DOI：10.13718/j.cnki.xdzk.2020.11.007

基于多角度空间结构的超多类簇聚类方法

史欣蕊[1,3]，　钱宇华[1,2,3]，　李飞江[1,3]

1. 山西大学 大数据科学与产业研究院，太原 030006；
2. 山西大学 计算智能与中文信息处理教育部重点实验室，太原 030006；
3. 山西大学 计算机与信息技术学院，太原 030006

摘要：为应对超多类簇聚类问题，提出了一个多角度空间结构的多类簇聚类方法 MS^{2}BC，基于空间结构表示方法与 bagging 特征抽样技术从多个角度构建数据的空间结构并进行集成，然后利用集成后空间结构表示完成聚类．在 10 个真实数据上的聚类实验验证了 MS^{2}BC 方法的有效性．

关　键　词：聚类分析；多类簇；空间结构；聚类集成；数据挖掘

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673－9868(2020)11－0059－09

无监督学习[1]是机器学习中的一项重要技术，其中聚类分析与应用[2-30]在近年来得到了广泛关注．

传统的聚类方法大致可以分为：基于划分的聚类算法[9-12]、基于模糊理论的聚类算法[16-20]、基于密度的聚类算法[21-23]等．除此之外，受传统聚类算法的启发，为完成更广泛的任务，越来越多的新颖的聚类算法被提出[24-30]．但以上算法在解决超多类聚类问题时却并不理想[31-33]．

文献[34]提出了空间结构表示框架，可有效应对符号型数据空间结构模糊的问题，并应用于符号数据聚类问题中[35-36]．基于空间结构的表示方法将原始符号型数据映射到一个概率表示空间，在保持原有类结构信息的前提下，该方法提供了更加丰富的测度信息，从而使原始符号数据的类结构信息更加清晰．借助该思想，本文提出了一种多角度空间结构的数据聚类算法，从多个角度构建数据的空间结构，以期更全面地识别数据空间中存在的类结构，从而应对多类簇数据聚类问题．该方法利用特征抽样从多个视角刻画原始数据集，然后构建不同视角下的空间结构表示，再集成这些不同视角下的数据表示得到一个统一的表示矩阵，最后利用该矩阵完成聚类．

本文的主要贡献包括 3 个方面：

1）多角度空间结构表示方法．本文提出了一种从多个角度对原始数据集进行空间结构表示的方法，以更加准确地识别复杂的类簇分布结构；

2）多角度空间结构聚类方法．本文提出多角度空间结构聚类算法框架，集成多个视角所形成的空间结构来对数据进行更有效地聚类；

3）本文在 10 个真实数据集上验证了多角度空间结构聚类算法相较于传统聚类算法的优越性．

1　空间结构表示方法

本章内容将分别介绍符号型数据和数值型数据的空间结构表示方法．

收稿日期：2020－10－14
基金项目：国家自然科学基金项目(61672332)；山西省重点研发计划(国际科技合作) 项目(201903D421003)．
作者简介：史欣蕊(1994－)，女，硕士研究生，主要从事机器学习的研究．

1.1 符号型数据的空间结构表示方法

符号型数据是由一组有限和无序的特征向量来表示，所以无法像数值型数据那样直接度量样本间的相似度或距离，且难以准确刻画符号型数据的空间分布结构，这导致许多聚类算法无法处理符号型数据.

为更清晰地刻画符号型数据的空间结构，文献[34]基于样本间相似性概率提出了一种空间结构表示方法，具体如下：

假设 $U=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ 为数据集合，$A=\{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m\}$ 为特征集合，如果 A 中的特征均为符号型特征，则数据集 U 为符号型数据集.

假设 $T=(U, A)$ 为符号数据集，其中 U 为样本集，A 为属性集，表1给出了一个表示示例.

表1 符号数据表示示例

	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\cdots$	$\boldsymbol{a}_m$
$\boldsymbol{x}_1$	$\boldsymbol{a}_1(\boldsymbol{x}_1)$	$\boldsymbol{a}_2(\boldsymbol{x}_1)$	$\cdots$	$\boldsymbol{a}_m(\boldsymbol{x}_1)$
$\boldsymbol{x}_2$	$\boldsymbol{a}_1(\boldsymbol{x}_2)$	$\boldsymbol{a}_2(\boldsymbol{x}_2)$	$\cdots$	$\boldsymbol{a}_m(\boldsymbol{x}_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\boldsymbol{x}_n$	$\boldsymbol{a}_1(\boldsymbol{x}_n)$	$\boldsymbol{a}_2(\boldsymbol{x}_n)$	$\cdots$	$\boldsymbol{a}_m(\boldsymbol{x}_n)$

样本 $\boldsymbol{x}_i$ 和样本 $\boldsymbol{x}_j$ 的相似性概率为：

$$p_{ij}=\frac{1}{m}\sum_{i=1}^{m}\theta_i(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{1}$$

其中：$\boldsymbol{\alpha}_l(\boldsymbol{x})$ 为数据 $\boldsymbol{x}$ 的第 l 个特征值，

$$\theta_l(\boldsymbol{x}_i, \boldsymbol{x}_j)=\begin{cases}1 & \boldsymbol{a}_l(\boldsymbol{x}_i)=\boldsymbol{a}_l(\boldsymbol{x}_j)\\ 0 & \boldsymbol{a}_l(\boldsymbol{x}_i)\neq\boldsymbol{a}_l(\boldsymbol{x}_j)\end{cases} \tag{2}$$

通过计算两两样本间的相似性概率，可得到符号型数据的空间结构表示矩阵为：$\boldsymbol{S}_C=[p_{ij}]_{n\times n}$，在符号型数据的空间结构中，一个样本的特征为$\{b_i=\boldsymbol{x}_i, 1\leqslant i\leqslant n\}$，表2给出了符号型数据的空间表示示例.

表2 符号型数据的空间表示矩阵

	$\boldsymbol{x}_1(b_1)$	$\boldsymbol{x}_2(b_2)$	$\cdots$	$\boldsymbol{x}_n(b_n)$
$\boldsymbol{x}_1$	p_{11}	p_{12}	$\cdots$	p_{1n}
$\boldsymbol{x}_2$	p_{21}	p_{22}	$\cdots$	p_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\boldsymbol{x}_n$	p_{n1}	p_{n2}	$\cdots$	p_{nn}

空间结构表示矩阵可以将符号型数据映射到一个欧式空间.

1.2 数值型数据的空间结构表示方法

虽然空间结构表示方法最初是针对符号型数据提出，但该方法很容易扩展至处理数值型数据任务.

对于数值型特征 $\boldsymbol{a}_l$，样本 $\boldsymbol{x}_i$ 和样本 $\boldsymbol{x}_j$ 相似的概率为：

$$p_{\boldsymbol{a}^l}(\boldsymbol{x}_i, \boldsymbol{x}_j)=1-\frac{|\boldsymbol{a}_l(\boldsymbol{x}_i)-\boldsymbol{a}_l(\boldsymbol{x}_j)|}{\max(\boldsymbol{a}_l)-\min(\boldsymbol{a}_l)} \tag{3}$$

其中 $\max(\boldsymbol{a}_l)=\max\limits_{1\leqslant i\leqslant n}\{\boldsymbol{a}_l(\boldsymbol{x}_i)\}$，$\min(\boldsymbol{a}_l)=\min\limits_{1\leqslant i\leqslant n}\{\boldsymbol{a}_l(\boldsymbol{a}_i)\}$. 基于公式(3)，样本 $\boldsymbol{x}_i$ 和样本 $\boldsymbol{x}_j$ 相似的概率为：

$$p(\boldsymbol{x}_i, \boldsymbol{x}_j)=\frac{\sum\limits_{l}^{m}p_{\boldsymbol{a}_l}(\boldsymbol{x}_i, \boldsymbol{x}_j)}{m} \tag{4}$$

通过计算两两样本间的相似性概率，可得到数值型数据的空间结构表示矩阵为 $\boldsymbol{S}_N=[p_{ij}]_{n\times n}$.

本文借助上述空间结构表示方法的思想，提出了一种多角度空间结构聚类算法，通过从多个角度构建原始数据集的空间结构，使得类簇结构信息更加清晰准确，进而解决数值型数据的超多类聚类问题.

2　多角度空间结构

2.1　多角度数据表示

为从多个角度刻画原始数据集，本文采用装袋(Bagging)算法. 装袋算法被广泛地应用于监督学习任务中，通过有放回地抽样数据集，训练多个模型，这样可以降低模型泛化误差，其采用的策略为模型平均，即使用训练好的若干学习器来对新的未知样本预测. 这种算法是一种集成方法，通过集成多个弱学习器来提高模型的学习性能.

在这种集成思想的启发下，本文通过对原始数据集的特征进行有放回地抽样，以从多个不同的视角来刻画原始数据. 这样可以从不同的角度来描述一个样本，以期为后续揭示更清晰的类簇分布信息奠定基础. 令 $U=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ 为数据集合，$A=\{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m\}$ 为特征集合，通过对特征集 A 进行有放回采样可得到新视角下的特征集描述 $B=\{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_m\}$ 以及新视角下的数据集合 $U'=\{\boldsymbol{x}_1', \boldsymbol{x}_2', \cdots, \boldsymbol{x}_n'\}$.

本文所提算法每次随机有放回地抽取与原始数据集相同维数的特征，形成新视角下的数据集. 为进一步提升特征对原始数据结构的表达能力，本文在每次特征抽取完成之后，均对新的数据进行特征提取. 另外，当原始数据集的特征较多时，每次形成的数据也具有较高的维度，会导致存储空间的浪费以及降低计算效率. 因此，特征提取在减少存储空间的同时提高了计算效率.

本文所提算法采用 PCA 降维技术来对数据集进行降维处理. PCA 是一种被广泛使用的数据降维算法. 其主要思想是将数据从 n 维输入空间映射到 k 维特征空间. 即将每个 n 维数据点通过映射转换成另一个数据点. 其工作原理是，从原始数据空间中依次找出相互正交的坐标方向. 第一个坐标方向选择数据集中方差最大的维度所在的方向，第二个坐标方向选择与第一个方向呈正交的平面中与方差最大的坐标方向，第三个坐标方向选择与前两个坐标方向呈正交的平面中方差最大的坐标方向，依次类推. 最后会发现大部分数据基本集中在 k 个坐标方向中，所以后面的坐标方向可直接忽略，从而达到降维目的. 对于数据矩阵 $\boldsymbol{U}'_{n\times d}$，PCA 降维流程如下：

1) 求 $\boldsymbol{U}'_{n\times d}$ 的协方差矩阵 $\boldsymbol{C}_{n\times n}=\mathrm{COV}(\boldsymbol{U}')$；

2) 求解协方差矩阵的特征值和特征向量；

3) 选取最大的 k 个特征值所对应的特征向量组成的矩阵 $\boldsymbol{P}_{d\times k}$；

4) 计算 $\boldsymbol{B}_{n\times k}=\boldsymbol{X}_{n\times d}\boldsymbol{P}_{d\times k}$；

将 $\boldsymbol{B}$ 视为一个视角下的数据集. 通过多次对特征的重抽样和提取，可得到多个视角下的数据集描述.

2.2　多角度空间结构融合

为了可以提供更清晰的类簇分布结构信息，本文结合数据的空间结构表示和多角度表示提出一种多角度空间结构表示方法.

样本间距离度量可以反应更加精细的结构信息，由此本文通过集成多个视角下的空间结构度量信息来构建统一地、更精细的数据表示，同时将多角度数据映射到同一个欧式空间中，便于之后的聚类分析.

首先，定义一个单一视角下的数据空间结构表示矩阵：

假设 $B'=\{\boldsymbol{x}_1', \boldsymbol{x}_2', \cdots, \boldsymbol{x}_n'\}$ 为第 t 个视角下的数据集合，基于公式(1)可求得 B' 的空间结构表示 $\boldsymbol{S}'=[s_{ij}']_{n\times n}$，其中 $s_{ij}'=p(\boldsymbol{x}_i', \boldsymbol{x}_j')$.

由空间结构的构造方法易知，空间结构表示的数据由规范化的特征描述，数据存在于第一象限，且取值范围区间为[0, 1]. 另外，一个数据在多个视角下的空间结构表示的特征维数相同，均为样本个数 n. 因此，可通过对应位置求平均的方法对一个数据的多个空间结构表示进行融合，即：

$$\boldsymbol{S}_{\boldsymbol{P}}=\frac{1}{m}\sum_{t=1}^{m}\boldsymbol{S}^t \tag{5}$$

至此完成了原始数据的多角度空间结构的表示.

3 多角度空间结构

基于数据的多角度空间结构表示方法，提出了一个多角度空间结构聚类算法 MS^{2}BC(multi-view space structure-based clustering). 该方法一方面借助了空间结构表示方法在数据类结构清晰化上的优势，另一方面从多个角度刻画数据的空间结构. 因此，当类簇数量增多时，该方法有望缓解直接在原始数据上聚类难以识别类簇结构的问题，进而提升聚类性能. 具体步骤如下：

1) 对特征进行与原始特征空间维数相等的装袋算法抽样，得到映射到多个视角的数据集合；

2) 对每个视角下的数据集合进行空间结构表示，再集成所有的结构表示信息；

3) 对集成后的空间结构表示进行聚类，获得数据集合的类簇划分结果.

对于步骤 3)，可以采用不同的聚类方法对空间结构表示进行聚类. 本文采用的是谱聚类算法.

本小节给出本文所提算法的整体算法流程，伪代码描述如下：

算法　多角度空间结构聚类算法(MS^{2}BC)

输入：数据集 D、聚类数目 k、特征采样次数 m

输出：聚类结果 π

(1) begin

(2) 置 $t=0$，置 $\boldsymbol{S}$ 为 0 矩阵

(3) for $t<m$ do

(4) 对数据集 D 执行特征装袋算法抽样得到 D'

(5) 对 D' 进行 PCA 降维得到 B'

(6) 依据公式(5)计算 B' 的空间结构表示矩阵 $\boldsymbol{S}'$

(7) 置 $\boldsymbol{S}=\boldsymbol{S}+\boldsymbol{S}'$

(8) end for

(9) 置 $\boldsymbol{S}_P=\dfrac{\boldsymbol{S}}{m}$

(10) 对 $\boldsymbol{S}_P$ 进行谱聚类得到聚类结果 π

(11) end

算法整体分为 2 个阶段：第一阶段是构造多视角空间结构表示；第二阶段基于该空间结构表示聚类. 从算法框架可以看出，本文算法在第一阶段构造多视角空间结构表示时，循环 m 次，每轮循环将执行特征装袋算法、PCA 降维以及构建空间结构表示矩阵，其中构建空间结构表示矩阵的时间复杂度为 $O(n^2)$. 因此，算法第一阶段的时间复杂度为 $O(m(d+O(PCA)+n^2))$，其中 m 表示循环次数，d 表示数据维数，n 代表样本个数. 算法的第二阶段利用谱聚类对第一阶段得到的空间结构表示进行聚类. 因此第二阶段的时间复杂度为 $O(SC)$. 综上所述，本文所提的算法的时间复杂度为 $O(m(d+O(PCA))+n^2+O(SC))$.

此外，本文所提算法可并行性较强，且易扩展至处理大规模数据. 首先，多角度构建空间结构部分具有天然的可并行性，不同视角下构建空间结构可分布式运行；其次，在构造空间结构时可选取具有代表性的样本进行度量，构造空间结构表示矩阵 $\boldsymbol{S}_{[n\times n']}$(其中 $n'<n$ 为代表性样本数)，从而提升该部分的运行速率；最后，算法第二阶段所采用的谱聚类可借鉴现有快速算法来加速聚类过程，如 U-SPEC[37].

4 实验及结果分析

为验证 MS^{2}BC 算法的有效性，本文对比该方法与 6 个代表性聚类算法在 10 组真实数据上的聚类性能.

4.1　数据集描述

本文实验部分采用 10 个数据集，数据集的详细信息如表 3 所示. 其中 glass，solar，zoo 以及 segment 等 4 个数据集类簇较少，lsolet，ORL，corel_5k，Mpeg7，caltech101 和 Tdt2 等 6 个数据集类簇较多.

表 3　数据集描述

数据集	样例量	数据维数	类别数
glass	214	9	6
solar	323	12	6
zoo	101	16	7
segment	2 310	19	7
lsolet	1 560	617	26
ORL	400	1024	40
Core_5k	5 000	423	50
Mpeg7	1 400	6 000	70
Caltech101	8 641	256	101
Tdt2	10 212	36 771	96

4.2　评价指标

调整兰德指数(adjusted rand index，ARI)和标准互信息(normalized mutual information，NMI)是评价聚类算法好坏的两个常用指标，被广泛使用，两者均为外部指标.

NMI 值定义如下：

$$p_{\mathrm{NMI}}(P,\ Q)=\frac{I(P,\ Q)}{\sqrt{HH}} \tag{6}$$

其中：P，Q 为两种聚类结果；$I(\cdot)$为其互信息；$H(\cdot)$为信息熵.

ARI 值定义如下：

$$p_{ARI}=\frac{C_n^2\sum\limits_{i,\ j}C_{n_{ij}}^2-\sum\limits_i C_{b_i}^2\sum\limits_j C_{d_j}^2}{\frac{1}{2}C_n^2\left(\sum\limits_i C_{b_i}^2+\sum\limits_j C_{d_j}^2\right)-\sum\limits_i C_{b_i}^2\sum\limits_j C_{d_j}^2} \tag{7}$$

其中：n_{ij} 表示在分类结果 X，Y 中，X_i 类与 Y_j 类相同的对象个数；b_i 表示 X_i 类的对象个数；d_j 表示 Y_j 类的对象个数.

两者取值在 0～1 之间，聚类结果越接近数据集的真实分布，NMI 和 ARI 指标越高. 本文采用以上两个指标衡量所提算法的有效性.

4.3　实验结果及其分析

本文采用的对比方法包括 K-means、AP、HC、谱聚类(SC)、U-SPEC[37] 以及 AD-AP[38] 等 6 种方法. 表 4 列出了 7 种方法的 NMI 值，表 5 列出了 7 种方法的 ARI 值，其中粗体表示每个数据集上评价指标取得的最高值.

从表 4 可以看出本文所提算法在 8 个数据集上的 NMI 值达到了最高，其中在 zoo，lsolet，ORL 等 3 个数据集上 NMI 值均达到了 0.75 以上，说明本文所提算法的聚类性能优于其它算法. 另外，在数据集 caltech101 上，传统聚类算法的 NMI 值较低，但本文所提算法确取得了不错的聚类结果. 从表 5 可以看出本文所提算法在 8 个数据集上 ARI 值达到了最高，本文所提算法的聚类性能优于其它算法. 另外，在数据集 ORL 上，传统聚类算法的 ARI 值较低，但本文所提算法确取得了不错的聚类结果. 本文所提算法在 8 个数据集上两个指标均达到了最高.

上述分析表明我们的算法相比于传统聚类算法具有竞争优势，同时也反映出我们的算法的确克服了传统聚类算法在类簇数量增多时聚类性能下降的问题.

图1和图2分别表现了在不同数据集上，NMI和ARI随特征抽取次数的增大所发生的变化. 从图中可以清晰看到本文算法较为平稳，且这些折线也反应出当特征抽取次数较小时，聚类性能随抽取次数的增大而增大；当特征抽取次数较大时，聚类性能随抽取次数的增大而下降. 关于如何根据具体规模的数据集选择合适的特征抽取次数将作为本文未来研究工作.

表4 标准互信息结果比较

数据集	K-means	AP	HC	SC	U-SPEC	AD-AP	MS^{2}BC
glass	0.405 7	0.339 2	0.151 7	0.379 4	0.224 6	0.380 8	**0.427 4**
solar	0.370 0	0.357 3	0.294 5	0.305 3	0.264 1	0.382 1	**0.390 9**
zoo	0.733 3	0.691 4	0.740 3	0.752 9	**0.761 3**	0.691 4	0.752 8
segment	0.600 3	0.612 9	0.043 9	0.536 9	0.098 1	**0.625 0**	0.622 8
Isolet	0.756 1	0.724 6	0.694 9	0.783 9	0.793 4	0.724 6	**0.795 1**
ORL	0.737 7	0.757 3	0.744 0	0.799 1	0.829 1	0.761 6	**0.829 9**
corel_5k	0.268 5	0.273 6	0.139 4	0.271 5	0.281 8	0.273 6	**0.299 0**
Mpeg7	0.682 0	0.684 8	0.423 4	0.716 7	0.713 4	0.681 0	**0.736 5**
caltech101	0.530 9	0.500 7	0.398 8	0.089 9	0.529 5	0.499 1	**0.557 7**
Tdt2	0.393 9	0.366 7	0.061 5	0.615 3	0.546 8	0.390 1	**0.681 2**

表5 调整兰德信息结果比较

数据集	K-means	AP	HC	SC	U-SPEC	AD-AP	MS^{2}BC
glass	0.261 4	0.186 2	0.019 8	0.241 5	0.062 0	0.233 7	**0.284 1**
solar	0.290 0	0.281 7	0.189 2	0.217 2	0.174 6	0.298 5	**0.301 7**
zoo	0.624 0	0.480 6	0.615 9	0.541 6	**0.670 6**	0.480 6	0.621 6
segment	0.451 9	0.420 4	0.000 1	0.425 9	0.015 7	**0.565 0**	0.472 9
Isolet	0.516 8	0.482 9	0.212 0	0.614 5	0.560 7	0.482 9	**0.621 3**
ORL	0.357 7	0.394 7	0.276 8	0.496 8	0.538 4	0.417 1	**0.550 9**
corel_5k	0.193 0	0.204 2	0.032 3	0.168 7	0.294 0	0.063 3	**0.356 9**
Mpeg7	0.260 2	0.265 5	0.019 6	0.062 8	0.306 2	0.265 5	**0.406 0**
caltech101	0.216 1	0.165 1	0.095 2	0.001 2	0.172 8	0.158 8	**0.272 9**
Tdt2	0.023 1	0.010 3	0.000 4	0.138 8	0.135 5	0.068 8	**0.151 3**

5 总结和未来工作展望

大数据背景下的聚类任务中，类簇数量剧增，给传统聚类方法带来了巨大挑战. 针对该问题，本文提出了一种多角度空间结构聚类算法，通过集成数据的多个视角空间结构来加强算法识别类簇的能力，进而提升聚类性能. 从实验结果的对比中可以看出，相较于传统聚类算法，本文所提算法取得了较高性能指标. 未来我们重点研究本文所提算法的快速算法和并行算法，从而有效应对大规模超多类数据的聚类问题.

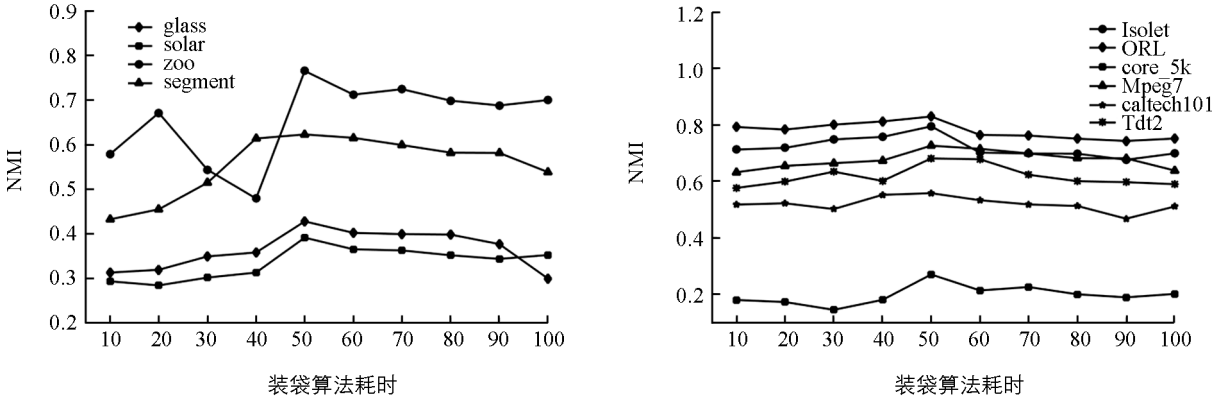

图 1　NMI 与特征抽样次数变化关系

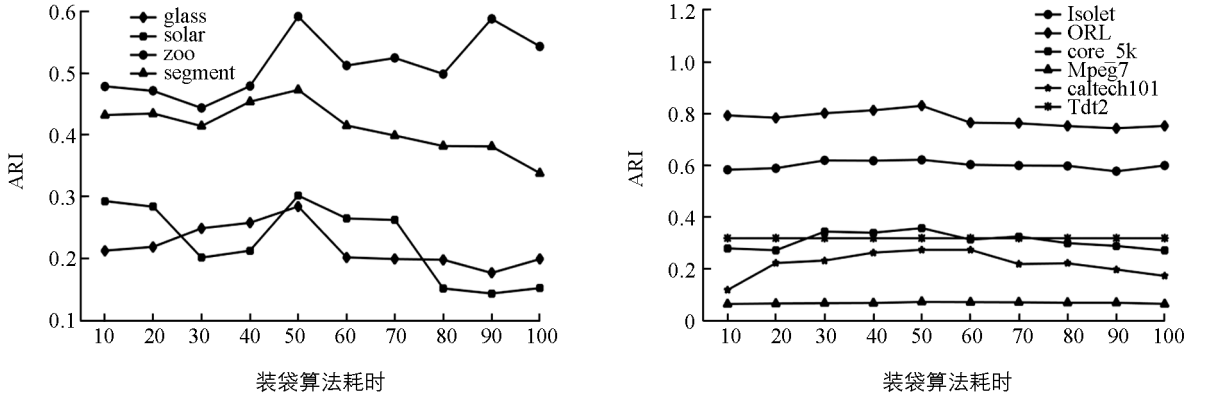

图 2　ARI 与特征抽样次数变化关系

参考文献：

[1] LECUN Y, BENGIO Y, HINTON G. Deep Learning [J]. Nature, 2015, 521(7553): 436-444.

[2] XU D K, TIAN Y J. A Comprehensive Survey of Clustering Algorithms [J]. Annals of Data Science, 2015, 2(2): 165-193.

[3] ZHENG L, YANG Y, TIAN Q. SIFT Meets CNN: a Decade Survey of Instance Retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1224-1244.

[4] ANNUNZIATA R, SAGONAS C, CALI J. Jointly Aligning Millions of Images with Deep Penalised Reconstruction Congealing [EB/OL]. 2019: arXiv: 1908. 04130 [cs. CV]. https: //arxiv. org/abs/1908. 04130.

[5] LUAN Y, LI H. Clustering of Time-course Gene Expression Data Using a Mixed-effects Model with B-splines [J]. Bioinformatics, 2003, 19(4): 474-482.

[6] DHILLON I S, MODHA D S. Concept Decompositions for Large Sparse Text Data Using Clustering [J]. Machine Learning, 2001, 42(1/2): 1-31.

[7] LI Z C, LIU J, YANG Y, et al. Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(9): 2138-2150.

[8] ERHAN D, BENGIO Y, COURVILLE A, et al. Why Does Unsupervised Pre-training Help Deep Learning? [J]. Journal of Machine Learning Research, 2010, 11(3): 625-660.

[9] MACQUEEN J B. Some methods for classification and analysis of multivariate observations [C] // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. California: University of California Press, 1967.

[10] PARK H S, JUN C H. A Simple and Fast Algorithm for K-medoids Clustering [J]. Expert Systems With Applications, 2009, 36(2): 3336-3341.

[11] KAUFMAN L, ROUSSEEUW P J. Partitioning around Medoids (Program PAM) [M] //Finding Groups in Data.

Hoboken: John Wiley & Sons, 2008: 68-125.

[12] SARLE W S, KAUFMAN L, ROUSSEEUW P J. Finding Groups in Data: an Introduction to Cluster Analysis [J]. Journal of the American Statistical Association, 1991, 86(415): 830.

[13] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: an Efficient Data Clustering Method for very Large Databases [J]. ACM SIGMOD Record, 1999, 25(2): 15-34.

[14] GUHA S, RASTOGI R, SHIM K. Cure: an Efficient Clustering Algorithm for Large Databases [J]. Information Systems, 2001, 26(1): 35-58.

[15] GUHA S, RASTOGI R, SHIM K. Rock: a Robust Clustering Algorithm for Categorical Attributes [J]. Information Systems, 2000, 25(5): 345-366.

[16] BEZDEK J C, EHRLICH R, FULL W. FCM: The Fuzzy C-means Clustering Algorithm [J]. Computers & Geosciences, 1984, 10(2-3): 191-203.

[17] DAVE R N, BHASWAN K. Adaptive Fuzzy C-shells Clustering and Detection of Ellipses [J]. IEEE Transactions on Neural Networks, 1992, 3(5): 643-662.

[18] YAGER R R, FILEV D P. Approximate Clustering via the Mountain Method [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1994, 24(8): 1279-1284.

[19] XU X, ESTER M, KRIEGEL H P, et al. A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases [C] // IEEE International Conference on Data Engineering. New York: IEEE Press, 1998.

[20] Rasmussen C E. The Infinite Hidden Markov Model [M] //Advances in Neural Information Processing Systems 14. Massachusetts: The MIT Press, 2002 .

[21] ESTER M. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [C] // Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining. California: AAAI Press, 1996.

[22] ANKERST M, BREUNIG M M, KRIEGEL H P, et al. OPTICS: Ordering Points to Identify the Clustering Structure [C] //Proceedings of the 1999 ACM SIGMOD international conference on Management of data - SIGMOD'99. May 31-June 3, 1999. Philadelphia, Pennsylvania, USA. New York: ACM Press, 1999.

[23] COMANICIU D, MEER P. Mean Shift: a Robust Approach Toward Feature Space Analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603-619.

[24] SCHÖLKOPF B, SMOLA A, MÜLLER K R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem [J]. Neural Computation, 1998, 10(5): 1299-1319.

[25] Macdonald D, Fyfe C. The kernel self-organising map [C] // International Conference on Knowledge-based Intelligent Engineering Systems & Allied Technologies. IEEE, 2000.

[26] BENHUR A, HORN D, SIEGELMANN H T, et al. Support Vector Clustering [J]. Journal of Machine Learning Research, 2002, 2(2): 125-137.

[27] Xu L, Neufeld J, Larson B, et al. Maximum Margin Clustering [C] // Neural Information Processing Systems 1. Massachusetts: MIT Press, 2004: 1537-1544.

[28] LI F J, QIAN Y H, WANG J T, et al. Clustering Ensemble Based on Sample's Stability [J]. Artificial Intelligence, 2019, 273: 37-55.

[29] FREY B J, DUECK D. Clustering by Passing Messages between Data Points [J]. Science, 2007, 315(5814): 972-976.

[30] RODRIGUEZ A, LAIO A. Machine Learning. Clustering by Fast Search and Find of Density Peaks [J]. Science, 2014, 344(6191): 1492-1496.

[31] BONNIER B. Random Sequential Adsorption Ofk-mers on a Square Lattice: The Largekregime [J]. Physical Review E, 1996, 54(1): 974-976.

[32] Chitta R, Jain A K, Jin R. Sparse Kernel Clustering of Massive High-Dimensional Data sets with Large Number of Clusters [C] // Proceedings of the 8th Workshop on Ph. D. Workshop in Information and Knowledge Management 2015. New York: ACM Press, 2015.

[33] CURTIN R R. A Dual-Tree Algorithm for Fast K-means Clustering with Large K [M] //Proceedings of the 2017 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017: 300-

308.

[34] QIAN Y H, LI F J, LIANG J Y, et al. Space Structure and Clustering of Categorical Data [J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(10): 2047-2059.

[35] CAO J, ZHENG Q, WENG N, et al. Low Dimensional Representation of Space Structure and Clustering of Categorical Data [C] //IEEE International Conference on Ubiquitous Computing. New York: IEEE Press, 2018: 1079-1086.

[36] 王　齐，钱宇华，李飞江. 基于空间结构的符号数据仿射传播算法 [J]. 模式识别与人工智能，2016，29(12)：1132-1139.

[37] HUANG D, WANG C D, WU J S, et al. Ultra-Scalable Spectral Clustering and Ensemble Clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(6): 1212-1226.

[38] FAN Z Y, JIANG J, WENG S Q, et al. Adaptive Density Distribution Inspired Affinity Propagation Clustering [J]. Neural Computing and Applications, 2019, 31(S1): 435-445.

MS^{2}BC-a Multi-view Space Structure-Based Clustering Algorithm

SHI Xin-rui[1,3], QIAN Yu-hua[1,2,3], Li Fei-jiang[1,3]

1. *Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China;*
2. *Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education (Shanxi University), Taiyuan 030006, China;*
3. *School of Computer and Information Technology, Taiyuan 030006, China*

Abstract: Clustering analysis is an important task in the field of machine learning and data mining. In recent years, a large number of clustering algorithms have been proposed and successfully used in many fields. However, the complex development of data at this stage has brought great challenges to the existing clustering algorithms, in which the rapid increasing number of potential clusters is very representative. To address this problem, a multi-view space structure-based clustering method (MS^{2}BC) is proposed in this paper. Space structure is a data representation method that can maintain the structure of data clusters and provide richer measurement information. Based on the space structure representation method and the bagging feature sampling technology, this paper constructs and integrates the space structure of the data from multiple views, and uses the integrated space structure representation to complete the clustering. Finally, the superiority of this method over other representative clustering methods in clustering performance is verified based on 10 real data.

Key words: clustering analysis; multi-cluster; space structure; clustering ensemble; data mining

责任编辑　张　枸