

文章编号:1001-9081(2016)06-1634-05

DOI:10.11772/j.issn.1001-9081.2016.06.1634

## 基于层次划分的密度优化聚类算法

逢琳<sup>1,2</sup>, 刘方爱<sup>1,2\*</sup>

(1. 山东师范大学 信息科学与工程学院, 济南 250014; 2. 山东省分布式计算机软件新技术重点实验室, 济南 250014)

(\*通信作者电子邮箱 lfa@sdnu.edu.cn)

**摘要:**针对传统的聚类算法对数据集反复聚类,且在大型数据集上计算效率欠佳的问题,提出一种基于层次划分的最佳聚类数和初始聚类中心确定算法——基于层次划分密度的聚类优化(CODHD)。该算法基于层次划分,对计算过程进行研究,不需要对数据集进行反复聚类。首先,扫描数据集获得所有聚类特征的统计值;其次,自底向上地生成不同层次的数据划分,计算每个划分数据点的密度,将最大密度点定为重心点,计算重心点距离更高密度点的最小距离,以重心点密度与最小距离乘积之和的平均值为有效性指标,增量地构建一条关于不同层次划分的聚类质量曲线;最后,根据曲线的极值点对应的划分估计最佳聚类数和初始聚类中心。实验结果表明,所提 CODHD 算法与预处理阶段的聚类优化(COPS)算法相比,聚类准确度提高了 30%,聚类算法效率至少提高 14.24%。所提算法具有较强的可行性和实用性。

**关键词:**聚类算法;层次划分;最佳聚类数;初始聚类中心;聚类有效性指标

**中图分类号:** TP301.6    **文献标志码:**A

### Optimized clustering algorithm based on density of hierarchical division

PANG Lin<sup>1,2</sup>, LIU Fang'ai<sup>1,2\*</sup>

(1. College of Information Science and Engineering, Shandong Normal University, Jinan Shandong 250014, China;

2. Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan Shandong 250014, China)

**Abstract:** The traditional clustering algorithms cluster the dataset repeatedly, and have poor computational efficiency on large datasets. In order to solve the problem, a novel algorithm based on hierarchy partition was proposed to determine the optimal number of clusters and initial centers of clusters, named Clusters Optimization based on Density of Hierarchical Division (CODHD). Based on hierarchical division, the computational process was studied, which did not need to cluster datasets repeatedly. First of all, all statistical values of clustering features were obtained by scanning dataset. Secondly, the data partitions of different level were generated from bottom-to-up, the density of each partition data point was calculated, and the maximum density point of each partition was taken as the initial center. At the same time, the minimum distance from the center to the higher density data point was calculated, the average of products' sum of the density of the center and the minimum distance was taken as the validity index and a clustering quality curve of different hierarchical division was built incrementally. Finally, the optimal number of clusters and the initial center of clusters were estimated corresponding to the partition of extreme points of curve. The experimental results demonstrate that, compared with Clusters Optimization on Preprocessing Stage (COPS), the proposed CODHD improved clustering accuracy by 30% and clustering algorithm efficiency at least 14.24%. The proposed algorithm has strong feasibility and practicability.

**Key words:** clustering algorithm; hierarchical division; optimal cluster number; initial cluster center; clustering validity index

## 0 引言

聚类分析是数据挖掘研究的重要内容之一,传统的聚类算法<sup>[1]</sup>通常需要用户根据经验或者具备相关领域的背景知识来给定聚类数和初始聚类中心,而初始中心的选择和确定的聚类数将直接影响算法的效率。

现有的聚类算法<sup>[2~9]</sup>是利用多次聚类反复评估有效性指标确定最佳聚类数,即在给定的数据集上,使用不同的参数  $k$  (通常是聚类数) 运行特定的聚类算法,对数据集进行不同的

划分,计算每种划分的有效性指标,最后比较分析各个指标值的大小或者变化情况,符合预定条件的指标值所对应的算法参数  $k$  被认为是最佳聚类数。 $k$ -means<sup>[10~11]</sup>最佳聚类数确定算法从样本几何结构的角度设计了一种新的聚类有效性指标,并在此基础上提出一种新的确定  $k$ -means 最佳聚类数的方法。然而这种方法为了产生最佳聚类数,通常需要对数据集进行  $k_{\max} - k_{\min} + 1$  次聚类,尤其当数据量较大时,算法效率极低;另一方面,不恰当的  $k_{\max}$  和  $k_{\min}$  值设置也会影响计算结果的准确性,传统的优化聚类数的方法如图 1 所示。预处理

收稿日期:2015-11-30;修回日期:2015-12-30。

基金项目:国家自然科学基金资助项目(61572301, 90612003);山东省自然科学基金资助项目(ZR2013FM008)。

作者简介:逢琳(1991—),女,山东青岛人,硕士研究生,CCF 会员,主要研究方向:数据挖掘、大数据分析; 刘方爱(1962—),男,山东青岛人,教授,博士生导师,博士,主要研究方向:无线网络、分布式计算。

阶段的聚类优化(Clusters Optimization on Preprocessing Stage, COPS)<sup>[12]</sup>,首先根据数据集的几何结构一次性地生成所有合理的划分,然后再评估聚类质量,通过聚类质量极值点对应找到最佳聚类数,不需要反复地运行特定的聚类算法,与特定的聚类算法无关,在很大程度上提高了计算效率和结果的准确性。基于层次划分的最佳聚类数确定方法如图2所示。

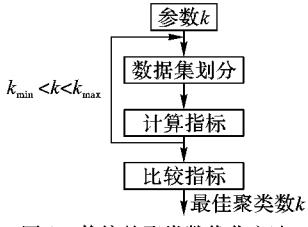


图1 传统的聚类数优化方法

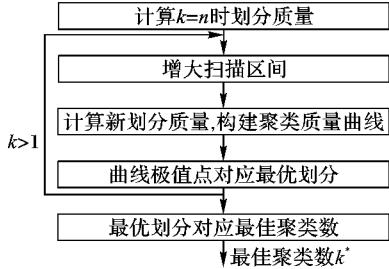


图2 基于层次划分的聚类数优化方法

常见的聚类有效性指标<sup>[13-15]</sup>有:划分系数( Partition Coefficient, PC)、划分熵(Partition Entropy, PE)、CH(Calinski-Harabasz)、DB(Davies-Bouldin)、Sil(Silhouette)和加权的类间类内相似度比率(Weighted inter-intra, Wint),它们是基于数据集几何结构的指标函数,考虑了聚类的基本特征,即一个“好”的聚类结果应使得 $k$ 个聚类内的数据点尽可能“紧凑”,而不同的聚类间的数据点应该尽可能“分离”,而以上的指标量化聚类内紧凑度和聚类间分离度并组合二者生成聚类有效性指标,通过有效性指标的极值点选择最佳聚类数目。在这个过程中,算法复杂度直接受到聚类过程的影响,尤其面对大数据量时,计算复杂度非常高。而基于层次划分的方法,可以很好地处理复杂性数据,并且计算复杂度低。

本文通过研究COPS算法的计算过程,在文献[12,16]的基础上对聚类算法的过程以及有效性指标进行改进,提出一种基于层次划分密度的聚类优化(Clusters Optimization based on Density of Hierarchical Division, CODHD)。首先通过顺序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;然后计算每个划分中数据点的密度,将每个划分中密度最大的数据点确定为中心点<sup>[16]</sup>,计算此中心点距离更高密度点的最小距离,计算每次划分的聚类质量;最后增量地构建一条关于不同层次划分的聚类质量曲线,曲线极大值点所对应的划分用于估计最佳的聚类数目和每个划分的初始聚类中心。本文算法避免了对大型数据集的反复聚类,并且不依赖特定的聚类算法,能够有效识别数据集中可能包含的孤立点,并且采用 $k$ -means基本算法进行聚类分析时,在给定初始聚类中心的基础上,计算效率得到很大程度的提高。

## 1 相关描述与问题定义

给定数据集 $X = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^{d \times n}$ ,其中: $d$ 为维度, $n$ 为样本个数。一个硬划分聚类算法将数据集划分为

$k(k > 1)$ 个子集的集合 $U^k = \{U_1, U_2, \dots, U_k\}$ 。

通过下列定义给出确定 $X$ 划分的计算过程<sup>[12]</sup>。

**定义1** 点的维度相似性。给定一个阈值 $s_i \geq 0, 0 \leq i \leq d$ ,若 $|x_i - y_i| \leq s_i$ ,则 $x$ 和 $y$ 关于 $s_i$ 第*i*维相似,其中 $x_i$ 和 $y_i$ 表示 $x$ 和 $y$ 的第*i*维属性值。

**定义2** 相似度。即给定一个阈值向量 $S = (s_1, s_2, \dots, s_d)$ ,如果点 $x$ 和 $y$ 在所有数据维度上都关于 $s_i$ ( $i = 1, 2, \dots, d$ )相似,则称数据点 $x$ 和 $y$ 是关于 $S$ 相似的。

相似的数据点组成 $X$ 的簇,如果 $S$ 的各个分量相同,则 $\|S\| = \sqrt{s_1^2 + s_2^2 + \dots + s_d^2}$ 相当于基于密度的聚类算法中点的邻域半径,但是,为了反映数据集各维度属性值分布的差异性,允许 $S$ 的各个分量不相同。

**定义3**  $S$ 的比较。即给定两个阈值向量 $S^a = (s_1^a, s_2^a, \dots, s_d^a)$ 和 $S^b = (s_1^b, s_2^b, \dots, s_d^b)$ ,如果满足以下条件,则称 $S^a > S^b$ 。

1)  $s_i^a \geq s_i^b, i = 1, 2, \dots, d$ ;

2) 至少存在一个 $i \in [1, d]$ 使得 $s_i^a > s_i^b$ 。

**定义4 扩展长度**  $\Delta_i = 0.01 \times \max\{\lambda_1, \lambda_2, \dots, \lambda_d\}/\lambda_i, \lambda_i = \sqrt{\sum_{j=1}^n (x_{ji} - u_i)^2/(n-1)}$ ,其中: $x_{ji}$ 是 $x_j$ 第*i*维属性的值, $u_i$ 是第*i*维属性的平均值。

在具有较高维度的实际应用数据中,数据集各维度属性值分布的差异性是常见的,因此 $S$ 的各个分量不相同。简化查找相似点的过程。首先按照每个维度*i*的属性值的大小进行排序,生成序列 $A_i$ ,通过顺序扫描 $A_i$ 可以得到所有与点 $x$ 第*i*维相似的点 $y$ ,扫描范围局限在 $|x_i - y_i| \leq s_i$ 条件的有限区间内,当 $s_i$ 增加 $\Delta_i$ 时,只需在原有范围的基础上扩展扫描区间 $s_i < |x_i - y_i| < s_i + \Delta_i$ 即可。

计算从 $S = S^0$ (每个分量值为0)开始,每个步骤 $S$ 增大一个量 $\Delta(\Delta = \{\Delta_1, \Delta_2, \dots, \Delta_d\})$ ,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,生成新的划分;计算每个划分的 $Q$ 值,直到 $S$ 增长到所有数据点被划分到同一个集合为止。

## 2 CODHD 算法

CODHD算法是基于层次划分的对聚类数据预处理的改进算法,CODHD算法采用新的聚类有效性指标,在算法过程中不仅获得最佳聚类数,而且确定了初始聚类中心。

### 2.1 算法有效性指标

CODHD算法虽然对计算过程进行优化,但是仍然需要定义有效性指标来判断聚类效果,本文的算法仍使用有效性指标 $Q(U)$ 来评估 $X$ 被划分的聚类质量。

CODHD算法的相关有效性指标<sup>[16]</sup>定义如下。

**定义5** 聚类中心。一个聚类中密度最大的数据点,即一个聚类中被最多数据点包围的中心点。密度的计算公式为 $N_x = \sum_y \rho(D_{xy} - D)$ ,其中: $D$ 为截断距离。 $\tau = D_{xy} - D$ ,若 $\tau < 0, \rho(\tau) = 1$ ;相反, $\rho(\tau) = 0$ 。在本文的算法中, $D$ 的选取通常使数据集中大约有1~2%的数据点在邻域内。

**定义6** 最小距离。即聚类中心 $x$ 到比它更高密度点 $y$ 的最小距离 $\delta_x = \min_{y: N_y > N_x} D_{xy}$ 。当 $x$ 为最高密度点时, $\delta_x =$

$\max_y(D_{xy})$ 。根据数据点的几何分布,最小距离即类间分离度,最小距离大即聚类中心远离其他聚类,聚类效果好。

**定义 7 聚类有效性指标。**  $Q = \sum_{m=1}^k N_x \delta_x / k$ , 即聚类有效性指标  $Q$  是聚类中心点密度与此点到更高密度点最小距离乘积之和的平均值,最优聚类质量对应的聚类数  $k$  为最佳聚类数,对应的各个聚类中心点为初始中心点。

## 2.2 算法描述

1) 设定初始值  $k = n$ ,  $S = S^0$ ,  $U^n = \{x_1, x_2, \dots, x_n\}$ ,  $Q^n = Q(U^n)$ 。

2) 将  $n$  个数据点按照第  $i$  维属性排序,生成序列  $A_i$ 。

3) 从  $k = n$  循环至  $k = 1$ ;

① 扫描满足  $|x_i - y_i| < s_i$  条件的有限区间,寻找序列  $A_i$  中数据点  $x$  的第  $i$  维相似点  $y$ ;

② 生成新划分;

③ 计算密度,确定聚类中心点;

④ 计算最小距离和聚类有效性指标;

⑤ 画聚类质量曲线。

4) 提取聚类质量曲线极大值点,此时对应的划分为最佳划分,根据剪枝<sup>[17-18]</sup>原理,此时的  $k$  为最佳聚类数,聚类中心为初始聚类中心。

## 2.3 算法示例

图 3 给出本算法在 2 维数据集上的例子。首先,所有数据点均构成独立的簇如图 3(a)所示;然后,随着  $S$  的增大,数据点被逐渐合并,假设在某个步骤形成了如图 3(b)所示的椭圆形区域所代表的若干小簇;当  $S$  进一步增大时,原本分别属于两个小簇的数据点被合并,基于这样的策略可以生成任意形状的簇,如图 3(c)所示(实际的聚类过程可能不止这三个层次,但作为例子,这里只给出三层)。在这个过程中生成两个聚类,最下面的点是孤立点,示例表明此算法可以有效地识别包含的噪声点。

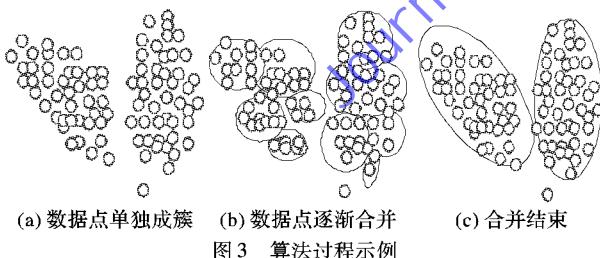


图 3 算法过程示例

## 2.4 最佳聚类数和初始聚类中心的确定

根据聚类质量曲线,曲线最大值对应的划分为最优划分,  
 $U^* = \arg \max_{U^* \in \{U^1, U^2, \dots, U^n\}} Q(U^k)$ , 此时, 根据剪枝原理  
 $initial\_centers = \theta(U^*)$ ,  $k^* = \phi(U^*)$ , 最优划分对应的中心点为初始聚类中心点,对应的聚类数为最佳聚类数。

## 3 实验结果与分析

本文实验采用 Matlab 开发环境编程实现,在 CPU 3.1 GHz, RAM 4 GB 的 Windows 7 操作系统计算机上运行通过。为验证本文提出的 CODHD 算法的有效性,在加州大学尔湾分校(University of California Irvine, UCI)真实数据集上进行仿真实验,并进行不同算法的对比仿真实验。

### 3.1 数据集

本文数据集选自 UCI 真实数据集,分别为 magic04、Normal7、abalone、dna、air、mssplice、diabetes、sonar、blood、australian,真实数据集各个参数特征如表 1 所示。

### 3.2 算法有效性

CODHD 算法和 COPS 算法在层次式的簇合并过程中计算不同划分的聚类质量,合并过程中产生的簇的数目取决于数据集本身的结构,这正是两个算法在识别复杂形状聚类方面的优势。但本文的 CODHD 算法与 COPS 方法不同之处在于本文的算法计算每个划分的聚类中心,并且将聚类中心的密度值与中心点距离更高密度点的最小距离乘积之和的平均值作为聚类质量评估指标,在这个过程中,聚类中心点密度越大,该点周围数据点越多,聚类内数据点越紧凑;与其他聚类最小距离越大,此聚类越独立,聚类之间越分离。

表 1 数据集参数

数据集	维数	数据个数	聚类数
magic04	10	19 020	2
Normal7	2	7 000	7
abalone	7	4 177	23
dna	180	2 000	3
air	64	359	3
mssplice	240	3 175	3
diabetes	8	768	2
sonar	60	208	2
blood	4	748	2
australian	14	690	2

基于层次划分的改进算法 CODHD 在 magic04、Normal7、abalone、dna 这 4 个数据集都得到正确的聚类数,实验结果如图 4 所示。

对于 magic04 和 Normal7, CODHD 检测到聚类数趋于最优数目时,聚类质量逐渐增加。受噪声影响,abalone 中有两个边界模糊的簇,如图 4(c)显示,对应于聚类数 23 和 24 的聚类质量只存在很小的差异。对以上 4 个较为复杂的数据集,CODHD 没有检测到连续变化的  $k$  值,例如在图 4(d)中,  $k$  从 8 跳变到最优聚类数 3。这是因为 CODHD 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,CODHD 完成准确的区分,表明 CODHD 可以有效地识别噪声并区分簇间的密度差异。

几种有效性指标估计出的数据集最佳聚类数情况如表 2 所示。

由表 2 可以看出,对于聚类明显分离的数据集 Normal7,所有方法都得到正确的最优聚类数 7。对于复杂数据集 magic04,只有 CODHD 算法和 COPS 算法得到正确的最优聚类数 2。COPS 和 CODHD 能识别带有噪声的数据集 abalone,其他有效性指标对孤立点无法有效识别。然而对高维数据集 ms splice,只有 CODHD 算法得到正确最优聚类数 3,这在一定程度上验证了 CODHD 算法正确处理高维数据的能力。经过多次实验,CODHD 在 10 个数据集上得到 9 个正确的最优聚类数,COPS 算法得到 6 个正确的最优聚类数,实验结果表明 CODHD 算法比 COPS 算法具有更好的聚类有效性和稳定性。

通过反复聚类,计算有效性指标 CH、DB,根据聚类质量曲线的极值点确定最佳聚类数的方法,在多次实验后没有得到更好的结果。

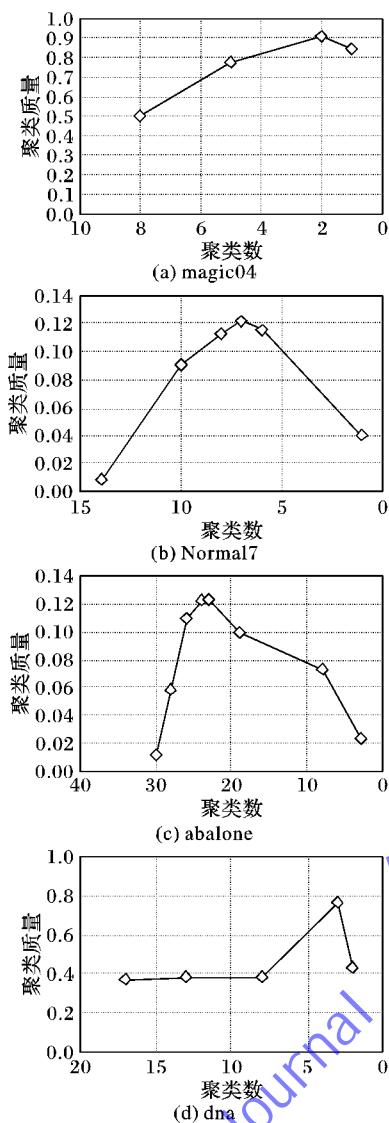


图4 CODHD 在4个数据集上的实验结果

表2 不同方法得到的最佳聚类数

数据集	正确值	计算值			
		CODHD	COPS	CH	DB
magic04	2	2	3	3	4
Normal7	7	7	7	7	7
abalone	23	23	23	5	24
dna	3	3	3	4	4
air	3	3	3	4	3
mssplice	3	4	4	5	5
diabetes	2	2	2	2	3
sonar	2	2	3	3	3
blood	2	2	2	2	2
australian	2	2	3	2	3

### 3.3 算法效率

为了验证算法效率,进行如下两种操作方法:方法一,运行 CODHD 算法,通过剪枝原理得到最佳聚类数和初始聚类中心,继续运行  $k$ -means 算法;方法二,运行 COPS 算法,通过

剪枝原理得到最佳聚类数,继续运行  $k$ -means 算法。两种算法以及在此基础上的  $k$ -means 算法总运行时间对比如表3。实验表明,在适当的参数配置下,即指定了正确的聚类数和初始聚类中心, $k$ -means 算法可以很好地区分出这 10 个数据集,并在一定程度上提高计算效率。

表3 结果表明:CODHD 算法大幅度提高了聚类算法的效率,在高维数据集 dna、mssplice 上尤为明显,在 dna 数据集上,两种方法运行时间分别为 65.6 s 和 76.5 s,方法一比方法二缩短时间超过 10 s;在 msplice 数据集上,两种方法运行时间分别为 109.1 s 和 140.5 s,方法一比方法二缩短时间超过 31 s,在高维数据集聚类数确定的同时,初始聚类中心的确定大大降低中心点选取的计算量。在大型数据集 magic04、Normal7 和 abalone 上,方法一和方法二的运行时间分别为 10.8 s 和 17.8 s、3.4 s 和 4.8 s、1.6 s 和 5.1 s,对同样的大型数据集,已知聚类数和初始聚类中心的前提下, $k$ -means 聚类避免对初始聚类中心的迭代查找,降低  $k$ -means 算法复杂度,因此方法二运行聚类时间也得到明显提高。两种优化算法在最佳聚类数一致的情况下,CODHD 算法基础上的聚类运行时间低于 COPS 算法基础上的聚类运行时间,由此看出,本文 CODHD 算法在一定程度上提高了聚类算法运行效率,并能达到较好的聚类效果,由此证明本文算法具有较强的可行性和实用性。

表3 两种方法的运行时间比较

数据集	CODHD/s	COPS/s
magic04	10.8	17.8
Normal7	3.4	4.8
abalone	1.6	5.1
dna	65.6	76.5
air	1.8	3.3
mssplice	109.1	140.5
diabetes	0.3	0.9
sonar	0.5	1.9
blood	0.3	0.4
australian	0.4	1.4

### 4 结语

本文主要针对的是聚类数据集预处理算法的研究,针对已有算法依赖于特定的聚类算法,且在聚类准确度和聚类算法效率不高方面的问题,提出了 CODHD 算法,通过理论分析和实验结果验证了该算法的可行性和有效性。目前聚类算法优化的理论研究逐渐成熟,但是本文扩展长度  $\Delta$  的选取与某维度属性值的平均值有关,在数据集中容易受孤立点的影响,在今后的工作中,研究扩展长度的取值是必要的。

#### 参考文献:

- [1] BERKHIN P. A survey of clustering data mining techniques [M]// Grouping Multidimensional Data. Berlin: Springer, 2006: 25–71.
- [2] TIBSHIRANI R, WALTHER G, HASTIE T. Estimating the number of clusters in a data set via the Gap statistic [J]. Journal of the Royal Statistical Society, 2000, 63(2): 411–423.
- [3] 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[J].模糊系统与数学,2001,15(1):89–92.(SUN C Z, WANG J D, PAN J. Research on the method of determining the optimal class number of fuzzy cluster [J]. Fuzzy Systems and Mathematics,

- 2001, 15(1): 89–92.)
- [4] 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.
- [5] HALKIDI M, BATISTAKIS Y, VAZIRGIANNIS M. Clustering validity checking methods: part II [J]. *ACM SIGMOD Record*, 2002, 31(3): 19–27.
- [6] 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[J]. *电子与信息学报*, 2002, 24(8): 1017–1021. (FAN J L, WU C M. Clustering validity function based on probabilistic partition coefficient combined with fuzzy variation [J]. *Journal of Electronics and Information Technology*, 2002, 24(8): 1017–1021.)
- [7] YU J, CHENG G. Search range of the optimal number of clusters in fuzzy clustering [J]. *Science in China (Series E)*, 2002, 32(2): 274–280.
- [8] SUN H, WANG S, JIANG Q. FCM-based model selection algorithms for determining the number of clusters [J]. *Pattern Recognition*, 2004, 37(10): 2027–2037.
- [9] BOUGUESSA M, WANG S, SUN H. An objective approach to cluster validation [J]. *Pattern Recognition Letters*, 2006, 27(13): 1419–1430.
- [10] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. *软件学报*, 2008, 19(1): 48–61. (SUN J G, LIU J, ZHAO L Y. Clustering algorithms research [J]. *Journal of Software*, 2008, 19(1): 48–61.)
- [11] CELEBI M E, KINGRAVI H A, VELA P A. A comparative study of efficient initialization methods for the  $k$ -means clustering algorithm [J]. *Expert Systems with Applications*, 2013, 40(1): 200–210.
- [12] 陈黎飞, 姜青山, 王声瑞. 基于层次划分的最佳聚类数确定方法
- [J]. *软件学报*, 2008, 19(1): 62–72. (CHEN L F, JIANG Q S, WANG S R. A hierarchical method for determining the number of clusters [J]. *Journal of Software*, 2008, 19(1): 62–72.)
- [13] PAKHIRIA M K, BANDYOPADHYAY S, MAULIK U. A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification [J]. *Fuzzy Sets and Systems*, 2005, 155(2): 191–214.
- [14] WANG W, ZHANG Y. On fuzzy cluster validity indices [J]. *Fuzzy Sets and Systems*, 2007, 158(19): 2095–2117.
- [15] REZAEE B. A cluster validity index for fuzzy clustering [J]. *Fuzzy Sets and Systems*, 2010, 161(23): 3014–3025.
- [16] ALEX R, ALESSANDRO L. Machine learning. Clustering by fast search and find of density peaks [J]. *Science*, 2014, 344(6191): 1492–1496.
- [17] AGRAWAL R, GEHRKE J, GUNOPULOS D, et al. Automatic subspace clustering of high dimensional data [J]. *Data Mining & Knowledge Discovery*, 2005, 11(1): 5–33.
- [18] MEDEIROS C M S, BARRETO G A. A novel weight pruning method for MLP classifiers based on the MAXCORE principle [J]. *Neural Computing & Applications*, 2013, 22(1): 71–84.

#### Background

This work is partially supported by the National Natural Science Foundation of China (61572301, 90612003), the Shandong Provincial Natural Science Foundation (ZR2013FM008).

**PANG Lin**, born in 1991, M. S. candidate. Her research interests include data mining, big data analysis.

**LIU Fang'ai**, born in 1962, Ph. D., professor. His research interests include wireless network, distributed computation.

(上接第 1633 页)

- [4] 张明亮, 吴俊, 李凡长. 五子棋机器博弈系统评估函数的设计 [J]. *计算机应用*, 2012, 32(7): 1969–1972, 1990. (ZHANG M L, WU J, LI F C. Design of evaluation-function for computer gobang game system [J]. *Journal of Computer Applications*, 2012, 32(7): 1969–1972, 1990.)
- [5] 朱全民, 陈松乔. 五子棋算法的研究与思考 [J]. *计算技术与自动化*, 2006, 25(2): 71–74. (ZHU Q M, CHEN S Q. Gobang algorithm research and think [J]. *Computing Technology and Automation*, 2006, 25(2): 71–74.)
- [6] 张小川, 候鑫磊, 涂飞. 博弈机器人的行为规划 [J]. *重庆理工大学学报(自然科学版)*, 2014, 28(4): 99–103. (ZHANG X C, HOU X L, TU F. Behavior planning for the game robot [J]. *Journal of Chongqing University of Technology (Natural Science Edition)*, 2014, 28(4): 99–103.)
- [7] 蒋加伏, 陈蒿祥, 唐贤英. 基于知识推理的博弈树搜索算法 [J]. *计算机工程与应用*, 2004, 40(1): 74–76. (JIANG J F, CHEN A X, TANG X Y. Search algorithm for game of checkers based on knowledge inference [J]. *Computer Engineering and Applications*, 2004, 40(1): 74–76.)
- [8] 杨云强, 吴姣. 一种改进的基于博弈树模型的五子棋系统 [J]. *科学技术与工程*, 2012, 12(5): 1052–1055, 1060. (YANG Y Q, WU J. An Improved Gobang system based on game-playing tree [J]. *Science Technology and Engineering*, 2012, 12(5): 1052–1055, 1060.)
- [9] 张海峰, 白振兴, 张登福. 五子棋中的博弈智能设计 [J]. *现代电子技术*, 2004, 27(7): 25–27. (ZHANG H F, BAI Z X, ZHANG D F. Design of playgame intelligence in Gobang [J]. *Modern Electronics Technique*, 2004, 27(7): 25–27.)
- [10] 金元郁, 李新, 刘国建. 基于图像处理的人和机械手象棋对弈系统实现 [J]. *青岛科技大学学报(自然科学版)*, 2007, 28(1): 73–75, 93. (JIN Y Y, LI X, LIU G J. Development of man and manipulator's chess-playing system based on image processing [J]. *Journal of Qingdao University of Science and Technology (Natural Science Edition)*, 2007, 28(1): 73–75, 93.)
- [11] 黄立波, 夏庭锴, 王春香, 等. 实时环境下的对弈机器人控制系统设计与分析 [J]. *机械*, 2004, 31(6): 50–52. (HUANG L B, XIA T K, WANG C X, et al. Design and analysis of the Chinese-chess robot in real time environment [J]. *Machinery*, 2004, 31(6): 50–52.)
- [12] 吕艳辉, 宫瑞敏. 计算机博弈中估值算法与博弈训练的研究 [J]. *计算机工程*, 2012, 38(11): 163–166. (LYU Y H, GONG R M. Study on valuation algorithm and game training in computer game [J]. *Computer Engineering*, 2012, 38(11): 163–166.)

#### Background

This work is partially supported by the Suzhou Science and Technology Project (SYG201504).

**MAO Limin**, born in 1981, M. S., lecturer. His research interests include robot control, target tracking.

**ZHU Peiyi**, born in 1980, Ph. D., lecturer. His research interests include intelligent control, data fusion.

**LU Zhenli**, born in 1974, Ph. D., lecturer. His research interests include intelligent control, machine vision.

**PENG Weiwei**, born in 1993, engineer. His research interests include data acquisition, signal processing.