聚类结果易受初始聚类中心影响,导致聚类结果不稳定。

2 新聚类有效性指标

评价聚类结果优劣的过程称为聚类有效性分析。一般来

收稿日期: 2010 - 02 - 23; 修回日期: 2010 - 03 - 21。
基金项目: 国家 863 计划项目(2007AA1Z158) ; 国家自然科学基金资助项目(60703106) 。
作者简介: 周世兵(1972 -) ,男,江苏盐城人,讲师,博士研究生,主要研究方向: 人工智能、模式识别、生物信息学; 徐振源(1946 -) ,男,上海人,教授,博士生导师,主要研究方向: 混沌、同步控制、人工智能、生物信息学; 唐旭清(1963 -) ,男,安徽望江人,副教授,博士,主要研究方向: 计算智能、生物信息学。

说,一个好的聚类划分应尽可能反映数据集的内在结构,使类内样本尽可能相似,类间样本尽可能不相似。从距离测度考虑,就是使类内距离极小化而类间距离最大化的聚类是最优聚类。目前已提出了一些聚类有效性指标,由于这些指标自身的缺陷,一般难以找到正确的最佳聚类数。鉴于这种情况,本文设计了一种新的聚类有效性指标,该指标可以对 *K*-means 算法的聚类结果进行评估并可用来确定最佳聚类数。

2.1 新指标及相关概念定义

定义 1 令 $K=\{X,R\}$ 为聚类空间,其中 $X=\{x_1,x_2,\cdots,x_n\}$,假设 n 个样本对象被聚类为 c 类,定义第 j 类的第 i 个样本的最小类间距离 $b(j,i)$ 为该样本到其他每个类中样本平均距离的最小值,即:

$$b(j,i)=\min_{1\le k\le c,k\ne j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\|x_p^{(k)}-x_i^{(j)}\|^2\right) \tag{1}$$

其中:k 和 j 表示类标,$x_i^{(j)}$ 表示第 j 类的第 i 个样本,$x_p^{(k)}$ 表示第 k 类的第 p 个样本,n_k 表示第 k 类中的样本个数,$\|\cdot\|^2$ 表示平方欧氏距离。

定义 2 令 $K=\{X,R\}$ 为聚类空间,其中 $X=\{x_1,x_2,\cdots,x_n\}$,假设 n 个样本对象被聚类为 c 类,定义第 j 类的第 i 个样本的类内距离 $w(j,i)$ 为该样本到第 j 类中其他所有样本的平均距离,即:

$$w(j,i)=\frac{1}{n_j-1}\sum_{q=1,q\ne i}^{n_j}\|x_q^{(j)}-x_i^{(j)}\|^2 \tag{2}$$

其中:$x_q^{(j)}$ 表示第 j 类中的第 q 个样本,并且 $q\ne i$,n_j 表示第 j 类中的样本个数。实际使用中,无需保证 $q\ne i$,因为 $q=i$ 时,欧氏距离为 0,并不影响算法的正确性。

定义 3 令 $K=\{X,R\}$ 为聚类空间,其中 $X=\{x_1,x_2,\cdots,x_n\}$,假设 n 个样本对象被聚类为 c 类,定义第 j 类的第 i 个样本的聚类距离 $baw(j,i)$ 为该样本的最小类间距离和类内距离之和,即:

$$baw(j,i)=b(j,i)+w(j,i)=\min_{1\le k\le c,k\ne j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\|x_p^{(k)}-x_i^{(j)}\|^2\right)+\frac{1}{n_j-1}\sum_{q=1,q\ne i}^{n_j}\|x_q^{(j)}-x_i^{(j)}\|^2 \tag{3}$$

定义 4 令 $K=\{X,R\}$ 为聚类空间,其中 $X=\{x_1,x_2,\cdots,x_n\}$,假设 n 个样本对象被聚类为 c 类,定义第 j 类的第 i 个样本的聚类离差距离 $bsw(j,i)$ 为该样本的最小类间距离和类内距离之差,即:

$$bsw(j,i)=b(j,i)-w(j,i)=\min_{1\le k\le c,k\ne j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\|x_p^{(k)}-x_i^{(j)}\|^2\right)-\frac{1}{n_j-1}\sum_{q=1,q\ne i}^{n_j}\|x_q^{(j)}-x_i^{(j)}\|^2 \tag{4}$$

定义 5 令 $K=\{X,R\}$ 为聚类空间,其中 $X=\{x_1,x_2,\cdots,x_n\}$,假设 n 个样本对象被聚类为 c 类,定义第 j 类的第 i 个样本的类间类内划分(Between-Within Proportion,BWP)指标 $BWP(j,i)$ 为该样本的聚类离差距离和聚类距离的比值,即:

$$BWP(j,i)=\frac{bsw(j,i)}{baw(j,i)}=\frac{b(j,i)-w(j,i)}{b(j,i)+w(j,i)}=\frac{\min_{1\le k\le c,k\ne j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\|x_p^{(k)}-x_i^{(j)}\|^2\right)-\frac{1}{n_j-1}\sum_{q=1,q\ne i}^{n_j}\|x_q^{(j)}-x_i^{(j)}\|^2}{\min_{1\le k\le c,k\ne j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\|x_p^{(k)}-x_i^{(j)}\|^2\right)+\frac{1}{n_j-1}\sum_{q=1,q\ne i}^{n_j}\|x_q^{(j)}-x_i^{(j)}\|^2} \tag{5}$$

2.2 新指标分析

为了反映了聚类结构的类内紧密性和类间分离性,我们提出了 BWP 指标。BWP 指标基于样本的几何结构,以数据集中的某个样本作为研究对象,对聚类结果进行有效性分析。为便于说明该指标及相关概念的意义,我们借鉴了文献[8]的指标说明方法,结合图 1 的聚类结构分布示意图进行说明。在图 1 中,数据集中的所有样本被分为 4 类,分别是 j,x,y,z,在第 j 类中有一个样本 i。在样本 i 的类内结构方面,根据定义 2,样本 i 到类 j 中所有样本距离的平均值,称为样本 i 的类内距离。相比于把样本 i 到类 j 的中心之间的距离称为类内距离来说,定义 2 更准确,更能反映样本 i 和类 j 中其他样本的结构关系。在样本 i 的类间结构方面,为了反映类间的分离性,我们研究样本 i 的近邻聚类与样本 i 的关系。近邻聚类可以通过样本 i 的最小类间距离所对应的聚类得到,在图 1 中类 x 就是样本 i 的近邻聚类。因为如果类 x 满足类间远离的要求,那么其他聚类也一定满足要求。另外,样本 i 如果没有聚类到类 j,那么类 x 就是它的最佳选择。因此,研究样本 i 所在的聚类 j 和样本 i 的近邻聚类 x 具有重要意义。

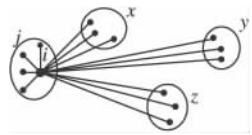

图 1 聚类结构分布示意图

从某一样本所属的聚类和它的近邻聚类的关系出发,我们提出了 BWP 指标。确定聚类有效性的标准是使聚类结果达到类内紧密、类间远离。从类内紧密的角度出发,我们希望样本的类内距离 $w(j,i)$ 越小越好;从类间远离的角度出发,我们希望样本离近邻聚类的距离,也就是最小类间距离 $b(j,i)$ 越大越好。为了综合这两种因素,我们使用线性组合方式平衡二者,并使函数的目标一致。使用 $b(j,i)+(-w(j,i))$,也就是聚类离差距离 $bsw(j,i)$ 来评价聚类结果,显然 $bsw(j,i)$ 越大,说明该样本聚类效果好。为了使指标能够对所有样本进行有效性分析,并使指标不受量纲影响,我们引入了样本聚类距离的概念。通过样本聚类距离对单个样本的聚类离差距离进行压缩,使指标成为无量纲,指标的值为样本单位聚类距离上的离差距离,指标值的范围为 $[-1,1]$。为了扩大指标的适用范围,使指标能够处理类间距较小的数据集,我们采用平方欧氏距离作为距离测度。

为了更好地使用 BWP 指标,我们分析几种特例情况。当样本的类内距离和样本的最小类间距离相比可以忽略时,BWP 指标的值近似为 1,说明此时该样本被正确聚类。当样本的最小类间距离和样本的类内距离相比可以忽略时,BWP 指标的值近似为 -1,说明此时该样本被错误聚类。由于最小类间距离需要聚类数至少为两类,因此 BWP 指标不适用于聚类数为 1 的情况。

2.3 新指标与最佳聚类数确定

BWP 指标反映了单个样本的聚类有效性情况,BWP 指标值越大,说明单个样本的聚类效果越好。我们通过求某个数

据集中所有样本的 BWP 指标值的平均值，来分析该数据集的聚类效果。显然，平均值越大，说明该数据集的聚类效果越好，其最大值所对应的聚类数是最佳聚类数。由此我们得到如下公式，其中 $avg_{BWP}(k)$ 表示数据集聚成 k 类时的平均 BWP 指标值，k_{opt} 表示最佳聚类数。

$$avg_{BWP}(k) = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} BWP(j,i) \quad (6)$$

$$k_{opt} = \underset{2\leqslant k<n}{\operatorname{argmax}}\{avg_{BWP}(k)\} \quad (7)$$

3　确定最佳聚类数的算法

本文结合 K-means 算法以及式(5)定义的 BWP 聚类有效性指标，提出一种新的分析聚类效果，确定最佳聚类数的算法。算法归纳如下。

1) 选择聚类数的搜索范围 $[k_{min}, k_{max}]$。

2) 从 k_{min} 循环至 k_{max}:

① 调用 K-means 算法;

② 利用式(5)计算单个样本的 BWP 指标值;

③ 利用式(6)计算平均 BWP 指标值。

3) 利用式(7)计算最佳聚类数。

4) 输出最佳聚类数、有效性指标值和聚类结果。

4　实验与分析

为了检验有效性指标 BWP 和最佳聚类数确定算法的性能，本文通过两组实验共 6 个数据集进行测试，并与常用指标 CH 指标、DB 指标、KL 指标、Wint 指标以及 IGP 指标等进行比较。实验中聚类数的搜索范围为 $[2, k_{max}]$，根据普遍使用的经验规则 $k_{max}\leqslant\sqrt{n}$，取 $k_{max}=\mathrm{Int}(\sqrt{n})$。为避免初始聚类中心对 K-means 算法的聚类结果产生影响，我们使算法分别运行 50 次，分析聚类效果和最佳聚类数产生情况。

实验 1　人工数据集实验。该实验包括 3 个人工数据集，分别是 SM1、SM2 和 Y3c。SM1 数据集由中心分别为(0,0)，(30,30)的二维两高斯分布数据组成，其中(0,0)类有 100 个样本，(30,30)类有 300 个样本，它们的协方差矩阵分别为 $\boldsymbol{I}_2$ 和 $50\boldsymbol{I}_2$，$\boldsymbol{I}_2$ 为 2 阶单位矩阵，该数据集的结构特征为两个聚类的类密度不同。SM2 数据集由中心分别为(0,0)，(5,5)，(10,10)，(15,15)的二维四高斯分布数据组成，每个类有 600 个样本，每个类的协方差矩阵为 $2\boldsymbol{I}_2$，该数据集的结构特征为某些类的类间距离很近并且存在大量重叠的聚类结构。Y3c 数据集[1]是二维三类的人工合成数据集，其结构特征为轻微重叠、松散的聚类结构。

对 SM1 数据集采用不同有效性评价指标估计出的最佳聚类数实验情况如表 1 所示。

表 1　几种有效性指标估计出的 SM1 数据集最佳聚类数

指标	最佳聚类数/%					最终聚类数
	2	3	4	5	其他	
CH	100	0	0	0	0	2
DB	100	0	0	0	0	2
KL	78	0	8	2	12	2
Wint	44	34	8	8	6	2,3
IGP	100	0	0	0	0	2
BWP	100	0	0	0	0	2

表 1 中百分数值代表算法运行 w 次，得到相应最佳聚类数的次数与 w 的比值。本文实验取 $w=50$，其中带下划线的数值表示得到的最终聚类数是正确的最佳聚类数。从总体情况看，除了 Wint 指标，其他五种有效性评价指标都得到了正确的最佳聚类数。从具体情况看，CH 指标、DB 指标、IGP 指标和 BWP 指标性能较好，评估结果稳定，每次都能得到正确的最佳聚类数; KL 指标稍差，正确率为 78%; Wint 指标评估结果不稳定，每个最佳聚类数都未超过 50%。SM1 数据集的结构分布和聚类结果如图 2 所示。

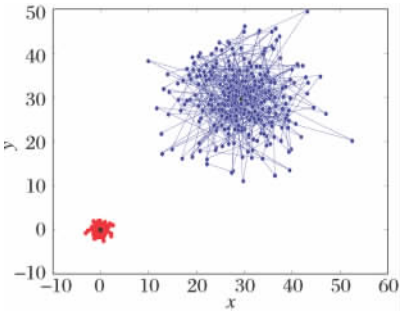

图 2　$k=2$ 时数据集 SM1 的聚类结果

对 SM2 数据集采用不同有效性评价指标估计出的最佳聚类数实验情况如表 2 所示。从总体情况看，CH 指标、DB 指标和 BWP 指标都得到了正确的最佳聚类数，其他指标不能得到正确的最佳聚类数。从具体情况看，CH 指标和 BWP 指标性能较好，评估结果稳定，每次都能得到正确的最佳聚类数，DB 指标稍差，正确率为 86%。SM2 数据集的结构分布和聚类结果如图 3 所示。

表 2　几种有效性指标估计出的 SM2 数据集最佳聚类数

指标	最佳聚类数/%					最终聚类数
	2	3	4	5	其他	
CH	0	100	0	0	0	4
DB	0	14	86	0	0	4
KL	4	0	0	12	84	不确定
Wint	0	56	32	6	6	3
IGP	58	42	0	0	0	2,3
BWP	0	0	100	0	0	4

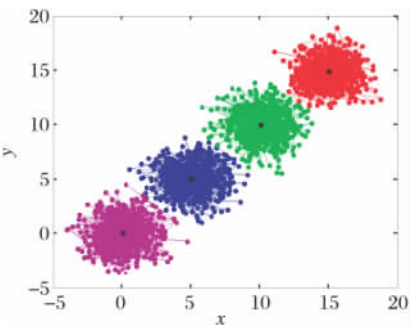

图 3　$k=4$ 时数据集 SM2 的聚类结果

对 Y3c 数据集采用不同有效性评价指标估计出的最佳聚类数实验情况如表 3 所示。从总体情况看，CH 指标、DB 指标、Wint 指标和 BWP 指标都得到了正确的最佳聚类数，其他指标不能得到正确的最佳聚类数。从具体情况看，CH 指标和 BWP 指标性能较好，评估结果稳定，每次都能得到正确的最佳聚类数; DB 指标和 Wint 指标稍差，正确率分别为 72% 和 66%。

实验 2　UCI 真实数据集实验。该实验包括 3 个 UCI 真实数据集，分别是 BUPA、Pima-indians-diabetes(本文简称 Pid)、Breast-cancer-wisconsin(本文简称 Bcw)，来源于 UCI

Machine Learning Repository (http: //www. ics. uci. edu/ ~ mlearn/MLRepository. html) 。基于 *K*-means 算法,对 BUPA 数据集采用几种有效性指标得到的聚类评估结果如表 4 所示。其中带下划线的指标值所对应的聚类数为该列指标得到的最佳聚类数。由于 BUPA 数据集的真实类数为 2 类,所以 CH 指标、DB 指标、IGP 指标和 BWP 指标得到的最佳聚类数是正确的,而 KL 指标和 Wint 指标得到的最佳聚类数是错误的。

表 3 几种有效性指标估计出的 Y3c 数据集最佳聚类数

指标	最佳聚类数/%					最终聚类数
	2	3	4	5	其他	
CH	0	100	0	0	0	3
DB	0	72	0	0	28	3
KL	0	6	10	8	76	不确定
Wint	0	66	16	12	6	3
IGP	74	0	26	0	0	2
BWP	0	100	0	0	0	3

表 4 BUPA 数据集的聚类有效性指标值

聚类数	CH	DB	KL	Wint	IGP	BWP
2	322.2691	0.7679	4.1907	0.6105	0.9849	0.7442
3	264.7511	0.9176	1.3429	0.8210	0.9151	0.5647
4	244.5438	0.8694	1.9649	0.6646	0.8623	0.3527
5	222.9263	0.9540	2.9106	0.5742	0.8535	0.3706
6	199.3688	0.9397	0.4004	0.5857	0.9193	0.2345
7	191.9995	1.0160	1.9707	0.5467	0.8454	0.2448
8	181.0314	0.9770	5.3781	0.5524	0.8212	0.2426
9	167.2158	1.1174	0.0333	0.5234	0.8103	0.2275
10	198.3169	0.9332	187.6784	0.5367	0.8510	0.2558
11	184.5119	0.9794	0.0467	0.5213	0.8640	0.2147
12	178.4358	0.9634	2.0807	0.5446	0.8427	0.2286
13	170.7596	1.0044	0.3405	0.5072	0.8377	0.2317
14	169.0864	0.9911	4.8775	0.5244	0.8315	0.2381
15	162.2242	0.9319	0.7533	0.5136	0.7980	0.2420
16	152.8220	1.1007	0.4163	0.5178	0.8112	0.2045
17	150.5617	1.0267	0.8854	0.5036	0.8215	0.2245
18	149.3795	1.0275	0.8854	0.4996	0.8373	0.2005

对真实类数为 2 类的 Pid 数据集,运用 CH 指标、DB 指标和 BWP 指标确定最佳聚类数的实验情况分别如图 4 ~ 图 6 所示。从中可以看出 BWP 指标得到的最佳聚类数 2 是正确的,而 CH 指标得到的最佳聚类数 3 和 DB 指标得到的最佳聚类数 4 是错误的。

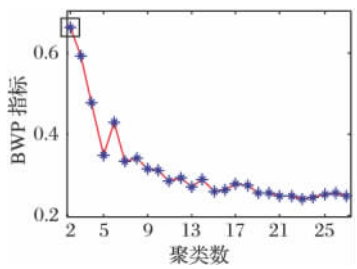

图 4 Pid 的聚类数—BWP 指标关系图

真实数据集信息以及几种有效性评价指标估计出的最佳聚类数实验结果如表 5 所示。从表 5 可知,BWP 指标和 IGP 指标对 3 个真实数据集都能够得到正确的最佳聚类数,CH 指标和 DB 指标对 BUPA 和 Bcw 数据集能够得到正确的最佳聚类数,KL 指标和 Wint 指标对每个真实数据集都无法得到正确的最佳聚类数。

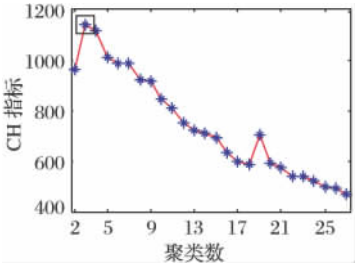

图 5 Pid 的类数—CH 指标关系图

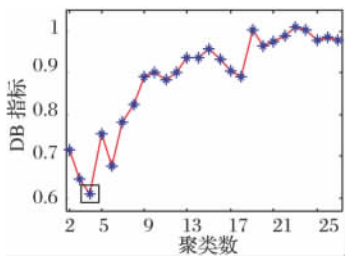

图 6 Pid 的类数—DB 指标关系图

表 5 几种有效性评价指标估计出的 UCI 数据集最佳聚类数

数据集	样本数目	样本维数	真实类数	最佳聚类数					
				CH	DB	KL	Wint	IGP	BWP
BUPA	345	6	2	2	2	10	3	2	2
Pid	768	8	2	3	4	7	3	2	2
Bcw	699	9	2	2	2	3	3	2	2

5 结语

K-means 聚类算法需要用户根据先验知识提供聚类数 k,但多数情况下,聚类数 k 事先无法确定。本文针对 *K*-means 算法的有效性问题,从样本几何结构的角度定义了样本聚类距离和样本聚类离差距离,设计了一种新的聚类有效性指标——BWP 指标。在此基础上提出了一种新的确定 *K*-means 算法最佳聚类数的方法。理论研究和实验结果验证了以上算法方案的有效性和良好性能。

参考文献:

[1] CALINSKI R, HARABASZ J. A dendrite method for cluster analysis [J]. Communications in Statistics, 1974, 3(1): 1 – 27.

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

[3] DUDOIT S, FRIDLYAND J. A prediction-based resampling method for estimating the number of clusters in a dataset[J]. Genome Biology, 2002, 3(7): 1 – 21.

[4] DIMITRIADOU E, DOLNICAR S, WEINGESSEL A. An examination of indexes for determining the number of cluster in binary data sets[J]. Psychometrika, 2002, 67(1): 137 – 160.

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

[6] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48 – 61.

[7] 孙即祥. 现代模式识别[M]. 长沙: 国防科技大学出版社, 2002.

[8] ROUSSEEUW P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics, 1987, 20(1): 53 – 65.

[9] DEMBÉLÉ D, KASTNER P. Fuzzy C-means method for clustering microarray data[J]. Bioinformatics, 2003, 19(8): 973 – 980.