DOI：10.7652/xjtuxb201410017

两阶段密度意识子空间聚类模型

李长路[1,2]，王劲林[1]，郭志川[1]，潘梁[1]
（1. 中国科学院声学研究所国家网络新媒体工程技术研究中心，100190，北京；2. 中国科学院大学，100049，北京）

摘要：针对网格聚类方法在高维子空间聚类中网格规模随着维度急剧升高的问题，以及差别阈值方法引入干扰小聚簇的问题，提出一种具有两个网格划分阶段的密度意识子空间聚类模型。该模型第一阶段采用粗网格找出可能存在聚类的子空间区域，第二阶段在这些区域中进行等效精度更高的网格划分并找出所有致密单元。该模型在两个阶段处理的网格规模均远低于密度意识子空间聚类模型在相同划分精度下的网格规模，同时利用第一阶段对网格空间的筛选作用降低小聚簇干扰，提高聚类质量。合成数据集实验表明：该模型聚类精准率和查全率性能明显优于原模型；基于真实数据集实验，相比一次划分模型，该模型以损失 0.4% 数据点的代价提高输出聚类密度 19.4%，聚类质量大幅提升。

关键词：数据挖掘；子空间聚类；网格聚类；高维数据
中图分类号：TP393 **文献标志码**：A **文章编号**：0253-987X(2014)10-0108-07

A Density Conscious Subspace Clustering Model with Two Steps

LI Changlu[1,2], WANG Jinlin[1], GUO Zhichuan[1], PAN Liang[1]
(1. National Network New Media Engineering Research Center, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: A grid-based density conscious subspace clustering model with two grid partition steps is proposed to focus on the problem that the amount of grid cells grows rapidly with the increase of dimension, and a model with different thresholds for different subspaces brings in a large number of small interferential clusters. The first stage of the model uses coarse grids to find out possible regions for clusters in all subspaces, and the second stage finds all the dense units in these regions using an equivalent partition with accuracy higher than that in the first stage. The total number of grids in any stage of the proposed model is much lower than that of the density conscious subspace clustering model when the precision is the same. Screening to coarse grids in the first stage reduces the interference of small clusters, and improves the quality of clustering. Experimental results on a synthetic data set show that both the improved model precision rate in clustering and the recall rate performance are obviously better than those of the original model. Experiment results on a real data set and an experiment data set show that the model improves the output cluster density 19.4% at the expense of the loss 0.4% data points, and improves the quality of output clusters.

Keywords: data mining; subspace clustering; grid-based clustering; high-dimensional data

---

收稿日期：2014-03-05。 作者简介：李长路(1987—)，男，博士生；王劲林(通信作者)，男，教授，博士生导师。 基金项目：国家高技术研究发展计划资助项目(2011AA01A102)；国家科技支撑计划资助项目(2012BAH73F01)；中国科学院战略性先导科技专项资助项目(XDA06010301)。
网络出版时间：2014-09-23 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20140923.1747.004.html

随着生物信息学和电子商务应用的迅速发展，在这些领域积累了大量高维数据。对高维数据的子空间聚类[1-11]能够发现隐藏在不确定子空间的聚类及其所在子空间，是实现高维数据集聚类的有效途径。由于高维数据空间存在维度灾难现象[12-14]，数据越来越稀疏，数据间距离趋于均匀，这使得基于传统相似度量的聚类挖掘方法在高维空间无法有效工作。

根据是否使用网格结构，现有的子空间聚类方法主要分为两类。以 PROCLUS[1] 和 ORCLUS[2] 为代表的非网格方法基于一定的距离公式，对整个数据集进行多次遍历，不断改进聚类结果，但在高维数据集聚类应用中，非网格类算法在聚类效率、发现任意形状聚类以及聚类结果的可解释性等方面表现较差。CLIQUE[3] 及其改进方法 ENCLUS[4]、MAFIA[5] 则采用网格结构子空间聚类。这类方法通常结合传统聚类中基于密度的方法和基于网格的方法，理论上能够发现任意形状的聚簇，且不依赖距离，处理数据集效率高，因而受到青睐。这类算法将各个特征维度分为若干区间，从而在所有子空间上都形成了网格结构，之后的所有聚类操作都在网格上进行，从而降低计算代价，提高聚类效率。

这些网格类方法[3-5] 以及非网格的方法 SUBCLU[6] 都设计了类 Apriori 的算法来发现聚簇，充分利用了密度的向下闭包性质，即如果 k 维子空间中存在稠密区域，那么在投影到该子空间的所有 $k-1$ 维子空间中，也一定存在稠密区域；相反，如果给定的 $k-1$ 维子空间中不存在致密区域，则其对应任意的 k 维空间也不存在致密区域。这一性质的基础是对所有子空间的网格单元，采用全局统一的阈值来判断其是否属于组成聚类的致密单元。

高维子空间的网格单元远多于低维子空间，因为数据密度被严重稀释，采用全局阈值会在高维和低维子空间之间造成不公平，因此相关研究又提出在不同维度空间采用差别阈值的密度意识子空间聚类算法（DENCOS）[15]。该方法在不同子空间采用与子空间平均密度相关联的阈值，即高维子空间的阈值较低维子空间阈值更低，从而在高维数据集中实现公平的子空间聚类。

数据集维度升高还带来以下问题：①干扰小聚簇。少量数据会随机聚集在一些较小区域，在局部形成较高的数据密度，但这类小聚簇并不代表整个数据集的分布规律。在高维子空间，由于数据稀疏性，这类随机的小规模数据聚集更为常见，而采用更小的阈值会增强这一现象的影响。同时，更细的网格划分使得这些小规模聚集的区域更可能作为一个独立的致密单元存在，干扰聚类结果。②资源消耗。网格类聚类算法处理的数据规模由网格单元规模决定，随着维度上升和划分精度提高，网格单元数量按照幂函数增长，随之带来的时间空间资源消耗惊人。

通过降低网格划分精度可以在一定程度上缓解上述问题，但效果有限，对聚类质量的牺牲很大。由于网格法聚类本身的特点和高维数据稀疏性，高维网格聚类对网格划分十分敏感，降低精度很容易导致高维聚类的结果失去意义。

本文设计的两阶段密度意识子空间聚类模型保留了现有密度意识聚类模型的理想网格特性，能够方便地运用现有密度意识子空间聚类算法。两阶段网格划分在每一个阶段分别实现较少的网格划分，极大地降低了网格单元规模。由于在高密度区域进行实际精度较高的第二阶段划分，使得模型具有很好的聚类精度和聚类边界识别能力，而第一阶段的筛选有效提升了模型抵御小聚簇干扰的能力。

1　相关研究工作

1.1　网格聚类的一般步骤

基于网格的聚类方法，其模型和算法的设计与网格划分有着紧密关联[3-5,15-17]，其聚类过程一般分为以下步骤。

步骤 1　模型设计/网格结构建立。划分每一个属性维，在所有子空间形成网格结构，并给出该网格结构下致密单元定义。

步骤 2　算法设计/致密单元搜索。根据网格结构特征和致密单元的定义，分析总结致密单元在不同子空间中分布规律，并据此提出快速发现所有致密单元的算法，找出致密单元。

步骤 3　聚类输出。将所有致密单元按照所属子空间以及相互之间的紧密程度（必须连通或者不同程度地靠近）分组，形成各个聚类。

为了提高子空间聚类的质量，相关研究[4-5,15-17]提出了不同的网格划分方式和致密单元定义，并设计相应算法，以使得网格聚类能够在高维数据聚类、任意形状聚类等方面有更好的表现；同时控制高维空间聚类时处理网格单元的规模，因为这一规模直接影响算法运行过程中的时间、空间资源消耗。例如密度意识子空间聚类算法 DENCOS，由于采用差别阈值定义致密单元，使得向下闭包性质不再成立，但 DENCOS 通过采用规则的网格结构，同时通过用户指定的全局的参数将子空间阈值与子空间平均密

度关联起来，使得同维数的不同子空间阈值相同，不同维数的子空间阈值随着维数增长而单调递减，并据此设计了基于频繁项树的致密单元快速搜索算法。

1.2 密度意识子空间聚类模型

为了加速致密单元搜索过程，密度意识子空间聚类模型[15]对所有维度采取相同数量的均匀间隔划分。

定义 1 多维空间。$A=\{A_1,A_2,\cdots,A_d\}$ 表示数据集的 d 维属性集，$S=A_1\times A_2\times\cdots\times A_d$ 是由这些属性确定的 d 维空间，而 $A_1,A_2,\cdots,A_d$ 为 S 的属性维，任意 k 个 S 的维组成空间 S_k 成为 S 的一个 k 维子空间，而 S 被称为属性集的满空间。S 的任何 k 维子空间由 A 中选取的 k 个属性确定，因此 $k\leqslant d$。

定义 2 网格划分。输入一个参数 δ，将每一个属性维的取值区间均匀划分为 δ 个间隔区间，δ 取值越大，划分精度越高。因此数据空间第 i 维的取值区间为前闭后开空间$[\min_i,\min_i+\delta\varepsilon_i)$，其中每个区间间隔长度，亦即网格分辨率

$$\varepsilon_i=(\max_i-\min_i)/\delta \tag{1}$$

式中：$\max_i$ 和 $\min_i$ 分别为第 i 维上取的极大值与最小值。在第 i 维上第 j 个间隔，亦即该维度上第 j 个一维网格单元的取值范围为

$$u_{ij}=[\min_i+(j-1)\varepsilon_i,\min_i+j\varepsilon_i),\quad 1\leqslant i\leqslant d,\quad 1\leqslant j\leqslant\delta \tag{2}$$

第 i 维上的所有间隔集合

$$U_i=\{u_{ij}\mid 1\leqslant j\leqslant\delta\},\quad 1\leqslant i\leqslant d \tag{3}$$

满空间 S 被划分为 δ^d 个 d 维空间单元，每一个 d 维单元都是 d 个维度上间隔形成的 d 维交叠区域

$$u_d=u_{1p_1}u_{2p_2}\cdots u_{dp_d},1\leqslant p_l\leqslant\delta,1\leqslant l\leqslant d \tag{4}$$

同理，任一 k 维子空间 S_k 被划分为 δ^k 个 k 维空间单元 u_k。

定义 3 网格密度。对于包含 N 个数据点的 d 维空间数据集 $V=\{v=(v_1,v_2,\cdots,v_d)\mid v_i\in A_i\}$ 一个数据点 $v=(v_1,v_2,\cdots,v_d)$，落在所在 k 维空间的网格单元 $u=u_{1p_1}u_{2p_2}\cdots u_{kp_k}$，当且仅当

$$v_i\in u_{ip_i} \tag{5}$$

落在一个多维网格单元 u 中的数据点个数 $D(u)$，称为该单元的计数值，等于这个单元的数据密度。由于一个 k 维子空间被分为 δ^k 个网格单元，k 维子空间的平均密度为

$$D(k)=N/\delta^k \tag{6}$$

定义 4 致密单元。输入一个由用户指定的参数 α，表示 k 维子空间阈值 τ_k 与子空间平均计数值的比率

$$\tau_k=\alpha D(k)=\alpha(N/\delta^K) \tag{7}$$

对于一个 k 维子空间网格单元 u，当且仅当 $D(u)\geqslant\tau_k$ 时，被认为是致密单元。当 α 取值较大时，牺牲查全率，以取得较高的精准率；当 α 较小时，牺牲精准率，以取得较高的查全率。

密度意识子空间聚类模型不基于距离发现聚类，且在较高维子空间自适应地采用较小的阈值来判定其致密单元，一定程度上解决了高维数据空间的典型问题：数据间距离相似度量失效；数据分布稀疏。同时，因为该模型在将数据分布转换为网格密度分布之后，直接在网格结构上操作，致密单元发现过程所消耗的时间与空间资源并不随着数据规模 N 的增加而增加。

该模型的设计存在一定局限性：当 δ 取值较小时，网格划分粗糙，严重降低对任意形状聚类的精准率；对于拥有不规则边缘的聚类，会在聚簇边缘造成识别的混乱，遗漏某些聚簇内区域，并错误纳入非聚簇区域。为了适应非矩形聚类的情况，应使每个网格单元在一定程度上小于聚类尺寸，并与聚类边界精度相当，以满足对聚类边界识别程度的要求。为了识别任意形状聚簇的边界，必须增大 δ 取值，使网格划分更精细。但 δ 取较大值时，会大幅提升高维数据空间需要搜索的网格单元规模，造成资源消耗膨胀，同时在高维子空间引入更多无意义的干扰小聚簇，限制该算法在实际系统中无法用于发现较高维度空间聚类。

2 本文模型设计

2.1 两阶段网格划分模型(TSDENCOS)描述

由于在高维子空间数据稀疏，存在大量空白区域、数据稀薄区域，本文认为不值得在这些区域做较高精度的网格划分和致密单元搜索。另外，当在一片区域内只有少数甚至个别第二阶段网格单元密度超过阈值时，我们认为这些体积过小的密集单元不是一个有效聚类的组成部分，而是子空间内的无效干扰小聚簇。

在本文的模型中，致密单元的搜索被分为两个阶段：聚类区域搜索利用较粗的网格划分，发现可能存在聚类的子空间区域；在找出的聚类区域采用较细的网格划分，找出区域内致密单元。

第一阶段网格划分输入每个维度划分间隔数 δ_1，包含 4 个定义，可参考密度意识子空间定义 1～4。其中定义 4 修改如下。

定义 5 致密区域。输入一个由用户指定的参数 α，表示 k 维子空间阈值 τ_{1k} 与子空间平均计数值的比率

$$\tau_{1k} = \alpha D_1(k) = \alpha(N/\delta_1^k) \tag{8}$$

式中:$D_1(k)=N/\delta_1^k$ 为第一阶段划分得到的网格结构中 k 维子空间的平均密度。对于一个 k 维子空间一阶段网格单元 u_k,当且仅当 $D_1(u_k)\geqslant\tau_{1k}$ 时,被认为是 k 维致密区域。

当 α 取值较大时,牺牲查全率,以取得较高的精准率;当 α 较小时,牺牲精准率,以取得较高的查全率。

定义 6　候选致密单元。第二阶段划分在致密区域中进行,将所有维度上第一次划分的间隔 δ_2 等分,任一 k 维致密区域 u_k 被划分为 $(\delta_1)^k$ 个 k 维二阶段网格单元 u'_k,一个 k 维子空间被划分为 $(\delta_1\delta_2)^k$ 个第二阶段网格单元。对于 k 维第二阶段网格单元 $u'_k=u'_{1p_1}u'_{2p_2}\cdots u'_{kp_k}$ 以及同一子空间的致密区域 $u_k=u_{1p_1}u_{2p_2}\cdots u_{kp_k}$,当且仅当

$$u'_{ip_i}\subseteq u_{ip_i},\quad 1\leqslant i\leqslant k \tag{9}$$

时,认定 u'_k 落在 u_k 内。落在致密区域中的第二阶段网格单元作为候选致密单元。

定义 7　致密单元。由于每个 k 维子空间被分为 $(\delta_1\delta_2)^k$ 个第二阶段网格,其平均密度为

$$D_2(k) = N/(\delta_1\delta_2)^k \tag{10}$$

对于第二阶段网格单元的阈值

$$\tau_{2k} = \alpha D_2(k) = (\alpha N)/(\delta_1\delta_2)^k \tag{11}$$

对于一个 k 维子空间候选密集单元 u'_k,当且仅当 $D_1(u'_k)\geqslant\tau_{2k}$ 时,被认为是 k 维致密区域。为保证模型有效,$\tau_{1k}>\tau_{2k}>1$,$k\in[1,d]$。

当 α 取值较大时,牺牲查全率,以取得较高的精准率;当 α 较小时,牺牲精准率,以取得较高的查全率。

第一次划分参数 δ_1 时,根据聚簇尺寸确定,以便发现聚簇,δ_1 划分网格的尺寸应大致与聚类尺寸相当;第二次划分为了进一步确定存在聚类的大区域中聚簇的边界,参数 δ_2 根据对于聚类结果的边界挖掘精度需求确定。两个阶段在同一时刻处理网格规模上限分别不得低于 $(\delta_1)^k$ 和 $(\delta_2)^k$,代表了模型的空间资源需求。因此,第二阶段网格模型要求同时处理网格能力不低于 $\max((\delta_1)^k,(\delta_2)^k)$,原密度意识子空间聚类模型要求同时处理网格能力不得低于 $(\delta_1\delta_2)^k$,而两个模型的实际划分精度均为 $(\delta_1\delta_2)$。

2.2　规模系数的引入

由于只有落入某个第一阶段网格的所有第二阶段网格密度都超过 τ_{2k} 时,其密度才能保证超过 τ_{1k}。实际上致密区域中只有部分的第二阶段网格为致密单元,因此修正式(8)为

$$\tau_{1k} = (\alpha/\beta)D_1(k) = (\alpha/\beta)(N/(\delta_1)^k) \tag{12}$$

当 β 无限大时,τ_{1k} 趋近于 0,第一阶段无法发挥筛选致密区域、淘汰稀薄区域的作用。β 越小,淘汰稀薄区域越多,筛选致密区域越严格,聚类质量越高,但在提高聚类精准率的同时会牺牲一定的查全率。因此,通过调整参数 β,可以控制聚类过程中认定聚类的规模大小。

综上所述,本文模型在降低对内存空间要求的同时,能够通过参数 α 有效地控制聚类强度,通过参数 β 控制聚类的大小,从而抵御无效小聚簇的干扰,实现对聚类结果的控制,提高聚类结果质量。

3　测试评估与性能分析

3.1　测试数据集

为了展示本文改进模型对于原一阶段网格模型的比较优势,合成 10 维数据集 DS1:其中 80% 为随机分布的噪声点,10% 的数据点构成若干随机分布的聚类,10% 的数据点构成大量的随机分布的无效小聚簇,所有有效聚类和无效聚簇均为规则的矩形。

为了对本文描述的密度意识聚类模型进行有效评估,我们采用前述合成数据集,以及从加州大学欧文分校提供的用于机器学习的数据库(http://archive.ics.uci.edu/ml/)选取的真实数据集,来验证本文模型性能。为了更好地证明本文模型的比较优势,采用前文所述 UCI 真实数据集 Daily and Sports Activities Data Set,选取其中若干数据点和维度,构成实验数据空间。对数据空间每一个维度上的数据进行预处理,通过等比拉伸和平移,将数据变化范围统一整理到区间[0,1 000),得到数据集 DS2。

3.2　评价指标

评价聚类效果的指标众多,本文着重考虑在相同划分精度前提下模型在面临小聚簇干扰时的表现。对已知聚类的合成数据集,着重考察模型聚类结果的精准率和查全率,同时考察子空间内聚类区域与非聚类区域的密度。对于未知聚类的真实数据集,考察子空间内聚类区域与非聚类区域的密度。为了研究模型抵御真实数据集中干扰小聚簇的效果,考察输出致密单元的紧密程度。本文考察输出聚类数据点损失与聚害密度的关系,说明高维子空间内无效小聚簇对聚类结果的干扰以及本文模型的有效性。

在合成数据集中,精准率定义为在聚类结果的所有数据点中,实际上属于已知聚类的数据点所占百分比;查全率定义为在所有已知聚类的数据点中,

属于聚类结果中某个聚类的数据点所占百分比。

密度比(D_R):对于任一子空间 S',根据聚类结果分为聚类区域和非聚类区域,聚类区域平均数据密度为 $D_c^{S'}$,非聚类区域平均数据密度为 $D_{non\_c}^{S'}$,则

$$D_R(S') = D_c^{S'} / D_{non\_c}^{S'} \tag{13}$$

指标 $D_R(S')$直观地衡量聚类效果,反映了聚类结果中数据点的集中程度,但是这一指标不能反映有效聚类与干扰小聚簇的差异,在真实数据集中存在一些相对孤立的、尺寸过小的聚簇,在小范围内形成较高的数据密度。有效的聚类不仅要求区域内数据密度较高,还要具备一定的规模和尺寸,这就要求组成聚类的致密单元分布紧凑而不能过于稀疏。因此,本文利用致密单元的紧密程度来反映高维子空间聚类结果中干扰小聚簇的含量。

致密单元紧密程度(D_U):考虑到在高维子空间 S',基于距离的聚类失效,设计一种同样基于网格密度的参数,衡量被确定为致密的小单元之间的紧密程度

$$D_U(S') = C_{2d\_unit} / C_{1d\_unit} \tag{14}$$

将整个数据空间划分成较粗的网格结构,形成若干较大的网格区域,统计落入总计 C_{1d_unit} 个致密区域的 C_{2d_unit} 个致密单元。越多的致密单元分布在越少的区域时,说明致密单元分布更紧凑;相反则说明网格单元分布稀疏。

3.3 实验结果与分析

3.3.1 *合成数据集* 为了体现本文模型的比较优势,在本文模型(TSDENCOS)中两个阶段划分分别采用参数 δ_1、δ_2,在原密度意识子空间聚类模型(DENCOS)中采用参数($\delta_1\delta_2$),以取得相同的实际网格划分精度。在合成数据集 DS1 上对比两个模型聚类结果的精准率与密度比,见表 1。

表 1 *合成数据集 DS1 上的聚类精准率与密度比*

模型	k	精准率/%	密度比
TSDENCOS	8	99.50	308.1
	9	99.35	577.5
	10	95.78	1 182.1
DENCOS	8	97.93	292.1
	9	67.01	400.7
	10	52.37	774.5

可以看出,相比 DENCOS 模型,本文模型始终具有更高的精准率和密度比。在数据集包含干扰小聚簇时,本文模型的精准率、密度比优势更加明显。随着维度上升,DENCOS 逐渐失去抵御小聚簇干扰能力,而本文模型仍然具有良好的聚类精度,能够有效抵御小聚簇干扰。

取 DS1 数据集的 10 维满空间,改变用户输入参数 α,考察 DENCOS 模型与本文模型的精准率-查全率曲线,结果如图 1 所示。

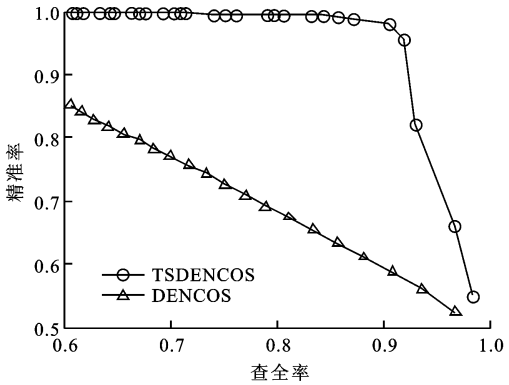
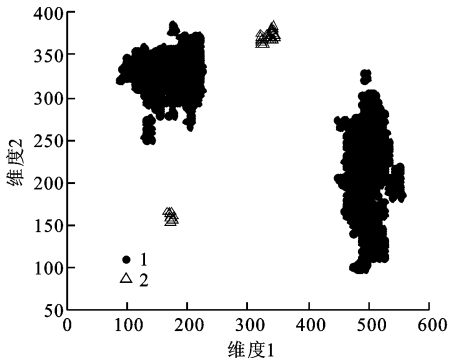


图 1 合成数据集上的精准率-查全率曲线

可以看出,若 α 取值过小,DENCOS 模型与本文模型均取得较高的查全率,但均不具备抵御小聚簇干扰能力;随着 α 取值增大,两个模型均损失一定查全率,同时精准率获得提高,但本文模型精准率提高明显优于 DENCOS 模型。若查全率相同,本文模型具有更高的精准率;若精准率相同,本文模型具有更高的查全率。实验表明,本文模型在面临小聚簇干扰时具有明显优势,而实际数据集通常是高维的,且含有不同程度的干扰小聚簇。

3.3.2 *真实数据集* 在真实数据集上比较 DENCOS 模型与本文模型的聚类质量。在真实数据中抽取 7 维实验数据集 DS2,首先在 DS2 中任选一个二维空间进行聚类,并比较两个模型的结果,以说明在实际数据集中抵御小聚簇干扰的意义,如图 2 所示。由图可见,本文模型能够有效区分干扰小聚簇和有效聚类,在采用高精度网格划分的同时抵御小



1:TSDENCOS 与 DENCOS 的聚类结果重叠部分;
2:DENCOS 聚类结果独有部分

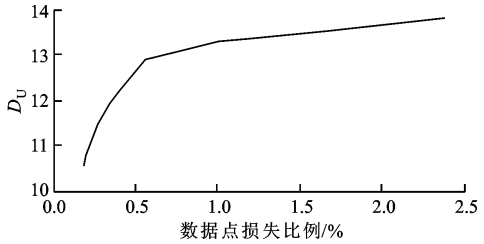
图 2 DS1 数据集二维空间聚类结果

聚簇干扰。

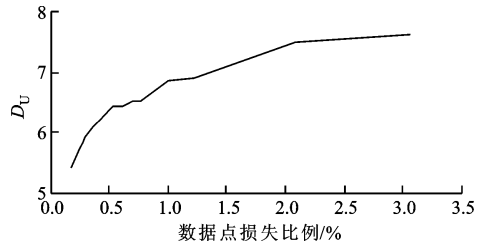
分别考察 DS2 中 4 维、5 维、6 维子空间和 7 维全空间的聚类结果质量，实验结果见表 2。TSPENCOS 模型的聚类结果相比 DENCOS 的聚类结果，在 4～7 维子空间上，数据点依次损失了 0.00%、1.20%、0.58%、0.39%。可以看出，随着维度上升，本文模型发现的致密单元的紧密度明显高于 DENCOS 模型，同时拥有更高的聚类密度 $D_c^{S'}$，也带来小部分数据损失。在较高维子空间，由于数据的稀疏性，干扰小聚簇更多，本文模型能以牺牲更少数据为代价，提高聚类质量。例如在 7 维空间，损失 0.4%数据构成了 64 个小干扰单元，而超过 99%的数据落在了有效聚类的 320 个有效致密单元内部，使得 $D_U(S')$ 提高将近一倍，输出聚类区域密度提高 19.4%。

比较模型聚类结果的 $D_c^{S'}$、$D_U(S')$、$D_R(S')$3 个指标，发现本文模型发现的聚类普遍具有更高的数据密度，组成聚类的致密单元分布更紧凑，拥有更好的聚类质量。但是，考察 $D_R(S')$指标，即聚类区域密度与非聚类区域的密度之比，本文模型相比 DENCOS 模型偏低。原因在于，非聚类区域存在空白网格和低密度网格的规模远大于聚类区域的致密网格，使得非聚类区域密度极低，干扰小聚簇相比这些稀薄或空白区域，在较少的网格中聚集了较多的数据点，明显提高了非聚类区域的平均密度，从而降低了密度比指标。鉴于数据稀疏现象和干扰小聚簇现象在高维更加明显，在衡量高维子空间聚类质量时，密度比参数 $D_R(S')$只能粗略评估聚类区域的数据聚集程度，需要结合另外两项指标才能正确评价聚类质量。

3.3.3　小聚簇对网格聚类的干扰　最后，我们通过一组数据来观察在任意指定的一个高维子空间，随着规模系数 β 的改变，聚类质量的变化，以此进一步说明，在实际的高维数据集子空间聚类过程中，采用本文模型抵御小聚簇干扰的必要性和可行性。图 3 反映了改变规模系数 β（即损失的数据点比例改变）得到的聚类结果。



(a)子空间 1



(b)子空间 2

图 3　数据点损失比例-致密单元紧密度曲线

当规模系数 β 无穷大时，相当于直接进行 $(\delta_1\delta_2)$粒度的网格划分，其效果等同于一阶段划分模型，而致密单元的紧密度很小，聚类质量差；随着规模系数 β 的减小（$\beta>0$），第一阶段滤除的致密单元会越来越多，损失的数据点比例增大，同时输出致密单元的紧密度也提高，如图 3 所示。随着规模系数增大，图中被滤除的数据点所占比例降低到 1%以下时，曲线斜率明显增大，即牺牲很少的数据点即可排除大量孤立的干扰致密单元，大幅提升聚类质量。大量实验表明，真实数据集高维子空间存在大量孤立的致密小单元，它们的数据规模总和很小，却对聚类结果带来极大干扰，造成大量的孤立无效小聚簇，通过本文模型滤除这部分干扰小聚簇，能够有效提升聚类质量。

表 2　真实高维数据集上的聚类结果对比

模型	k	$D_c^{S'}$	$D_U(S')$	$D_R(S')$	u_k 区域数	u'_k 单元数
TSDENCOS	4	6 037.10	1.00	7.05×10^3	7	7
	5	1 824.41	2.30	8.56×10^4	23	30
	6	579.67	5.94	1.19×10^6	17	101
	7	184.43	14.63	1.08×10^7	22	320
DENCOS	4	6 037.12	1.00	7.05×10^3	7	7
	5	1 730.10	2.13	9.29×10^4	25	32
	6	545.25	4.91	1.46×10^6	22	108
	7	154.45	8.21	1.46×10^7	47	384

4 结 语

本文设计了一个两阶段网格划分的密度意识子空间聚类模型,首先找出存在聚类的子空间区域,再在这些区域中确定聚类的高精度边界。模型在每一个阶段的网格规模远小于现有的密度意识子空间聚类模型在同精度下的网格规模,同时避免了引入干扰小聚簇,提高了聚类质量。在合成数据集以及真实数据集上的实验表明,本文模型有效提高了聚类质量,尤其在高维数据空间,面对小聚簇干扰时优势明显。

参考文献:

[1] AGGARWAL C C, WOLF J L, YU P S, et al. Fast algorithms for projected clustering [J]. ACM SIGMOD Record, 1999, 28(2): 61-72.

[2] AGGARWAL C C, YU P S. Finding generalized projected clusters in high dimensional spaces [J]. ACM SIGMOD Record, 2000, 29(2): 70-81.

[3] AGGARWAL R, GEHRKE J, GUNOPULOS D, et al. Automatic subspace clustering of high dimensional data for data mining applications [J]. ACM SIGMOD Record, 1998, 27(2): 94-105.

[4] CHENG Chun-hung, FU Ada Wai-chee, ZHANG Yi. Entropy-based subspace clustering for mining numerical data [C]// Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 1999: 84-93.

[5] NAGESH H, GOIL S, CHOUDHARY A. Adaptive grids for clustering massive data sets [C]// Proceedings of the first SIAM International Conference on Data Mining, Philadelphia, PA, USA: SIAM, 2001: 1-17.

[6] KAILING K, KRIEGEL H P, KROGER P. Density-connected subspace clustering for high-dimensional data [C]// Proceedings of the 2004 SIAM International Conference on Data Mining. Philadelphia, PA, USA: SIAM, 2004: 246-256.

[7] JIAO Hongzan, ZHONG Yanfei, ZHANG Liangpei, et al. Subspace clustering based on decision fusion strategy for hyperspectral imagery [C]// Proceedings of the 2013 IEEE International Digital Object on Geoscience and Remote Sensing Symposium. Piscataway, NJ, USA: IEEE, 2013: 1485-1488.

[8] IAM-ON N, BOONGOEN T. New soft subspace method to gene expression data clustering [C]// Proceedings of the 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics. Piscataway, NJ, USA: IEEE, 2012: 984-987.

[9] HECKEL R, BOLCSKEI H. Noisy subspace clustering via thresholding [C]// Proceedings of the 2013 IEEE International Symposium on Information Theory. Piscataway, NJ, USA: IEEE, 2013: 1382-1386.

[10] 朱林, 雷景生, 毕忠勤, 等. 一种基于数据流的软子空间聚类算法 [J]. 软件学报, 2013, 24(11): 2610-2627.
ZHU Lin, LEI Jingsheng, BI Zhongqin, et al. Soft subspace clustering algorithm for streaming data [J]. Journal of Software, 2013, 24(11): 2610-2627.

[11] 夏虎, 庄健, 于德弘. 面向高维特征故障数据的进化软子空间聚类算法 [J]. 西安交通大学学报, 2013, 47(5): 115-120.
XIA Hu, ZHUANG Jian, YU Dehong. Evolutionary soft subspace clustering method for fault diagnosis of high-dimensional features [J]. Journal of Xi'an Jiaotong University, 2013, 47(5): 115-120.

[12] AGGARWAL C C, HINNEBURG A, KEIM D A. On the surprising behavior of distance metrics in high dimensional space [M]. Berlin, Germany: Springer-Verlag, 2001: 420-434.

[13] AGGARWAL C C, Yu P S. The IGrid index: reversing the dimensionality curse for similarity indexing in high dimensional space [C]// Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2000: 119-129.

[14] HINNEBURG A, AGGARWAL C C, KEIM D A. What is the nearest neighbor in high dimensional spaces? [M]. San Mateo, CA, USA: Morgan Kaufmann, 2000: 506-515.

[15] CHU Yi-hong, HUANG Jen-wei, CHUANG Kun-ta, et al. Density conscious subspace clustering for high-dimensional data [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(1): 16-30.

[16] SARMAH S, BHATTACHARYYA D K. A grid-density based technique for finding clusters in satellite image [J]. Pattern Recognition Letters, 2012, 33(5): 589-604.

[17] LEI G, YU X, YANG X, et al. An incremental clustering algorithm based on grid [C]// Proceedings of the 2011 8th International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway, NJ, USA: IEEE, 2011: 1099-1103.

(编辑 武红江)