ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.1, January 2008, pp.62−72
DOI: 10.3724/SP.J.1001.2008.00062

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于层次划分的最佳聚类数确定方法*

陈黎飞[1], 姜青山[2+], 王声瑞[3]

[1](厦门大学 计算机科学系,福建 厦门 361005)

[2](厦门大学 软件学院,福建 厦门 361005)

[3](Department of Computer Science, University of Sherbooke, J1K 2R1, Canada)

A Hierarchical Method for Determining the Number of Clusters

CHEN Li-Fei[1], JIANG Qing-Shan[2+], WANG Sheng-Rui[3]

[1](Department of Computer Science, Xiamen University, Xiamen 361005, China)

[2](School of Software, Xiamen University, Xiamen 361005, China)

[3](Department of Computer Science, University of Sherbooke, J1K 2R1, Canada)

+ Corresponding author: Phn: +86-592-2186707, E-mail: qjiang@xmu.edu.cn, http://software.xmu.edu.cn/View/shizi/jqs.htm

Chen LF, Jiang QS, Wang SR. A hierarchical method for determining the number of clusters. *Journal of Software*, 2008,19(1):62−72. http://www.jos.org.cn/1000-9825/19/62.htm

Abstract: A fundamental and difficult problem in cluster analysis is the determination of the "true" number of clusters in a dataset. The common trail-and-error method generally depends on certain clustering algorithms and is inefficient when processing large datasets. In this paper, a hierarchical method is proposed to get rid of repeatedly clustering on large datasets. The method firstly obtains the CF (clustering feature) via scanning the dataset and agglomerative generates the hierarchical partitions of dataset, then a curve of the clustering quality w.r.t the varying partitions is incrementally constructed. The partitions corresponding to the extremum of the curve is used to estimate the number of clusters finally. A new validity index is also presented to quantify the clustering quality, which is independent of clustering algorithm and emphasis on the geometric features of clusters, handling efficiently the noisy data and arbitrary shaped clusters. Experimental results on both real world and synthesis datasets demonstrate that the new method outperforms the recently published approaches, while the efficiency is significantly improved.

Key words: clustering; clustering validity index; statistics; number of cluster; hierarchically clustering

摘 要: 确定数据集的聚类数目是聚类分析中一项基础性的难题.常用的 trail-and-error 方法通常依赖于特定的聚类算法,且在大型数据集上计算效率欠佳.提出一种基于层次思想的计算方法,不需要对数据集进行反复聚类,它首先扫描数据集获得 CF(clustering feature,聚类特征)统计值,然后自底向上地生成不同层次的数据集划分,增量地构建一条关于不同层次划分的聚类质量曲线;曲线极值点所对应的划分用于估计最佳的聚类数目.另外,还提出一种新的

* Supported by the National Natural Science Foundation of China under Grant No.10771176 (国家自然科学基金); the National 985 Project of China under Grant No.0000-X07204 (985 工程二期平台基金); the Scientific Research Foundation of Xiamen University of China under Grant No.0630-X01117 (厦门大学科研基金)

Received 2007-04-01; Accepted 2007-10-09

聚类有效性指标用于衡量不同划分的聚类质量.该指标着重于簇的几何结构且独立于具体的聚类算法,能够识别噪声和复杂形状的簇.在实际数据和合成数据上的实验结果表明,新方法的性能优于新近提出的其他指标,同时大幅度提高了计算效率.

关键词: 聚类;聚类有效性指标;统计指标;聚类数;层次聚类

中图法分类号: TP18 **文献标识码:** A

聚类是数据挖掘研究中重要的分析手段.迄今,研究者已提出了多种聚类算法[1],在商务智能、Web 挖掘等领域中得到了广泛的应用.然而,许多聚类算法需要用户给定聚类数,在实际应用中,这通常需要用户根据经验或具备相关领域的背景知识.确定数据集的聚类数问题目前仍是聚类分析研究中的一项基础性难题[2−4].

现有的研究[2−13]是通过以下过程(一种迭代的 trial-and-error 过程[9])来确定数据集最佳聚类数的,如图 1 所示.在给定的数据集或通过随机抽样得到的数据子集上,使用不同的参数(通常是聚类数 k)运行特定的聚类算法对数据集进行不同的划分,计算每种划分的统计指标值,最后比较分析各个指标值的大小或变化情况,符合预定条件的指标值所对应的算法参数 k 被认为是最佳的聚类数 k^*.

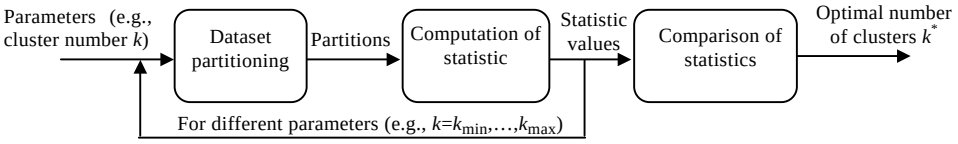


Fig.1 A typical process for determining the number of clusters in a dataset

图 1 典型的确定数据集最佳聚类数的计算过程

各种类型的统计指标从不同角度出发衡量数据集划分的聚类质量.聚类有效性指标(cluster validity index,简称 CVI)是一类常见的指标,在为数众多的 CVI[6−13]中,基于数据集几何结构的指标函数具有代表意义,它们考虑了聚类的基本特征,即一个“好”的聚类结果应使得 k 个簇的簇内数据点是“紧凑”的,而不同簇的点之间是尽可能“分离”的,指标量化聚类的簇内紧凑度和簇间分离度并组合二者[6,7],代表性的指标包括 Xie-Beni 指标[8]V_{xie} 和 S.Wang-H.Sun-Q.Jiang 指标[9]V_{wsj} 等.对应于最大或最小指标值的 k 被认为是最佳聚类数 k^*.其他类型的统计指标包括 Gap statistic[2]、信息熵[3]和 IGP(in-group proportion)[4]等,其中,IGP 是一种新近提出的指标,它使用簇内数据点的 in-group 比例来衡量聚类结果的质量,取得了优于现有其他指标的性能[4].

然而,现有的工作[2−13]多集中在对统计指标的改进上,而忽略了对计算过程的研究.图 1 所示的计算过程存在两个问题:首先,由于需要多次地对整个数据集进行聚类,其效率显然与所选用聚类算法本身的效率密切相关,且将随着数据集的增大而显著下降.尽管可以通过减少聚类次数来提高效率,然而估计准确的 $k_{\min}$ 和 $k_{\max}$ 也是不容易的[14];其次,指标被特定的聚类算法联系在一起.例如,实际中许多 CVI[6−12]是与 FCM(fuzzy C-mean)算法(或融合 GA(genetic algorithm)算法[13])结合在一起的,上述的其他指标出于计算的需要总要选择以聚类数 k 为参数的算法,如 k-means 或层次型聚类算法[2−4].这导致了指标性能依赖于聚类算法的问题.例如,FCM 和 k-means 算法存在不能有效识别噪声和只能发现凸形簇的局限性,使得这些指标自然地存在同样的缺陷.

本文提出的新方法 COPS(clusters optimization on preprocessing stage)采用与图 1 完全不同的两阶段计算方案,首先通过扫描一遍数据集一次性地构造出数据集所有合理的划分组合,进而生成一条关于不同划分的聚类质量曲线;在第 2 阶段抽取出对应曲线极小值点的划分来估计数据集的最佳聚类数目,从而避免了对大型数据集的反复聚类,且不依赖于特定的聚类算法,能够有效识别数据集中可能包含的噪声和复杂形状的簇.在实际数据和合成数据上的测试结果表明,COPS 的性能优于 IGP,同时大幅度提高了计算效率.

本文第 1 节给出 COPS 方法.第 2 节提出和分析 COPS 使用的聚类有效性新指标.第 3 节进行实验验证和分析.最后在第 4 节作出总结.

1 COPS 方法

给定 d 维数据集 $DB=\{X_1,X_2,\ldots,X_n\}$,$X=\{x_1,x_2,\ldots,x_d\}$为一个数据点,$n$ 为数据点数目.一个硬划分聚类算法[1]将 DB 划分为 $k(k>1)$个子集的集合 $C^k=\{C_1,C_2,\ldots,C_k\}$,且$\forall j\neq l,1\le j,l\le k,C_j\cap C_l=\Phi,C_j$ 称为 DB 的簇.常用的 trail-and-error 方法为产生 $C^k(k=k_{\min},\ldots,k_{\max})$需要对数据集进行 $k_{\max}-k_{\min}+1$ 次聚类,这影响了算法效率,尤其当数据量较大时;另一方面,不恰当的 $k_{\min}$ 和 $k_{\max}$ 设置也会影响计算结果的准确性.因此,若能根据数据集的几何结构一次性地生成所有合理的划分,同时评估它们的聚类质量,就可以在很大程度上提高计算效率和结果的准确性.COPS 借鉴层次聚类的思想来达到这个目的.其原理是,首先将每个数据点看作单独的簇,然后在自底向上层次式的簇合并过程中生成所有合理的划分,同时计算它们的聚类质量,保存具有最优聚类质量的划分为 C^*,最后根据 C^*的有关统计信息来估计聚类数 k^*.这里使用新的聚类有效指标性指标函数 $Q(C)$来评估划分 C 的聚类质量,其最小值对应最优的质量.COPS 计算过程的形式化表示如下:

$$C^*=\arg\min_{C^k\in\{C^1,C^2,\ldots,C^n\}}Q(C^k),\ k^*=\theta(C^*).$$

COPS 的处理对象是可能包含噪声和复杂形状(非凸形)簇的数据集,这样的数据在实际应用中是常见的,如空间数据和一些高维度的数据.过程θ剔除噪声的影响,识别 C^*中有意义的簇的数目.

1.1 算法原理

下面,首先结合距离和维度投票(dimension voting)[15]思想提出数据点间相似度的定义,以此为基础给出确定 DB 最优划分 C^*的计算过程.

定义 1(点的相似维度[15]). 给定一个阈值 $t_j\ge 0,1\le j\le d$,若$|x_j-y_j|\le t_j$,则称点 X 和 Y 是关于 t_j 第 j 维相似的.

根据定义 1,可以定义相似的数据点.

定义 2(相似点). 给定一个阈值向量 $T=\{t_1,t_2,\ldots,t_d\}$,若点 X 和 Y 在所有数据维度上都是关于 $t_j(j=1,2,\ldots,d)$相似的,则称数据点 X 和 Y 是关于 T 相似的.

彼此相似的数据点组成 DB 的簇.若 T 的各分量相等,则$\|T\|$相当于基于密度聚类算法[16]中点的邻域半径.这里,我们允许 T 的各分量不相等,它反映了数据集各维度属性值分布的差异.显然,T 的取值"大小"决定了簇结构.鉴于 T 是一个向量,定义 3 用于比较不同 T 之间的相对关系.

定义 3(T 的比较). 给定两个阈值向量 $T^a=\{t_1^a,t_2^a,\ldots,t_d^a\}$ 和 $T^b=\{t_1^b,t_2^b,\ldots,t_d^b\}$,称 $T^a>T^b$ 若满足:

(1) $t_j^a\ge t_j^b,\ j=1,2,\ldots,d$;

(2) 至少存在一个 $j\in[1,d]$,使得 $t_j^a>t_j^b$.

给定数据集 DB,根据定义 2 和定义 3,一个很小的 T 令大多数数据点不相似,极端情形是每个数据点(设 DB 中没有相同的数据点)都构成单独的"簇",此时,DB 的划分数目达到最大值,$k=n$;记这样的 T 为 T^0.相反地,一个足够大的 T 将使得所有数据点彼此相似而组成一个大簇,此时,k 达到最小值,$k=1$;用 T^m 表示使得 $k=1$ 的最小的 T.

至此,可以把确定数据集 DB 最优划分 C^*的问题转换为求解最优阈值向量 $T^*(T^0<T^*<T^m)$的问题,T^*是使得 Q 在 T 从 T^0 开始增大到 T^m 过程中取得最小值的阈值向量.由此可以给出 COPS 的计算过程,如图 2 所示.

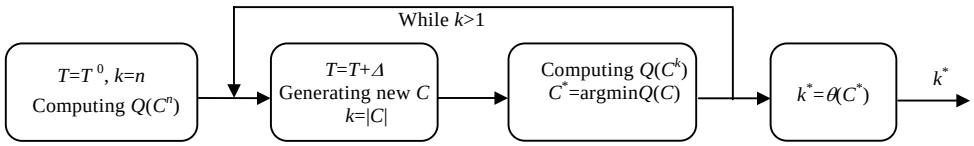


Fig.2 Flowchar of the COPS

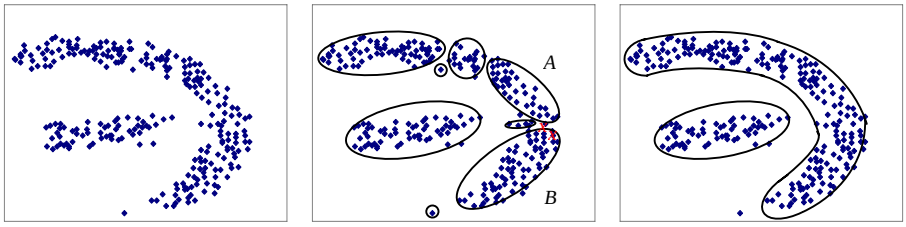
图 2 COPS 的计算流程图

计算从 $T=T^0$(每个分量值为 0)开始,每个步骤 T 增大一个量$\Delta(\Delta=\{\Delta_1,\Delta_2,\ldots,\Delta_d\})$,根据定义 2,此时将有部分原本属于不同子集的点变得相似,这些子集被合并,生成了新的划分;对每个划分计算其 Q 的值,直到 T 增长到所有

点被划分到同一个集合为止.这里子集的合并,也就是簇的合并,以类似 single link 的方式[1]进行.图 3 给出一个例子,说明图 2 的计算过程如何自底向上地进行簇的合并.

图 3 给出了 COPS 在一个 2 维数据集上若干步骤的结果,该数据集包含 2 个簇和 1 个噪声点,其中一个簇是非凸形的.如初始状态图 3(a)所示,所有的数据点均构成独立的簇;随 T 的增大数据点被逐渐合并,假设在某个步骤形成了图 3(b)所示的椭圆形区域所代表的若干个小簇;当 T 进一步增大使得两个分属于不同簇的点(如图 3(b)中分属于簇 A 和 B 的两个标志为'x'的点)变得相似时,两个原本为凸形的簇被合并.基于这样的策略可以生成任意形状的簇,如图 3(c)所示,最终合并成了一个 banana 型的非凸形簇.图 3 中,3 个阶段的簇结构组成聚类树的 3 个层次(实际的聚类树可能不止这 3 个层次,作为一个例子,这里只给出其中的 3 层).对于每个层次上的簇集合,分别计算它们的 Q 值,抽取出其中使得 Q 取值最小的层次,识别其中的噪声点.考虑图 3(c)所示的簇集合,位于下方只包含一个数据点的"簇"被识别为噪声,这样就得到了该数据集的最佳聚类数目为 2.

以上簇的合并方法与 single link[1]的区别是,传统的 single link 方式以全空间的欧氏距离为基础,而 COPS 依据定义 2 来衡量簇间的相似度.Single link 已被证明可以识别数据集中非凸形的簇结构[1],我们在此基础上增加考虑了数据集各维度属性值分布的差异因素,这种差异在具有较高维度的实际应用数据中是常见的[17].Δ的选取以及随 T 的变化如何快速地生成新的划分和计算 Q 值是影响算法性能的重要环节,以下章节将分别阐述.



(a) (b) (c)

Fig.3 An example of the working flow of the COPS

图 3 COPS 计算过程示例

1.2 算法过程和参数设置

COPS 的计算过程类似于凝聚型层次聚类算法中层次聚类树[17]的构建过程.然而,一个典型的层次聚类算法具有 $O(n^2)$的计算复杂度[1].性能的瓶颈在于为每个数据点 X 查找与其相似点的集合 $Neighbors(X)$,这是 range queries 问题[18],通常这需要遍历整个数据集,一些改进方法可参见文献[18].

基于定义 2 可以简化查找相似点的过程.首先将数据点按照每个维度的属性值大小进行排序(每个维度 j 有一个排序的序列 A_j);根据定义 1,通过顺序扫描 A_j 可以得到所有与点 X 第 j 维相似的点 Y,扫描范围局限在符合$|x_j-y_j|\le t_j$条件的有限区间内.当 t_j 增加Δ_j时,只需在原有范围的基础上扩展扫描区间 $t_j<|x_j-y_j|\le t_j+\Delta_j$即可.基于此优化方法的 COPS 伪代码如图 4 所示.其中 *MergePartitions* 的功能是在 C^{k+1} 的基础上合并两个相似点 X 和 Y 所在的子集生成新的划分 C^k;*UpdateQ* 在 $Q(C^{k+1})$的基础上根据 X 和 Y 所在子集的统计信息计算得到新的值 $Q(C^k)$.第 2 节将阐述基于 CF 的子集合并和 $Q(C)$的计算方法.

算法参数Δ的选取与 T 的一个隐含性质有关.考察定义 2,T 可以看作是聚类数据集的分辨率(clustering resolution)[19],而分辨率可以被想象为一个看待数据点是否构成簇的"望远镜".由此可知,T 的各分量间应成一定的比例关系,这个比例与数据点投影到各维度上时点的分布密度有关.同理,Δ的各分量间也应具有这个性质.定义 4 通过度量维度的稀疏度来量化这种比例关系.

定义 4(维度稀疏度). 数据集 DB 在第 j 维的分布稀疏度为λ_j,

$$\lambda_j=\sqrt{\frac{\sum_{i=1}^{n}(x'_{ij}-\mu_j)^2}{n-1}},$$

其中,x'_{ij} 是数据点 X_i 第 j 维属性的[0,1]规范化值,μ_j 表示第 j 维的中心,即

$$x'_{ij}=\frac{x_{ij}-\min\limits_{l=1,\ldots,n}\{x_{lj}\}}{\max\limits_{l=1,\ldots,n}\{x_{lj}\}-\min\limits_{l=1,\ldots,n}\{x_{lj}\}},\ \mu_j=\frac{1}{n}\sum_{i=1}^{n}x'_{ij}\ .$$

```
Algorithm. COPS (DB,Δ).
begin
  k=n,T=T^0,C^n={{X1},{X2},…,{Xn}},Q^n=Q(C^n)
  For each dimension j∈[1,d] do
    Aj=Points sorted on the values of jth-attributes
  {1. Generating Q-sequence}
  Repeat
    For each dimension j∈[1,d] do
      For each point X∈Aj do
        For each point Y∈Neighbors(X,Aj,tj,tj+Δj)
        begin
          Flag X and Y are jth-similar
          If X and Y are full-dimensional similar
          begin
            k=k–1
            C^k=MergePartitions(C^{k+1},X,Y)
            Q^k=UpdateQ(Q^{k+1},X,Y)
          end;
        end; {for Y∈Neighbors()}
    T=T+Δ
  Until k=1
  {2. Computing k*}
  C*=Partitions having the minimum of Q-sequence
  Return k*=θ(C*)
end;
```

Fig.4 The pseudocodes of the COPS

图 4 COPS 的伪代码

λ_j 实际上是数据集第 j 维规范化的标准偏差.在高维数据的投影聚类[20]中,正是以标准偏差为基础度量维度与簇之间的相关程度.λ_j值越大,表明第j维属性值分布得越稀疏,与其相关的簇也可能就越多.COPS利用这些维度上属性值的变化来揭示数据集潜在的簇结构,因此可用以下公式来确定算法的参数Δ:

$$\Delta_j=\varepsilon\times\frac{\max\{\lambda_1,\lambda_2,\ldots,\lambda_d\}}{\lambda_j},$$

其中,ε(ε>0)是给定的一个具有很小的数值的算法参数,用于控制计算Q序列的精度.显然,ε越小,COPS在每个维度上的搜索步数(即每个维度被分割成的用于计算的区间数)就越多,因而也就扩大了算法搜索数据集最优划分的搜索空间,其结果也就越有可能是最优的结果.另一方面,ε越小,将使得算法的时间开销越大,因而需要在这两者之间取一个平衡点.经过实验环节的反复验证,我们设定ε=0.01.

1.3 确定k^*的值

$|C^*|$是候选的聚类数 k^*,但由于噪声的影响,$k^*=|C^*|$并不完全成立.在 COPS 中,噪声数据点也是 C^*的组成部份,其特点是这些子集所包含的数据点数目较少[19].设$|C^*|>2$,过程θ采用基于 MDL(minimal description length)的剪枝方法[21]识别出 C^*中“有意义”的子集.MDL 的基本思想是,对输入的数据进行编码,进而选择具有最短编码长度的编码方案.在 COPS 中,其输入的数据是各个子集包含的数据点数目,簇的重要性由其包含数据点的数目来决定.

令 $C^*=\{C_1^*,C_2^*,\ldots,C_k^*\}$,$|C_i^*|$为 C_i^* 包含的数据点数目.首先按照$|C_i^*|$从大到小排序生成一个新的序列 C_1,C_2,…,C_k;然后将这个序列以 C_p(1<p<k)为界分为两个部分:$S_L(p)=\{C_1,C_2,\ldots,C_p\}$和 $S_R(p)=\{C_{p+1},C_{p+2},\ldots,C_k\}$,求得数据的编码长度 $CL(p)$.$CL(p)$的定义[21]为

$$CL(p)=\log_2(\mu_{S_L(p)})+\sum_{1\le j\le p}\log_2(\|C_j|-\mu_{S_L(p)}|)+\log_2(\mu_{S_R(p)})+\sum_{p+1\le j\le k}\log_2(\|C_j|-\mu_{S_R(p)}|),$$

其中,$\mu_{S_L(p)}=\left\lceil\sum_{1\le j\le p}|C_j|/p\right\rceil$,$\mu_{S_R(p)}=\left\lceil\sum_{p+1\le j\le k}|C_j|/(k-p)\right\rceil$.上式的第 1 项和第 3 项分别表示以 p 为界的两个

序列的平均编码长度;其余两项衡量$|C_j|$与平均数据点数之间的差异.实际计算中若出现$|C_j|=\mu_{S_L(p)}$或$|C_j|=\mu_{S_R(p)}$而使得 log 函数没有定义,则直接忽略该子集,即设定此时的差异为 0 bit.

最短的编码长度 $CL(p)$对应的分割位置 p 被看作数据序列的最优分割点,根据 MDL(minimal description length)剪枝方法的思想,此时 $S_L(p)$所包含的数据点可以认为代表了对 DB 的覆盖[21].在 COPS 中,$S_R(p)$所包含的数据点就识别为噪声.至此,我们得到了数据集的最佳聚类数,$k^*=p$.

1.4 算法复杂度

最坏情况下,COPS 的空间复杂度为 $O(n^2)$,实际中采用以下策略降低算法的空间使用量:对任意两个不相似的数据点 X_i 和 X_j,若 X_i 和 X_j 至少在一个维度上是相似的,则通过一个 $hash$ 函数映射到一个线性表中的单元 $HASH(i,j)$,$HASH(i,j)$记录 X_i 和 X_j 的维度相似情况.算法开始时,该表的所有单元为空(未被使用),随着 T 的增大,一些点对变得相似时,其映射在线性表中的单元亦被释放,从而有效降低了算法的实际空间占有量.

采用快速排序(quicksort)方法对数据点进行排序的时间复杂度为 $O(dn\log n)$.生成 Q 序列部分算法的时间复杂度为 $O(\bar{k}dn\bar{N})$,其中,$\bar{k}$ 为外层循环的执行次数,是一个与 n 无关的量,在数值上,$\bar{k}<<n$.这是因为随着 T 的增大,越来越多的点变得相似,k 在内层循环中将迅速减少,其数值只与数据点的分布和ε的取值有关;$\bar{N}$ 是数据点在Δ邻域内的平均相似点数目,在数值上 $\bar{N}<<n$,它也只与数据点分布和ε有关.计算初始值 $Q(C^n)$的复杂度为 $O(dn)$(参见第 2.3 节分析),MDL 剪枝方法的复杂度为 $O(k^2)$.综上,COPS 的时间复杂度为 $O(dn\log n)$.

2 COPS 的聚类有效性指标

COPS 用有效性指标 $Q(C)$评估 DB 被划分为 C 时的聚类质量.本节提出的指标 $Q(C)$主要考虑数据集的几何结构,即通过衡量簇内数据点分布的紧凑度以及簇间的分离度,并保持二者之间的平衡.$Q(C)$不依赖于具体的聚类算法.

2.1 新的有效性指标

设$||X-Y||$表示点 X 和 Y 之间的欧氏距离,给定 DB 的一个划分 $C^k=\{C_1,C_2,\ldots,C_k\}$,$Scat(C^k)$衡量 C^k 的簇内紧凑度,$Sep(C^k)$对应 C^k 的簇间分离度.具体地,

$$Scat(C^k)=\sum_{i=1}^{k}\sum_{X,Y\in C_i}\|X-Y\|^2 \tag{1}$$

$$Sep(C^k)=\sum_{i=1}^{k}\left(\sum_{l=1,l\neq i}^{k}\frac{1}{|C_i|\cdot|C_l|}\sum_{X\in C_i,Y\in C_l}\|X-Y\|^2\right) \tag{2}$$

以上两式的定义原理如下:$Scat$ 是簇内任意两个数据点之间距离的平方和;Sep 的原理是将每个簇看作是一个大“数据点”,大“数据点”间的“距离”通过簇间点对的平均距离来衡量.这样,$Scat$ 和 Sep 保持了度量上的一致性.另一方面,$Scat$ 和 Sep 基于“点对”的平均距离定义,可用于度量非凸形簇结构的聚类质量.传统的基于几何结构的聚类有效性指标(如 V_{xie}[8])通常基于簇质心(centroids)使用簇的平均半径和质心之间的距离来定义 $Scat$ 和 Sep,这样的指标往往只对球(超球)形的簇结构有效[7].

代入欧氏距离公式再做简单的变换,$Scat(C^k)$和 $Sep(C^k)$可分别表示为

$$Scat(C^k)=2\sum_{j=1}^{d}\sum_{i=1}^{k}(|C_i|SS_{ij}-LS_{ij}^2),$$

$$Sep(C^k)=2\sum_{j=1}^{d}\left((k-1)\sum_{i=1}^{k}\frac{SS_{ij}}{|C_i|}-\left(\sum_{i=1}^{k}\frac{LS_{ij}}{|C_i|}\right)^2+\sum_{i=1}^{k}\frac{LS_{ij}^2}{|C_i|^2}\right),$$

其中,$SS_{ij}=\sum_{X\in C_i}x_j^2$,$LS_{ij}=\sum_{X\in C_i}x_j$.直观上,$Scat$ 的值越小,表明簇越紧凑;Sep 的值越大,表明簇间的分离性越好.在下式中使用线性组合平衡二者,$\beta(\beta>0)$为组合参数,用于平衡 $Scat$ 和 Sep 取值范围上的差异:

$$Q_1(C)=Scat(C)+\beta\cdot Sep(C).$$

这里将数据集的划分 C 看作一个变量,其定义域为$\{C^1,C^2,\ldots,C^n\}$.根据定理 1 可以推定β=1.

定理 1. 给定数据集 DB,$Scat(C)$和 $Sep(C)$具有相同的值域范围.

证明:在初始状态 $k=n$,$C^n=\{\{X_1\},\{X_2\},\ldots,\{X_n\}\}$,由公式(1)可知 $Scat(C^n)=0$;根据公式(2)有

$$Sep(C^n)=2\sum_{j=1}^{d}\left(n\cdot\sum_{X\in DB}x_j^2-\left(\sum_{X\in DB}x_j\right)^2\right)=M \tag{3}$$

设在某个步骤 C_u 和 $C_v(u,v\in[1,k],u\neq v,k>1)$合并:

$$\begin{aligned}Sep(C^{k-1})-Sep(C^k)=&-2\sum_{j=1}^{d}\frac{LS_{uj}LS_{vj}}{|C_u||C_v|}-\sum_{j=1}^{d}\left((k-2)\frac{|C_v|^2SS_{uj}+|C_u|^2SS_{vj}}{(|C_v|+|C_u|)|C_u||C_v|}+\sum_{i=1}^{k}\frac{SS_{ij}}{|C_i|}\right)-\\&2\sum_{j=1}^{d}\left(\frac{|C_v|^2LS_{uj}+|C_u|^2LS_{vj}}{(|C_v|+|C_u|)|C_u||C_v|}\sum_{i=1,i\neq u,v}^{k}\frac{LS_{ij}}{|C_i|}\right)<0\end{aligned} \tag{4}$$

$$Scat(C^{k-1})-Scat(C^k)=2\sum_{j=1}^{d}(|C_u|SS_{vj}+|C_v|SS_{uj}+2LS_{uj}LS_{vj})>0 \tag{5}$$

因此,$Scat(C)$是单调递增函数,而 $Sep(C)$为单调递减函数.当 k=1 时,$C^1=\{\{X_1,X_2,\ldots,X_n\}\}$,容易求得 $Sep(C^1)$=0 和 $Scat(C^1)=M$. □

根据定理 1,COPS 使用的聚类有效性指标函数 $Q(C)$取以下形式:

$$Q(C)=\frac{1}{M}(Scat(C)+Sep(C)) \tag{6}$$

2.2 指标分析

最优的聚类质量对应于簇内紧凑度和簇间分离度的平衡点[9,10],在数值上反映为指标函数 $Q(C)$取得最小值.定理 2 表明,对于大多数(一种特例除外,见定理条件)数据集 $Q(C)$存在(0,1)区间的最小值.

定理 2. 给定数据集 $DB=\{X_1,X_2,\ldots,X_n\}$,若 $n>2$ 且至少存在一个 $i\in[2,n-1]$使得$\|X_{i-1}-X_i\|\neq\|X_i-X_{i+1}\|$,则 $Q(C)$存在小于 1 的极小值.

证明:考虑在 COPS 的初始状态 $k=n$,$C^n=\{\{X_1\},\{X_2\},\ldots,\{X_n\}\}$.令 $t_j=\min_{l=1,2,\ldots,n-1}\{x_{lj}-x_{(l+1)j}\}$,$j=1,2,\ldots,d$,若满足定理条件,根据定义 1,$\exists u,v\in[1,n],u\neq v$,使得 X_u 和 X_v 是相似的,且(X_{i-1},X_i)和(X_i,X_{i+1})中至少有 1 对是不相似的,后者确保对所有相似点做合并处理后 $k>1$.考虑点 X_u 和 X_v 合并后 Q 的变化,根据公式(3)~公式(6)有

$$Q(C^{n-1})-1=-\frac{n-2}{2M}\sum_{j=1}^{d}(SS_{uj}+SS_{vj})-\frac{1}{M}\sum_{j=1}^{d}\left(\left(\sum_{i=1,i\neq u,v}^{n}SS_{ij}\right)+(LS_{uj}+LS_{vj})\sum_{i=1,i\neq u,v}^{n}LS_{ij}\right)<0.$$

定理 1 已经证明 $Q(C^1)=Q(C^n)=1$,这意味着若满足定理条件,则 $Q(C)$存在小于 1 的极小值. □

定理 2 给出了 $Q(C)$无法取得(0,1)区间极小值的一种特殊结构的数据集,其直观情形是所有数据点均匀地分布在空间等分网格的节点上.对这样的数据集,COPS 将输出 $k^*=n$.这是合理的,因为此时,合理的 k^*取值为 1 或 n,而聚类算法通常要求 $k^*>1$.

2.3 指标计算

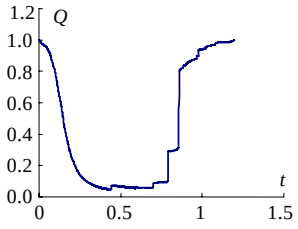


Fig.5 An example of a $Q(C)$ curve

图 5 $Q(C)$曲线的一个例子

根据公式(3)~公式(6)可以进行 $Q(C^k)$的增量计算(*UpdateQ* 过程).为了得到 $Q(C^k)$,只需在 $Q(C^{k+1})$的基础上使用公式(4)和公式(5)计算其增量即可.为此,在算法中为每个 $C_i(i=1,2,\ldots,k)$保存一个结构:

$$CF_i=(|C_i|,\langle SS_{i1},LS_{i1}\rangle,\langle SS_{i2},LS_{i2}\rangle,\ldots,\langle SS_{id},LS_{id}\rangle).$$

这正是 BIRCH 算法提出的聚类特征(clustering feature)[17].基于此,合并数据集划分的操作(*MergePartitions* 过程)可以转化成为相应的 $|C_i|$,SS_{ij} 和 LS_{ij} 数值之间简单的加法运算.在计算初始值 M 时,需获得每个数据点的 CF 结构,再按照公式(3)计算,其时间复杂度为 $O(dn)$.

一条典型的 $Q(C)$曲线例子如图 5 所示.图中,在最小值之后 Q 值出

现大幅跳变,这意味着若合并最优划分的一些子集将使得聚类质量急剧下降.利用这个特点,此时加大 T 的增量 Δ可以进一步提高 COPS 的性能.具体地,当数据集较大时(比如 n>1000),计算过程中若 Q 大于已出现的最小值,我们设定$\varepsilon=\varepsilon\times2$.

3 实验与分析

实验验证包括算法有效性和算法效率两方面.在众多的聚类有效性指标[6−13]里,选用基于几何结构的 V_{xie}[8] 和 V_{wsj}[9]这两种有代表意义的指标作为对比对象,其中,V_{xie} 是首个采用“紧凑度”和“分离度”概念的经典指标;V_{wsj}改进了线性组合方法的稳定性,可以有效地处理包含有重叠的簇和噪声的数据集[9,12].两种指标都基于FCM算法,实验设定 FCM 算法的模糊因子 w=2.在其他类型方法中,选用 Gap statistic[2]和 IGP[4]作比较.Gap statistic 的特点是通过检测聚类质量的“突变(dramatic change)”确定最佳的 k 值;IGP 是新近提出的一个指标,使用 in-group 比例衡量聚类的质量,其性能已被验证优于现有的其他统计指标[4].根据文献[2,4]的建议,使用 k-means 作为它们的基本算法,取参数 R=5;IGP 使用的 Cutoff 阈值设置为 0.90.使用 Greedy 技术[9]选择 FCM/k-means 的初始簇中心点以提高算法的收敛速度.实验在 CPU 2.6GHz,RAM 512MB 的计算机上进行,操作系统为 Microsoft Windows 2000.

3.1 实验数据

实验分别采用了真实数据和人工合成数据.以下报告 6 个有代表性的数据集上的实验结果,数据集的参数汇总见表 1.为比较起见,选取的前 2 个数据集 *DS*1 和 *DS*2 是常被类似研究引用的真实数据 X30 和 IRIS 数据[3,7−9,12].*DS*3 和 *DS*4 是两个具有较高维度的实际应用数据.*DS*3 来源于 Vowel Recognition(deterding data)(http://www.ics.uci.edu/~mlearn/databases/undocumented/connectionist-bench/vowel/),包含有 10 个说话人、11 个英语元音的发音数据,用于语音识别研究;*DS*4 来源于Wisconsin Breast Cancer Database(http://mlearn.ics.uci.edu/databases/breast-cancer-wisconsin/),是患者肺部 FNA 测试的临床数据,用于医疗诊断.

Table 1 Summarized parameters of datasets

表 1 测试数据参数汇总表

DB	Description of the dataset	Dimension (d)	Size (n)	True number of clusters
*DS*1	X30	2	30	3
*DS*2	IRIS	4	150	3
*DS*3	Vowel recognition (deterding data)	10	528	11
*DS*4	Wisconsin breast cancer database	9	699	2
*DS*5	Synthetic dataset	3	4 000	6
*DS*6	*t*5.8*k*	2	8 000	6

为测试各种方法处理大型数据集的性能,根据文献[17]提供的方法(在原方法基础上改进为随机簇中心)合成了含 4 000 个数据点的 3 维数据集 *DS*5.*DS*5 同时还包含少量的噪声,这些噪声模糊了簇的边界,其中两个簇存在明显的重叠.第 6 个数据集 *DS*6 包含有 8 000 个数据点,是命名为“*t*5.8*k*”的公用数据集,其特点是包含有大量的噪声和复杂形状的簇(其 6 个簇呈‘GEOAGE’字母形状).更为重要的是,我们通过实验发现,在适当的参数配置下,即指定了正确的聚类数,FCM 和 k-means 算法可以很好地区分出这 6 个簇.以此验证基于 FCM 或 k-means 算法的其他 4 种方法以及 COPS 识别复杂形状簇的性能.

3.2 有效性实验

COPS 在 6 个数据集上都得到了正确的聚类数,实验结果如图 6 所示.对 *DS*1 和 *DS*2,COPS 检测出在聚类数从最优数目 3 变为 2 时聚类质量大幅度下降.*DS*2 包含有两个重叠的簇[12],图 6(b)表明,COPS 能够有效区分重叠的簇.受噪声影响,*DS*5 中有两个边界模糊的簇,图 6(e)显示,对应于聚类数 6 和 5 的聚类质量只存在很小的差异,尽管如此,COPS 还是完成了准确的区分,得到最优聚类数 6,这说明 COPS 可以有效地识别噪声并区分簇间的密度差异.对 *DS*3~*DS*6 这 4 个较为复杂的数据,COPS 没有检测到连续变化的 k 值,例如在图 6(f)中,k 从 18 跳变到最优数 6.这是因为 COPS 采用了与其他方法不同的做法,它不是通过设定一个 k 的区间反复运行聚类算法来

计算和比较不同的聚类结果,而是在层次式的簇合并过程中计算不同划分的质量,合并过程中产生的簇的数目取决于数据集本身的结构,这正是 COPS 在识别复杂形状簇方面的优势.

不同方法的实验结果对比见表 2.实验中为 V_{xie} 和 V_{wsj} 设置的 k 值范围是[2,12],对 Gap 和 IGP 设置为[1,11].文献[24]建议设定 $k_{\max} \le \sqrt{n}$,根据本文测试数据集的大小则会得到很大的 $k_{\max}$ 值,因而作了以上统一的设置.由于 FCM/k-means 算法的聚类结果容易受初始簇中心的影响,根据 Greedy 技术[8]的原理,其第 1 个初始簇中心是随机选择的,使得它们在复杂数据集上产生不确定的结果;Gap statistic 使用的随机空分布数据(不含簇结构)和 IGP 使用的随机数据抽样是另一个因素.表 2 列出了它们在多次实验中最接近真实聚类数的结果.作为对比,COPS 使用的 $Q(C)$指标独立于具体的聚类算法,它在由数据集几何结构所确定的层次聚类树上搜索最优的划分,由此得到的结果具有确定性的特点.

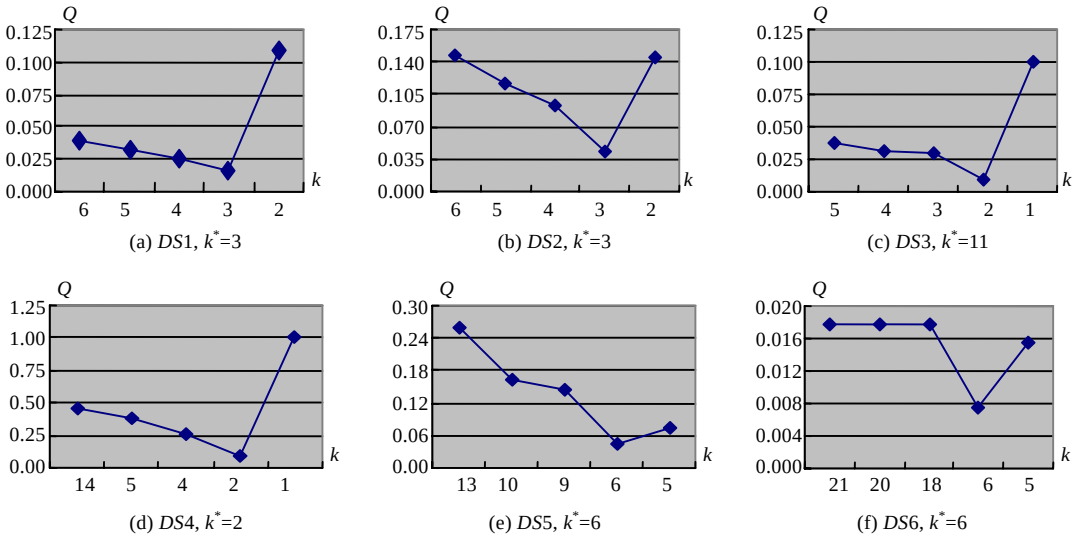


Fig.6 Experimental results of COPS on the datasets

图 6 COPS 在 6 个数据集上的实验结果

Table 2 The optimal number of clusters yielded by different methods

表 2 不同方法得到的最佳聚类数

DB	True number of clusters	COPS	V_{xie}	V_{wsj}	Gap statistic	IGP
*DS*1	3	**3**	**3**	**3**	**3**	**3**
*DS*2	3	**3**	2	**3**	4	**3**
*DS*3	11	**11**	2	**11**	2	3
*DS*4	2	**2**	**2**	5	4	**2**
*DS*5	6	**6**	4	5	5	**6**
*DS*6	6	**6**	2	4	3	9

从表 2 可以看出,对于 3 个簇明显分离的简单数据集 *DS*1,所有方法都得到了正确的最优聚类数 3.对于 *DS*2,只有 COPS,V_{wsj} 和 IGP 得到正确的结果,这在一定程度上验证了 V_{wsj} 和 IGP 能够正确处理重叠簇的能力.然而在 *DS*6 上,只有 COPS 得到正确的结果.需要指出的是,FCM/k-means 算法在设置 k=6 时可以对 *DS*6 的 6 个簇进行较好的区分,但 4 种对比方法均未能正确地计算,说明这些方法在识别非凸形簇方面存在缺陷.Gap statistic 需要随机生成的空分布数据作为对比以检测聚类质量存在跳变的 k 值,在多次尝试后我们没有得到更好的结果.实际上,除 *DS*1 外,Gap statistic 都返回错误的簇数目,这是由于 Gap statistic 只适合于处理簇间明显分离的数据集[2].表 2 显示 IGP 具有较好的性能,但在 *DS*3 和 *DS*6 上得到错误的结果.*DS*3 的数据点数较少(n=528),却具有较高的维度(d=10)和较多的簇(k^*=11),令 IGP 使用的 in-group 指标(对于簇内的每个数据点,计算离它最近的点也处在同一个簇内的比例)失效.

3.3 算法效率分析

COPS 与其他方法的运行时间对比见表 3.由于 *DS*1 和 *DS*2 分别只有 30 个和 150 个数据点,不同方法计算时间的差异较小(均小于 0.5s),表 3 略去 *DS*1 和 *DS*2 的时间比较.表 3 显示,COPS 大幅度提高了确定数据集最佳聚类数的计算效率,性能提升在大型数据集上(*DS*5 和 *DS*6)尤为明显.给定一个 k 值时,Gap statistic 和 IGP 需要执行 R(算法参数)次聚类和指标值计算,令它们的计算效率低于 V_{xie} 和 V_{wsj}.IGP 使用的 in-group 指标使得它可以获得比其他 3 种方法更为准确的结果(第 3.2 节),但却以更多的时间消耗为代价.根据 in-group 的定义,计算指标值时需要扫描数据集获得每个数据点的最近邻居点,并检查其是否属于同一个簇.

Table 3 Comparison of execution time

表 3 运行时间对比

DB	Size (n)	Execution time (s) of different methods				
		COPS	V_{xie}	V_{wsj}	Gap statistic	IGP
*DS*3	528	0.3	2.3	2.5	2.8	3.8
*DS*4	699	0.3	2.8	3.3	4.0	4.0
*DS*5	4 000	1.4	12.6	11.5	10.5	31.9
*DS*6	8 000	3.7	13.3	14.4	21.9	97.8

COPS 的性能优势来源于它在方法上与传统方法的根本区别.COPS 通过扫描一遍数据集获得数据集的统计信息,在增量生成的数据集划分中快速地计算划分质量,输出最优划分对应的聚类数.而其他方法需要对数据集多次反复地进行聚类,检测每遍聚类结果,其性能与算法的执行次数与算法本身的效率密切相关.

4 小结及工作方向

本文提出一种基于层次思想的计算方法 COPS,用于自动确定大型、复杂数据集的最佳聚类数目.COPS 与常用的 trial-and-error 型方法的不同之处在于,它不需要多次运行特定的聚类算法,而是首先使用新提出的聚类质量度量函数 $Q(C)$,在自底向上层次式的簇合并过程中确定使得 $Q(C)$达到最小值的数据集最优划分,再基于 MDL 剪枝原理估计最佳的聚类数.生成数据集划分和计算 $Q(C)$值均增量地进行,使得算法相对数据点数目具有接近线性增长的时间复杂度.$Q(C)$基于数据集的几何结构,独立于具体的聚类算法.在真实数据和合成数据上的实验结果表明,它可以有效地在含有噪声、重叠的和复杂形状簇的数据集上快速地计算得到正确的最佳聚类数.其性能优于新近提出的 IGP 方法,并大幅度提升了计算效率.我们通过定理 2 证明了 COPS 的普适性.

算法参数ε用于控制计算的精度,ε的取值可能会对计算结果产生影响.尽管实验结果(限于篇幅未列出)表明ε取 0.001~0.1 时 COPS 在测试数据集上都可以得到正确的聚类数目,但分析ε对计算结果的影响和研究如何确定ε的取值是必要的,这也将是我们下一步工作的重点.

致谢 在此,我们感谢对本文提供有价值评论的匿名审稿人.

References:

[1] Berkhin P. Survey of clustering data mining techniques. Technical Report, San Jose: Accrue Software, 2002.

[2] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a dataset via the gap statistic. Technical Report, 208, Stanford University, 2000.

[3] Still S, Bialek W. How many clusters? An information-theoretic perspective. Neural Computation, 2004,16(12):2483−2506.

[4] Kapp AV, Tibshirani R. Are clusters found in one dataset present in another dataset? Biostatistics, 2007,8(1):9−31.

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

[6] Halkidi M, Batistakis Y, Vazirgiannis M. Clustering validity checking methods: Part II. ACM SIGMOD Record Archive, 2002, 31(3):19−27.

[7] Bouguessa M, Wang S, Sun H. An objective approach to cluster validation. Pattern Recognition Letters, 2006,27(13):1419−1430.

[8] Xie X, Beni G. A validity measure for fuzzy clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1991,13(8): 841−847.

[9] Sun H, Wang S, Jiang Q. FCM-Based model selection algorithms for determining the number of cluster. Pattern Recognition, 2004, 37(10):2027−2037.

[10] Fan J, Wu C. Clustering validity function based on possibilistic partition coefficient combined with fuzzy variation. Journal of Electronics and Information Technology, 2002,24(8):1017−1021 (in Chinese with English abstract).

[11] Sun C, Wang J, Pan J. Research on the method of determining the optimal class number of fuzzy cluster. Fuzzy Systems and Mathmatics, 2001,15(1):89−92 (in Chinese with English abstract).

[12] Hong Z, Jiang Q, Dong H, Wang S. A new cluster validity index for fuzzy clustering. Computer Science, 2004,31(10):121−125 (in Chinese with English abstract).

[13] Zhu K, Su S, Li J. Optimal number of clusters and the best partition in fuzzy C-mean. Systems Engineering—Theory & Practice, 2005,25(3):52−61 (in Chinese with English abstract).

[14] Yu J, Cheng G. Search range of the optimal number of clusters in fuzzy clustering. Science in China (Series E), 2002,32(2): 274−280 (in Chinese with English abstract).

[15] Woo KG, Lee JH, Kim MH, Lee YJ. FINDIT: A fast and intelligent subspace clustering algorithm using dimension voting. Information and Software Technology, 2004,46(4):255−271.

[16] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han JW, Fayyad UM, eds. Proc. of the ACM-SIGKDD. Portland: AAAI Press, 1996. 226−231.

[17] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. In: Jagadish HV, Mumick IS, eds. Proc. of the ACM-SIGMOD. New York: ACM Press, 1996. 103−114.

[18] Brecheisen S, Kriegel HP, Pfeifle M. Multi-Step density-based clustering. Knowledge and Information Systems, 2006,9(3): 284−308.

[19] Foss A, Zaïane OR. A parameterless method for efficiently discovering clusters of arbitrary shape in large datasets. In: Kumar V, Tsumoto S, eds. Proc. of the ICDM. Los Alamitos: IEEE Computer Society Press, 2002. 179−186.

[20] Kim M, Yoo H, Ramakrishna RS. Cluster validation for high dimensional datasets. In: Proc. of the AIMSA. LNCS 3192, Berlin, Heidelberg, 2004. 178−187.

[21] Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data. Data Mining and Knowledge Discovery, 2005,11(1):5−33.

附中文参考文献:

[10] 范九伦,吴成茂.可能性划分系数和模糊变差相结合的聚类有效性函数.电子与信息学报,2002,24(8):1017−1021.

[11] 孙才志,王敬东,潘俊.模糊聚类分析最佳聚类数的确定方法研究.模糊系统与数学,2001,15(1):89−92.

[12] 洪志令,姜青山,董槐林,Wang S.模糊聚类中判别聚类有效性的新指标.计算机科学,2004,31(10):121−125.

[13] 诸克军,苏顺华,黎金玲.模糊 C 均值中的最优聚类与最佳聚类数.系统工程理论与实践,2005,25(3):52−61.

[14] 于剑,程乾生.模糊聚类方法中的最佳聚类数的搜索范围.中国科学(E 辑),2002,32(2):274−280.

陈黎飞(1972－),男,福建长乐人,博士生,主要研究领域为数据挖掘,模式识别.

姜青山(1962－),男,博士,教授,博士生导师,主要研究领域为数据挖掘,图像处理,数据库系统,模糊集理论与应用.

王声瑞(1963－),男,博士,教授,博士生导师,主要研究领域为模式识别,数据挖掘,人工智能,图像分析和理解,神经网络,信息系统,决策系统.