DOI: 10.13718/j.cnki.xdzk.2016.03.027

面向分类数据的重叠子空间聚类算法 SCCAT①

张辉荣，　唐　雁，　何　荧，　石教开，　徐平安

西南大学 计算机与信息科学学院，重庆 400715

摘要：改进了面向分类数据的重叠子空间聚类算法(ROCAT)，提出了面向分类数据的重叠子空间聚类算法(SCCAT). 利用数据凝聚力模型(DCC)代替 ROCAT 的数据压缩模型以提高算法精度；将源数据集分为样本内数据集和样本外数据集，采取对样本内数据聚类，对样本外数据分类的方法完成聚类来降低算法复杂度. 实验结果表明 SCCAT 在提高算法精度的同时，也降低了算法的时间复杂度和空间复杂度，适用于大规模数据的处理.

关　键　词：分类数据；复杂度；精度；凝聚力

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673－9868(2016)03－0171－06

随着科学技术在工业界与产业界的广泛推广和使用，近十年来，数据的规模正以数十倍于之前的增长速度快速膨胀，因此，如何处理这些规模庞大的数据成为了近年来研究者最关注的问题之一[1－16]. 这些数据的形式大致可以分为如下三种类型：数值型、分类型以及混合型. 所谓数值型，也叫连续型，指的是两个不同的值之间可取的值有无穷多个的其它取值，如长度、气温等. 这种类型的数据具有很多几何特性，可以进行各种运算. 分类型数据跟连续型数据相反，它的取值有限，如种族、国家、性别等. 分类型数据没有几何特点. 混合型数据指的是分类型数据和数值型数据的组合. 文献[4]从数据压缩的角度考虑，提出了面向分类数据的重叠子空间算法 ROCAT[4].

ROCAT 算法具有无须输入参数、能够高效处理重叠子空间聚类问题、对输入顺序不敏感等优点. 然而，由于 ROCAT 采用的是数据压缩模型，当簇的规模不够大导致对簇进行编码的代价比将其当作非簇处理时的编码代价更大时，这个簇将会被算法忽略，导致算法的精确度下降. 另外，ROCAT 算法在开始时需要将全部数据一次性载入内存中，算法的时间复杂度和空间复杂度非常高. 针对这些问题，提出面向分类数据的重叠子空间聚类算法(a novel subspace clustering method based on categorical data，SCCAT). 为了提高算法精度，利用数据凝聚力模型(DCC)代替 ROCAT 的数据压缩模型；为了降低算法复杂度，用随机取样算法将源数据集分为样本内数据集和样本外数据集，采取对样本内数据聚类，对样本外数据分类的方法完成聚类. 实验结果表明 SCCAT 在提高算法精度的同时，也降低了算法的时间复杂度和空间复杂度，适用于大规模数据的处理.

1　面向分类数据的重叠子空间聚类算法(SCCAT)

传统的引力模型[5－6]的关注点是对象与簇之间的关系，它们认为簇和数据对象之间存在引力，簇与对

① 收稿日期：2015－04－26
基金项目：教育部“春晖计划”项目(z2011149).
作者简介：张辉荣(1989－)，男，江西宜春人，硕士研究生，主要研究领域为数据挖掘.
通信作者：唐　雁，教授.

象的数据引力越大，对象越趋向于具有该簇的特点，在分类算法当中也越趋向于将对象归到这个簇当中. 凝聚力DCC(data cohesion cluster)模型的思想与这一观点类似，所不同的是，DCC的关注点放在簇本身，认为簇本身具有凝聚力，且凝聚力的大小由子簇的对象数和属性数共同决定，子簇包含的对象越多，所涉及的属性越多，簇越团结，所具有的凝聚力越大. 基于这个思想，本文构造了DCC模型，用以提高算法精度. 为了降低ROCAT算法的时间复杂度和空间复杂度，用随机取样算法将源数据集分为样本内数据集和样本外数据集，采取对样本内数据聚类，对样本外数据分类的方法完成聚类. 基于这一思想，提出了面向分类数据的重叠子空间聚类算法SCCAT. 包含以下5个步骤：

1) 随机取样. 采用随机取样算法，将源数据分解为样本内数据集和样本外数据集.

2) 生成样本簇. 用DCC模型对样本内数据集进行聚类，得到样本簇.

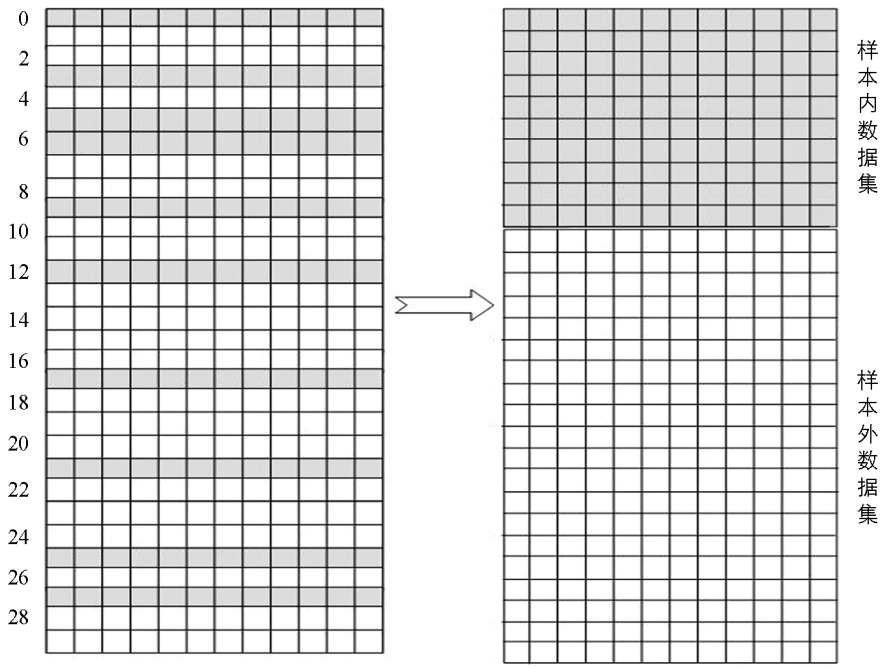
3) 生成簇的特征向量. 用互信息判断样本簇属性值的相似度确定样本簇特征向量.

4) 相似度计算. 将簇的特征向量与样本外数据集数据的夹角余弦值作为二者的相似程度以完成对样本外数据集的分类.

5) 重复步骤4)完成聚类.

1.1 随机取样

取样的好坏关系到数据样本是否完整从而影响聚类质量. 若样本集数据过于集中在部分数据，则数据样本不能代表全部数据，影响聚类质量. 为了避免这一问题，SCCAT算法采用的是随机取样的方法. 随机取样算法分为两步，首先根据数据大小生成一个随机序列，生成方法见图1. 得到随机序列后使用split()算法将原数据分割成样本内数据集和样本外数据集两部分，输入参数为源数据data，随机序列sequence以及样本大小sample_data.size，按样本大小在随机序列中取与样本大小相同的前n个样本点，数据大小为30，样本大小为10，随机序列为random_sequence. 取随机序列的前10个数据实例，即取第5,3,27,21,0,25,17,12,6,9点共10个实例，组成样本内数据集，其它数据实例作为样本外数据集.



DATA. size=30, IN_SAMPLE. size=10; OUT_OF_SAMPLE. size=20
Random_sequence={5, 3, 27, 21, 0, 25, 17, 12, 6, 9, 15, 8, 14, 20, 19, 10, 18, 7, 11, 24, 13, 2, 26, 28, 22, 4, 29, 1, 23, 16}

图1 随机取样图解

1.2 生成样本簇

用数据凝聚力DCC(datacohesion cluster)模型对1.1中产生的样本内数据集聚类，得到样本簇. DCC模型包含两个主要的步骤：计算机凝聚力和判定距离.

1.2.1　计算凝聚力

DCC 模型将簇定义为一个由凝聚力、属性、对象构成的三元组 $C=\{F, A, O\}$，每个属性和每个对象都是凝聚力的组成要素，其大小指的是一个对象要离开当前簇时受到的同簇中其它对象对它的引力. 按照万有引力定律，对象 A 受到对象 B 的引力大小可以描述为(1)式：

$$F_{A\leftrightarrow B}=G\frac{C_{A_a}\times C_{B_a}}{r_{A\leftrightarrow B}^{\ 2}} \tag{1}$$

其中：C_{A_a}，C_{B_a} 分别表示当前簇中对象 A，B 的属性个数；$r_{A\leftrightarrow B}$ 表示两者之间的距离. 本文只需比较引力大小而非具体数值，故引力常量 G 可以省略. 结合(1)式，对于含有 n 个对象的簇，一个对象受到其它 $n-1$ 个对象的引力可描述为：

$$F=\sum_{i=1}^{n-1}\frac{C_{i_o}\times C_o}{r_i^{\ 2}} \tag{2}$$

1.2.2　判定距离

差异度：是指在同一个子簇当中，两个对象属性值的差异程度，差异度的形式化描述如下：

$$D(x, y)=1+\sum_{k=1}^{m}d(x_k, y_k) \tag{3}$$

$$d(x_k, y_k)=\begin{cases}1 & (x_k\neq y_k)\\ 0 & (x_k=y_k)\end{cases} \tag{4}$$

其中：x，y 分别表示同一个簇中的两个不同对象；x_k，y_k 分别表示对象 x，y 的第 k 个属性值. 当两个相同属性的属性值相等时，距离为 0，否则为 1，依次比较所有的属性，计算两个对象的距离. 为了避免所有属性都相等，导致 $D(x, y)=0$ 这种情况的出现，对于所有的 $D(x, y)$ 都加上 1，使得 $r_i\neq 0$.

1.2.3　DCC 模型算法思想

DCC 模型的算法思想可以描述为下面 6 个步骤：

1) 将数据集 D 加入搜索队列.

2) 取出搜索队列的第一个元素作为搜索空间，迭代计算这个搜索空间每一个属性值的信息熵，找出最小的信息熵的属性列，同时记录这个属性的出现次数最多的属性值，记录下这个属性和属性值构成的矩阵，为第一个纯子簇[4](图 2)，第一列且属性值为 b 的矩形为第一个纯子簇.

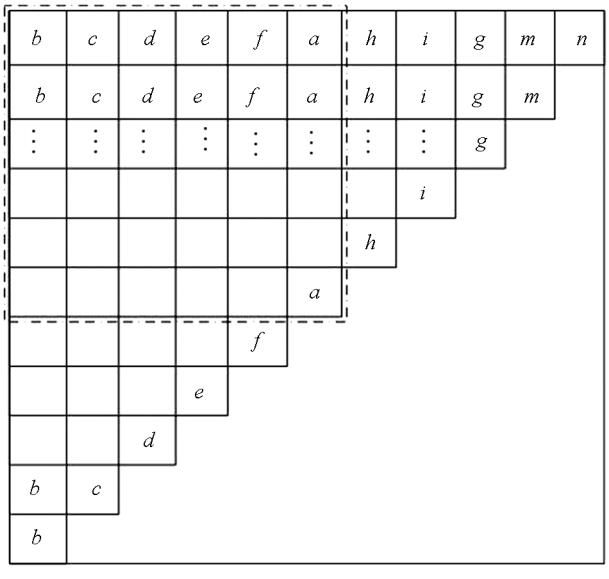


图 2　查找最佳纯子簇

3) 去除掉步骤 2)中没有涉及的对象，在已经搜索过的属性之外，重复步骤 2)，找到信息熵第二小的属性列，记录这个属性出现次数最多的值，与步骤 2)中找到的最小信息熵属性列和出现最多的值组合，在组合区域画出一个最大面积的矩形，此为第二个纯子簇. 例如在图 2 中，第一列属性值为 b、第二列属性值为 c 的元素所在的矩形为第二个纯子簇.

4) 重复步骤 3)，直到所有的纯子簇被查找出来.

5) 计算步骤 4)中所有纯子簇的情况下整个数据集的凝聚力((2)式)，取使得凝聚力达到最大的纯子簇作为最佳纯子簇，同时，将最佳纯子簇从原搜索空间中分离出来(这个簇不涉及的属性分为一部分，不涉及的实例也分为一部分)，并将它们加入搜索队列.

6) 重复步骤 2)到步骤 5)，直到所有的最佳纯子簇都被查找出来.

1.3 生成簇的特征向量

若把每一个属性的值当成一个词语，则相邻于这个属性值的属性值称为这个词语的上下文，也可以称为词组. 在样本簇生成以后，如何生成簇内的词组关系从而能够判断后面加入的数据是否也具有这一簇的词组特性，是生成特征向量的关键. SCCAT 算法采用的是使用互信息来进行词语特征学习，互信息的计算机公式[11]如下：

$$PMI(a,b)=\log_2\left\{\frac{p(a,b)}{p(a)\times p(b)}\right\} \tag{5}$$

其中：a,b 是同一个实例的两个相邻属性的属性值；$P(a,b)$ 表示点 P_m 的属性 A_i 的值为 a 且属性 A_j 的值为 b 的概率，$P(a)$,$P(b)$ 分别表示属性 A_i,A_j 的值为 a,b 的概率. 因为(5)式中 PMI 的值是关于 $\left\{\frac{p(a,b)}{p(a)\times p(b)}\right\}$ 递增而递增的，我们这里只讨论 PMI 的大小，对具体的值不关心，所以在 SCCAT 中，采用 $PMI(a,b)=\left\{\frac{p(a,b)}{p(a)\times p(b)}\right\}$ 来计算词语 a 和词语 b 的相关程度，显然，a 与 b 的相关性随 $PMI(a,b)$ 的增加而增加，故可由此计算 a 与 b 的相关程度，简化计算复杂程度，完成词语特征的学习，生成一个关于簇的词语特征向量.

1.4 相似度计算

在新的记录加入时，取相同的属性序列的属性值作为新记录的特征向量，与簇的特征向量比较，若对应位置的值相同，则新向量的值置为 1，否则置为 2，这样就生成了一个基于新记录且关于簇特征向量的向量，求这个向量与同维度单位向量的夹角余弦值，用以表示新向量与特征向量的相似度，从而判断新记录是否也具有该簇特征.

1.5 复杂度分析

对于一个由 N 个对象、M 个属性组成的数据集，ROCAT 算法的时间复杂度为 $O(M^2,N)$，而 SCCAT 算法对于样本内数据集的时间复杂度为 $O(M^2,N_1)$，对于数据外数据集的时间复杂度为 $O(M,N_2)$，其中：N_1 为样本内数据集的对象数，N_2 为样本外数据集的对象数，且 $N=N_1+N_2$. ROCAT 在算法开始时将所有数据读入内存，并在以后的迭代过程中多次遍历所有维度和数据点，使算法复杂度随记录集的大小呈指数增长. SCCAT 算法时间复杂度仅仅与样本时间复杂度有关，生成特征向量和求相似度的时间复杂度都非常低. SCCAT 算法的空间复杂度为 $O(M*N_1)$，而 ROCAT 算法的空间复杂度为 $O(M*N)$.

2 实 验

本节基于 SCCAT 算法在 3 个基准数据集上进行实验，将 SCCAT 算法与 ROCAT 算法从准确度、时间和空间复杂度 3 个方面进行对比.

2.1 实验数据

本文采用的是 UCI 数据库上的 3 个数据集，分别为 mushroom,vote 和 US Census Data (1990)数据集. UCI 数据库是加州大学欧文分校(University of CaliforniaIrvine)提出的用于机器学习的数据库，这个数据库目前共有 187 个数据集，其数目还在不断增加，UCI 数据集是一个常用的标准测试数据集.

2.2 评价标准

准确率和召回率是广泛应用于聚类、分类和机器学习领域的一个重要评价体系，又称为查准率和查全率. 本文采用了综合评价指标 F-Measure，用来评估算法的精度. F-measure 是准确率和召回率的加权平均值. 计算公式如下：

$$F=\frac{2\times p\times r}{p+r} \tag{6}$$

其中：p 指的是准确率；r 指的是召回率.

2.3 实验结果

实验分为两个部分，第一部分针对算法的时间复杂度和分类精度，用 mushroom 和 vote 两个数据集对算法在准确度、时间和空间复杂度上作了一个分析对比(表 1,2).

表 1　vote 数据集结果对比

	P	*R*	*F*-measure	时间/ms
ROCAT	0.81	0.52	0.62	97
SCCAT	0.79	0.83	0.81	72

表 2　mushroom 数据集结果对比

	P	*R*	*F*-Measure	时间/ms
ROCAT	1	0.27	0.42	1 478
SCCAT	0.64	0.85	0.75	932

由表 1，2 可知，ROCAT 算法的准确率很高，但召回率非常低，所以导致 *F*-Measure 这个综合评价指标的值偏低，而 SCCAT 算法在提升了聚类精度的前提下，大大缩短了运行时间，算法在聚类精度和时间复杂度上面都优胜于 ROCAT 算法.

第二部分采用 US Census Data (1990)数据集，对算法的时间复杂度和空间复杂度作了一个详细对比，实验结果如图 3－5 所示.

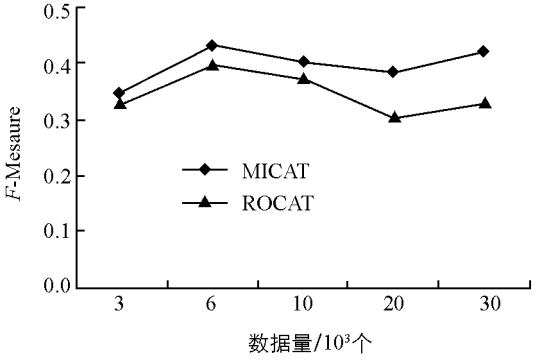


图 3　*F*-Measure 与数据规模的关系

从图 3 中可以看出，SCCAT 的 *F*-Measure 优于 ROCAT. 由图 4 可知，ROCAT 的运行时间与数据量大小的关系是趋向于指数级增长，而 SCCAT 随着数据量的增加，运行时间变化曲线较为平缓. 因此，SCCAT 算法在时间复杂度上优于 ROCAT 算法.

图 5 展示了 SCCAT 和 ROCAT 在空间使用方面随时间变化的曲线图. 在数据量为 3 000 时，MCIAT 的内存使用平均是 4.2 MB，而 ROCAT 的内存使用达到了 13 MB 之多；在数据量为 6 000 时，SCCAT 的内存使用平均为 5.9 MB，而 ROCAT 达到了 22 MB；在数据量分别为 10 000，20 000，30 000 时，差距更加明显. 此外，ROCAT 算法在运行时占用内存时间比 SCCAT 长很多，且 ROCAT 算法占用的时间随数据规模变化呈指数级增长，而 SCCAT 算法则相对较为稳定. 因此，SCCAT 算法在时间复杂度和空间复杂度上均优于 ROCAT.

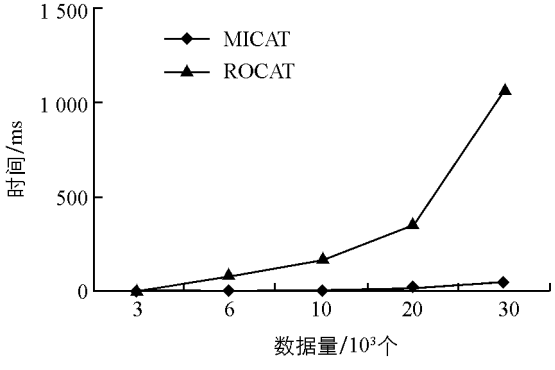


图 4　处理时间与数据规模的关系

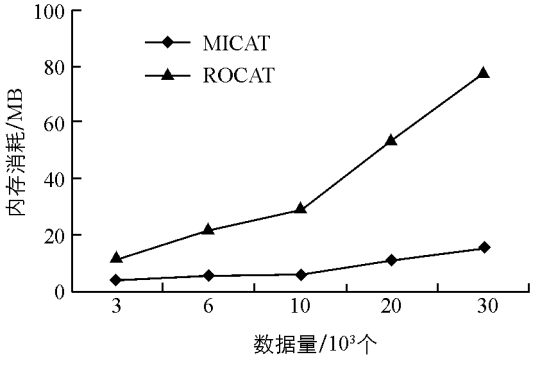


图 5　内存消耗与数据量的关系图

参考文献：

[1] RALAMBONDRAINY H. A Conceptual Version of The K-Means Algorithm [J]. Pattern Recognition Letters, 1995, 16(11): 1147－1157.

[2] HUANG Z. Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values [J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283－304.

[3] GAN G, WU J, YANG Z. A Genetic Fuzzy K-Modes Algorithm for Clustering Categorical Data [J]. Expert Systems with Applications, 2009, 36(2): 1615－1620.

[4] HE X, FENG J, KONTE B, et al. Relevant Overlapping Subspace Clusters on Categorical Data [C] //Proceedings of

the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. Providence: ACM Press, 2014: 213－222.

[5] PENG L, CHEN Y, YANG B. A Novel Classification Method Based on Data Gravitation [C] //Neural Networks and Brain, 2005. New York: IEEE Press, 2005, 2: 667－672.

[6] CANO A, ZAFRA A, VENTURA S. Weighted Data Gravitation Classification for Standard and Imbalanced Data [J]. IEEE Transactions onCybernetics, 2013, 43(6): 1672－1687.

[7] GRüNWALD P D. The Minimum Description Length Principle [M]. Massachusetts: MIT press, 2007.

[8] BARRON A, RISSANEN J, YU B. The Minimum Description Length Principle in Coding and Modeling [J]. IEEE Transactions on Information Theory, 1998, 44(6): 2743－2760.

[9] POLCZYNSKI M, POLCZYNSKI M. Using the k-Means Clustering Algorithm to Classify Features for ChoroplethMaps [J]. Cartographica: The International Journal for Geographic Information and Geovisualization, 2014, 49(1): 69－75.

[10] HUANG Z, NG M K. A Fuzzy K-Modes Algorithm for Clustering Categorical Data [J]. IEEE Transactions on Fuzzy Systems, 1999, 7(4): 446－452.

[11] LU H, REYES M, SERIFOVIC A, et al. Multi-Modal Diffeomorphic Demons Registration Based on Point-Wise Mutual Information [C] //2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. New York: IEEE Press, 2010: 372－375.

[12] BARBARá D, LI Y, COUTO J. COOLCAT: An Entropy-Based Algorithm for Categorical Clustering [C] //Proceedings of the Eleventh International Conference on Information and Knowledge Management. Providence: ACM Press, 2002: 582－589.

[13] 胡晨晓，邹显春，陈 武，等. 基于稀疏表示的近邻传播聚类算法 [J]. 西南大学学报(自然科学版)，2014，36(5)：220－224.

[14] 施成湘. 基于二次分水岭和近邻传播聚类的彩色图像分割算法研究与实现 [J]. 西南师范大学学报(自然科学版)，2013，38(8)：125－129.

[15] 赵 芳，马玉磊. 自训练半监督加权球结构支持向量机多分类方法 [J]. 重庆邮电大学学报(自然科学版)，2014，26(3)：404－408.

[16] 丰江帆，朱 毅. 云环境下的流式空间信息服务 [J]. 重庆邮电大学学报(自然科学版)，2012，24(6)：115－119.

A Novel Subspace Clustering Method Based on Categorical Data (SCCAT)

ZHANG Hui-rong, TANG Yan, HE Ying,
SHI Jiao-kai, XU Ping-an

College of Computer and Information Science, Southwest University, Chongqing 400715, China

Abstract: ROCAT (Relevant Overlapping Subspace Clusters on Categorical Data) has overcome a series of conventional problems, such as the difficulties in defining the parameters or a large amount of redundancies or other troubles caused by the conventional algorithm as CLIQUE, K-means and so on. Meanwhile, its data compression model reduces the clustering accuracy with high complexity. For dealing with these problems, we propose a novel Subspace Clustering Method based on Categorical Data (SCCAT). We put forward a data cohesion model (DCC) instead of the data compression model to improve the clustering accuracy, divide the data source into in-sample data and out-of-sample data, cluster the in-sample data to obtain sample clusters and spread the out-of-sample data to the sample clusters to finish clustering. As the experiment results show, SCCAT not only improves the accuracy but also reduces the time complexity and the space complexity, which makes it suitable for large-scale data processing.

Key words: categorical data; complexity; accuracy; cohesion

责任编辑 张 枸