



文章编号:1001-9081(2018)10-2844-06

DOI:10.11772/j.issn.1001-9081.2018020375

基于概率模型的非均匀数据聚类算法

杨天鹏¹, 陈黎飞^{1,2*}

(1. 福建师范大学 数学与信息学院, 福州 350117; 2. 福建师范大学 数字福建环境监测物联网实验室, 福州 350117)

(*通信作者电子邮箱 clfei@fjnu.edu.cn)

摘要:针对传统 K -means 型算法的“均匀效应”问题,提出一种基于概率模型的聚类算法。首先,提出一个描述非均匀数据簇的高斯混合分布模型,该模型允许数据集中同时包含密度和大小存在差异的簇;其次,推导了非均匀数据聚类的目标优化函数,并定义了优化该函数的期望最大化(EM)型聚类算法。分析结果表明,所提算法可以进行非均匀数据的软子空间聚类。最后,在合成数据集与实际数据集上进行的实验结果表明,所提算法有较高的聚类精度,与现有 K -means 型算法及基于欠抽样的算法相比,所提算法获得了 5%~50% 的精度提升。

关键词:聚类;概率模型;非均匀数据;均匀效应

中图分类号:TP311 **文献标志码:**A

Probability model-based algorithm for non-uniform data clustering

YANG Tianpeng¹, CHEN Lifei^{1,2*}

(1. College of Mathematics and Informatics, Fujian Normal University, Fuzhou Fujian 350117, China;

2. Digital Fujian Internet-of-Things Laboratory of Environmental Monitoring, Fujian Normal University, Fuzhou Fujian 350117, China)

Abstract: Aiming at the “uniform effect” of the traditional K -means algorithm, a new probability model-based algorithm was proposed for non-uniform data clustering. Firstly, a Gaussian mixture distribution model was proposed to describe the clusters hidden within non-uniform data, allowing the datasets to contain clusters with different densities and sizes at the same time. Secondly, the objective optimization function for non-uniform data clustering was deduced based on the model, and an EM (Expectation Maximization)-type clustering algorithm defined to optimize the objective function. Theoretical analysis shows that the new algorithm is able to perform soft subspace clustering on non-uniform data. Finally, experimental results on synthetic datasets and real datasets demonstrate that the accuracy of the proposed algorithm is increased by 5% to 50% compared with the existing K -means-type algorithms and under-sampling algorithms.

Key words: clustering; probability model; non-uniform data; uniform effect

0 引言

聚类分析作为数据挖掘的一种重要方法,目的是将给定数据划分成多个子集(每个子集为一个簇),使得簇内对象彼此相似,与其他簇对象不相似^[1]。传统的聚类算法可分为层次聚类、基于划分聚类、基于密度和网格聚类,以及其他聚类算法^[2-3]。目前聚类分析已广泛应用在 Web 搜索、图像处理、模式识别、医疗数据分析等众多领域。

作为数据挖掘十大算法之一, K -means 算法^[4]因其简单高效的优点得到广泛的研究和应用^[5]。然而,受“均匀效应(uniform effect)”的影响^[6], K -means 型算法在聚类医疗诊断等复杂数据时性能受限。这类数据的一个特点是同一数据集同时包含了样本数量和样本密度有较大差异的簇,这种数据称为非均匀数据(non-uniform data)。与不平衡数据(主要指簇样本量即簇大小差异较大的数据)聚类^[7]相比,非均匀数据聚类问题更具普遍性。例如,在含有“正常”和“患病”两个簇的疾病诊断数据中,两簇的大小差异明显(通常,“正常”簇比“患病”簇的样本数量大得多),更重要地,“患病”簇的样本

皆具特定的疾病模式,其密度比“正常”簇有显著区别(表现为“正常”簇样本分布的方差大得多)。

针对该问题研究者提出了多种方法^[8-12],可大致分为三类:第一类方法基于样本抽样,在聚类之前首先对样本集作欠采样或过采样的处理操作,文献[8-9]即是在这样预处理后的数据上进行 K -means 聚类的;第二类方法在聚类模型中考虑不同簇的样本量差异,例如,文献[10]引入簇的样本数量,给出了经典模糊聚类算法目标优化函数的两种改进方案;第三类法则侧重簇的密度差异,借助多代表点等方法^[11]以区分数据集中的不同密度区域。这些方法是分别针对簇样本数量不平衡特性或密度差异特性而提出的,未提供同时处置非均匀数据上述两个特性的解决方案。

从原理上说, K -means 型聚类是一种基于模型的方法,它所学习的概率模型是以相关参数为常数这一假设前提下的一种简化的高斯混合模型^[13],此简化模型并不能很好地刻画非均匀数据簇类的两个特点。为此,本文提出一种基于概率模型的非均匀数据聚类新算法——MCN(Model-based Clustering on Non-uniform data),以应对传统 K -means 型算法的“均匀效

收稿日期:2018-02-12;修回日期:2018-03-31;录用日期:2018-04-16。

基金项目:国家自然科学基金资助项目(61672157);福建师范大学创新团队项目(IRTL1704)。

作者简介:杨天鹏(1991—),男,湖北十堰人,硕士研究生,主要研究方向:数据挖掘; 陈黎飞(1972—),男,福建长乐人,教授,博士,主要研究方向:统计机器学习、数据挖掘、模式识别。



应”问题。本文的主要工作包括两个方面:其一,以高斯混合模型为基础,建立了非均匀数据簇的概率模型,新模型可以描述同一数据集中样本量和密度都存在差异的簇;其二,基于提出的模型推导了聚类目标函数,并给出优化目标函数的算法步骤,实现了非均匀数据的软子空间聚类。在合成数据和实际数据上的实验结果表明,与现有的非均匀数据聚类算法相比,本文 MCN 算法有效提高了聚类精度。

1 相关工作

首先给出文中使用的符号及定义。令待聚类数据集为 DB ,含 N 个 D 维样本,任一样本用 $\mathbf{x} = \langle x_1, x_2, \dots, x_j, \dots, x_D \rangle$ 表示,其第 $j(j=1, 2, \dots, D)$ 维属性为 x_j 。考虑硬聚类算法,它将 DB 划分成 K 个不相交的子集的集合 $C = \{c_1, c_2, \dots, c_k, \dots, c_K\}$,并称子集 c_k 为 DB 的第 $k(k=1, 2, \dots, K)$ 个簇, $|c_k|$ 表示该簇包含的样本数量。用 $\mathbf{v}_k = \langle v_{k1}, v_{k2}, \dots, v_{kD} \rangle$ 表示 c_k 的簇中心, $V = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_K\}$ 为全体簇中心的集合。

经典的 K -means 算法是一种划分型聚类算法,其优化目标定义为:

$$\min J_0(C, V) = \sum_{k=1}^K \sum_{x \in c_k} \|\mathbf{x} - \mathbf{v}_k\|^2 \quad (1)$$

K -means 通过类期望最大化(Expectation Maximization, EM)算法^[15]的学习过程求取式(1)的局部优解,过程如下:给定簇数目 K ,首先选择 K 个初始簇中心,然后计算每个样本与各簇中心点的距离,将样本划分至距离最小的簇,再为每个新划分生成的簇计算最优的簇中心;算法迭代执行上述“划分-簇中心优化”步骤,直到满足停止条件算法终止,得到对应式(1)局部优解的数据集聚类划分。

文献[6]分析了 K -means 聚类的“均匀效应”现象。以聚类图 1(a) 中的非均匀数据为例。图 1(a) 隐含有 3 个簇 Cluster1、Cluster2 和 Cluster3,它们不但在大小(样本数)上有差异,簇密度也显著不同,例如,Cluster1 和 Cluster2 中样本分布方差显然有较大差别。该数据的 K -means 聚类结果如图 1(b) 所示,其中样本数较少的 Cluster2 会“吞掉”样本较多的簇 Cluster1 的部分样本,使得两个簇的大小和密度趋向于

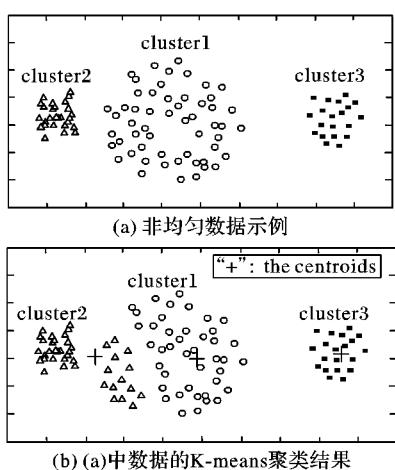


图 1 “均匀效应”的例子
Tab. 1 An example of “uniform effect”

相同,此即 K -means 型算法的“均匀效应”。

从统计学习^[16]的角度, K -means 可以看作是一种基于模型的统计聚类算法。这里,视簇 c_k 的每个样本 \mathbf{x} 源自如下高斯分布:

$$\text{Gauss}(\mathbf{x}; \mathbf{v}_k, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|\mathbf{x} - \mathbf{v}_k\|_2^2}{2\sigma^2}\right) \quad (2)$$

那么,给定数据集 DB ,划分聚类的目标就是搜索最小化下面负对数似然函数的模型参数 (C, V) :

$$\begin{aligned} \underset{(C, V)}{\operatorname{argmin}} & - \sum_{k=1}^K \sum_{x \in c_k} \ln \text{Gauss}(\mathbf{x} | \mathbf{v}_k, \sigma) = \\ & \underset{(C, V)}{\operatorname{argmin}} \sum_{k=1}^K \sum_{x \in c_k} \ln(\sqrt{2\pi}\sigma) + \sum_{k=1}^K \sum_{x \in c_k} \frac{\|\mathbf{x} - \mathbf{v}_k\|_2^2}{2\sigma^2} = \\ & \underset{(C, V)}{\operatorname{argmin}} \sum_{k=1}^K \sum_{x \in c_k} \|\mathbf{x} - \mathbf{v}_k\|_2^2 \end{aligned} \quad (3)$$

注意到式(3)的推导结果与 K -means 算法的优化目标是相同的,见式(1)。

上面推导过程基于如下基本假设:每个簇的样本方差 σ 是一个常数。如前所述, σ 体现了簇的密度。这从模型的角度解释了“均匀效应”产生的一个原因: K -means 型算法致力于求解密度相近的簇集合。此外,从式(3)还可以看出, K -means 算法的优化目标也没有体现不同簇中样本数量的差异,这也是其所假设的概率模型所决定的:对应不同簇的高斯分布分量以一种“平等”的方式进行混合建模。因此,为提高 K -means 型算法在非均匀数据上的聚类性能,下面首先提出一种新的高斯混合模型,以区分簇类在样本数量和密度上的差异;接着,以此为基础,推导出一种新型的非均匀数据聚类算法。

2 非均匀数据聚类模型及算法

本章首先建立用于非均匀数据聚类的高斯混合模型,然后定义基于模型的聚类目标优化函数,最后给出聚类算法。

2.1 非均匀数据聚类模型

如前所述,在一个非均匀数据集中,簇的密度通常存在差异。为刻画这种差异,引入两组记号:用 $\sigma_k^2(k=1, 2, \dots, K)$ 表示簇 c_k 的方差,其值越大,表明 c_k 的密度越小;进一步,引入向量 $\mathbf{w}_k = \langle w_{k1}, w_{k2}, \dots, w_{kj}, \dots, w_{kD} \rangle$,其各元素 $w_{kj} > 0$,用于区分簇 c_k 在不同属性上的密度差异,值越大表明 c_k 投影在相应属性上时数据分布的密度越小。由此, c_k 属性 j 上数据分布的方差可用 σ_k^2/w_{kj} 来表示。将这个方差表达式代入形如式(2)的高斯密度函数,得到任意样本 $\mathbf{x} \in c_k$ 投影在属性 j 上的概率密度函数,如下:

$$p(x_j; \mathbf{v}_{kj}, w_{kj}, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k/\sqrt{w_{kj}}} \exp\left(-\frac{(x_j - v_{kj})^2}{2\sigma_k^2/w_{kj}}\right) \quad (4)$$

在此基础上,基于数据集的 D 个属性是统计独立的这一“朴素”假设^[17]来建立簇的模型。虽然该假设在一些实际数据上并不现实,但它可以有效降低所构造模型的复杂性:简单地通过一组变量边缘分布的乘积来估计向量的概率密度。这样,令 $P(\mathbf{x})$ 表示 c_k 中任一样本的概率密度,有:



$$P(\mathbf{x}; \mathbf{v}_k, \mathbf{w}_k, \sigma_k) = \prod_{j=1}^D p(x_j; v_{kj}, w_{kj}, \sigma_k) \quad (5)$$

接下来,考虑非均匀数据的另一个特性:同一数据可能包含大小各异的簇。为此,引入代表簇大小的记号 $\alpha_k (k = 1, 2, \dots, K)$, 满足约束条件:

$$\forall k: \alpha_k > 0, \sum_{k=1}^K \alpha_k = 1 \quad (6)$$

其数值大小与簇所包含的样本数量相关,可以看作是赋予每个簇的一种权重。根据这些定义,非均匀数据的加权似然函数表示为:

$$L(\Theta) = \prod_{k=1}^K \prod_{x \in c_k} \alpha_k \times P(\mathbf{x}; \mathbf{v}_k, \mathbf{w}_k, \sigma_k) \quad (7)$$

其中: $\Theta = \{(c_k, \sigma_k, \mathbf{v}_k, \mathbf{w}_k) | k = 1, 2, \dots, K\}$ 为 K 组参数的集合。

基于上述模型,给定数据集 DB 和簇数 K ,聚类转变成了从 DB 求取优化的参数 Θ 以最大化加权似然的问题:

$$\max J_1(\Theta) = \sum_{k=1}^K |c_k| \ln \alpha_k + \sum_{k=1}^K \sum_{x \in c_k} \ln P(x; \mathbf{v}_k, \mathbf{w}_k, \sigma_k)$$

上式在式(7)基础上使用了对数变换,受条件式(6)约束。代入式(4)和(5),并略去其中的常数项,优化目标改写为:

$$\begin{aligned} \min J_2(\Theta) = & - \sum_{k=1}^K |c_k| \ln \alpha_k + \frac{1}{2} \sum_{k=1}^K |c_k| \sum_{j=1}^D \ln \frac{\sigma_k^2}{w_{kj}} + \\ & \frac{1}{2} \sum_{k=1}^K \sum_{j=1}^D \sum_{x \in c_k} \frac{w_{kj}(x_j - v_{kj})^2}{\sigma_k^2} \end{aligned} \quad (8)$$

对比式(1)可知:

1) 当所有的 α_k, σ_k 和 w_{kj} 都为常数, J_2 退化为 K -means 算法的优化目标函数 J_0 。这意味着 K -means 假定了所有簇具有相同的大小和相同的方差,且各簇每个属性上的数据分布密度也是相同的。而新的目标函数通过 σ_k, \mathbf{v}_k 和 \mathbf{w}_k 等参数可以区分簇类这些各异的特性;

2) 在 J_2 表达式中, w_{kj} 主要作用于 x_j 与 v_{kj} 间距离(实际上是二者间的平方误差,数值上等于二者欧氏距离值的平方)的计算。从效果上看,衡量属性密度差异的 $w_{kj} (j = 1, 2, \dots, D)$ 相当于赋予各属性的特征权重,其数值大小反映了各属性对距离度量的贡献程度。因此,优化 J_2 的过程可以看作是对非均匀数据集实施的软子空间聚类^[14]。

2.2 软子空间聚类算法

传统的软子空间聚类算法中对 w_{kj} 的约束条件通常设定为 $\sum_{j=1}^D w_{kj} = 1$, 其中 $0 \leq w_{kj} \leq 1 (j = 1, 2, \dots, D)$ 。这是一种“软”特征选择方法,可以较好地避免“硬”特征选择带来的样本特性信息丢失问题,但容易产生“平凡解”(例如,仅一个特征的权重为 1,其他特征的为 0),影响软子空间聚类的效果。为此,本文采用另外一种特征权重约束条件 $\prod_{j=1}^D w_{kj} = 1$, 其中 $w_{kj} > 0 (j = 1, 2, \dots, D)$ ^[18]。这样,当特征与簇关联较小时,其特征权重将获得较小值(例如取值为 0.001),当特征与簇有较大关联时可以获得较大的权重(例如 100)。这一约束

条件有效放大了特征之间的差异,利于区分不同特征的贡献度。

根据拉格朗日乘子法,将 w_{kj}, α_k 的约束条件引入到目标函数中,可得带约束条件的聚类优化目标函数为:

$$J_3(\Theta) = J_2(\Theta) + \sum_{k=1}^K \lambda_k \left(1 - \prod_{j=1}^D w_{kj} \right) + \eta \left(1 - \sum_{k=1}^K \alpha_k \right) \quad (9)$$

其中: λ_k 和 η 为拉格朗日乘子。

上述目标函数参数的求解是非线性函数的优化问题,难以求得全局最优解。本文 MCN 算法基于常用的 EM 算法结构求取其局部最优解。为叙述方便,引入符号 $W = \{w_{kj} | k = 1, 2, \dots, K; j = 1, 2, \dots, D\}$ 和 $A = \{\alpha_1, \alpha_2, \dots, \alpha_K, \sigma_1, \sigma_2, \dots, \sigma_K\}$ 。参数的求解可分为以下几个步骤:

1) 固定 W, V, A ,求 C 。对任意一个样本 x 根据以下公式进行簇划分:

$$\begin{cases} z = \arg \max_k \alpha_k G_k(x) \\ G_k(x) = \prod_{j=1}^D \frac{\sqrt{w_{kj}}}{\sqrt{2\pi} \sigma_k} \exp\left(-\frac{w_{kj}(x_j - v_{kj})^2}{2\sigma_k^2}\right) \end{cases} \quad (10)$$

式(10)通过比较样本 x 源自各高斯分量的概率将其划分到概率最大的簇中。

2) 固定 W, V, C ,求 A 。首先将 σ_k 固定,令 $\frac{\partial J_3}{\partial \alpha_k} = 0$ 得解为:

$$\alpha_k = |c_k| / N \quad (11)$$

其中: $|c_k|$ 为簇中样本数目; N 为样本总数。式(11)与簇中样本数量成正比,可以反映非均匀数据中不同簇样本数量存在差异的特点。然后,固定 α_k ,令 $\frac{\partial J_3}{\partial \sigma_k} = 0$ 可得:

$$\sigma_k^2 = \frac{1}{D \times |c_k|} \sum_{j=1}^D \sum_{x \in c_k} w_{kj} (x_j - v_{kj})^2 \quad (12)$$

从式(12)可知, σ_k^2 即是第 k 个簇中样本分布的加权散度,反映了非均匀数据中各簇有差异的密度信息。根据以上分析,算法的最优解 α_k 和 σ_k^2 能刻画非均匀数据中不同簇之间样本数量和密度都可能存在差异的特点。

3) 固定 W, A, C ,求 V 。令 $\frac{\partial J_3}{\partial v_{kj}} = 0$ 可得:

$$v_{kj} = \sum_{x \in c_k} \frac{x_j}{|c_k|} \quad (13)$$

式(13)为簇中心点求解公式,通过该式完成簇中心点的更新。

4) 固定 A, C, V ,求 W 。令 $\frac{\partial J_3}{\partial w_{kj}} = 0$ 可得:

$$w_{kj} = (|c_k| + \lambda_k) \frac{\sigma_k^2}{X_{kj}} \quad (14)$$

其中: $X_{kj} = \sum_{x \in c_k} (x_j - v_{kj})^2$, $\lambda_k = \left(\frac{1}{\sigma_k^2} \left(\prod_{j=1}^D \frac{1}{X_{kj}} \right)^{\frac{1}{D}} \right) - |c_k|$ 。

式(14)通过求解 w_{kj} 为各特征赋予不同的权重,效果上相当于将第 k 个簇的样本投影到相应的子空间中。

根据上述参数求解方法,可以得到基于概率模型的非均匀数据软子空间聚类算法如下。



输入 数据集 DB ,簇数目 K 。
 输出 簇划分 C 。
 初始化 随机生成初始簇中心 v_k ,并令 $w_{kj} = 1/D$, $\sigma_k = 1/K$,
 $\alpha_k = 1/K(k = 1, 2, \dots, K; j = 1, 2, \dots, D)$ 。
 Repeat:
 更新 C :利用式(10)更新簇划分;
 更新 v_{ij} :根据式(13),更新 v_{ij} ;
 更新 α_k 、 σ_k :根据式(11)、(12)更新 α_k 、 σ_k ;
 计算 w_{kj} :先计算 λ_k ,并将求得的 λ_k 代入到式(14)中求得 w_{kj} ;
 Until:满足迭代停止条件
 根据上述算法步骤可知本文 MCN 算法的时间复杂度为 $O(PKND)$,其中 P 为算法的迭代次数。

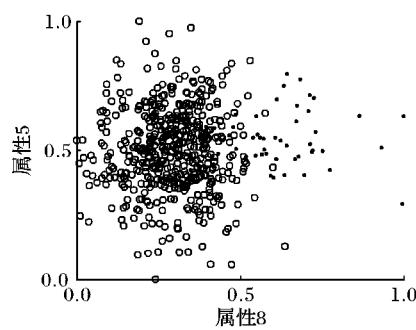
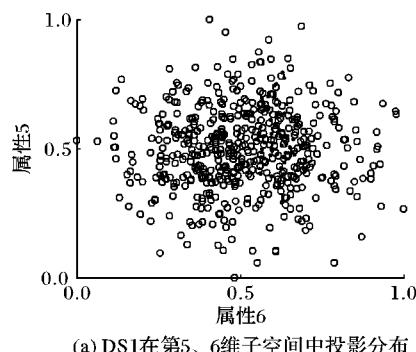
3 实验

实验平台为:Core i5-3470 3.2 GHz CPU,4 GB 内存,操作系统为 Windows 7。算法采用 Java 编写。

3.1 实验设置

实验选择了 GMM^[16]、Verify2^[19]、IFCM^[10]三种算法进行对比。GMM 作为基于概率模型的典型聚类算法,将其作为对比算法用来验证经典的概率模型和结合子空间技术的概率模型在非均匀数据上的表现;Verify2 为文献[19]提出的一种将欠采样和谱聚类结合对类不平衡数据进行聚类分析的方法,其中欠采样是非均匀数据预处理方法中的一种代表性方法;IFCM 为文献[10]中提出的基于样本数量加权的模糊聚类算法。

因为非均匀数据不同簇之间样本存在较大差异,合成数据能够从簇的数目、大小等控制数据集的簇结构,便于分析算法的性能及算法性能与簇结构之间的关系。首先在多个合成数据上进行测试,然后在 4 个真实数据上实验。由于各数据集已知类标签,选择两个外部评价指标 Macro-F1^[13] 和标准化



互信息(Normalized Mutual Information, NMI)^[20]来评估各种算法的聚类性能,指标的值越大表明聚类效果越好。

$$\text{Macro-F1} = \sum_{1 \leq k \leq K} F1(class_k)/K$$

$$F1(class_k) = \frac{2 \times R(class_k, c_i) \times P(class_k, c_i)}{R(class_k, c_i) + P(class_k, c_i)}$$

其中: $F1(class_k)$ 为第 k 个簇的 F1 值; $P(class_k, c_i)$ 和 $R(class_k, c_i)$ 分别表示数据集中真实的类 $class_k$ 与聚类结果中簇 c_i 相比的准确率和召回率; $class_k$ 表示数据集中第 k 个真实的类; n_k 表示 $class_k$ 包含的样本点数。

NMI 的计算公式如下:

$$NMI = \frac{\sum_{i=1}^R \sum_{j=1}^K n_{ij} \ln((N \times n_{ij}) / (n_i \times n_j))}{\sqrt{\sum_{i=1}^R n_i \ln(n_i/N) \times \sum_{j=1}^K n_j \ln(n_j/N)}}$$

其中: n_{ij} 表示真实数据集中类 i 与聚类结果中簇 j 相一致的样本点数目; n_i 表示属于类 i 的样本点数目; n_j 表示属于簇 j 的样本点数目; R 表示真实类别数,实验中设定 $K = R$ 。

3.2 合成数据实验结果

实验中利用 numpy 中的 random.multivariate_normal() 函数合成三个数据集。由于二类数据可以直观表现簇结构,因此,在合成数据时,将簇数目固定为两类;此外,使用方差 σ 测量各簇中样本的分布散度。合成数据的主要参数如表 1 所示。如表 1 所示,三个合成数据集的样本数量逐个递增,以此来验证本文 MCN 算法在不同数据量下的性能表现;同时,注意到同个数据集不同簇之间样本数量和样本方差都有较大差异。三个合成数据集的数据维度也逐个递增,以此测试不同数据维度下各算法的性能。

为直观地展现合成数据中样本的分布情况,将 DS1 投影

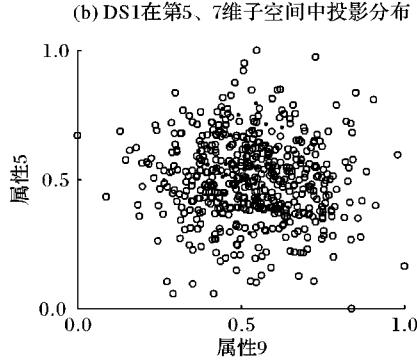
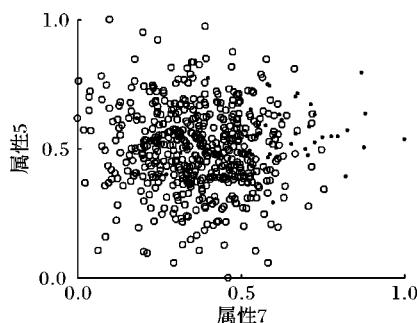


图 2 DS1 投影到部分低维空间中的数据分布

Tab. 2 Distribution of DS1 projected on some low-dimensional spaces



到部分维度所确定的低维空间中,投影结果如图 2 所示。从图 2 可知,DS1 中的多数类(样本数量较多的簇)的数据分布较为分散,少数类的分布则较为集中,且两个簇存在交叠现象。

表 1 合成数据集参数

Tab. 1 Parameters of synthetic datasets

数据集	簇大小	维度	方差
DS1	500;50	10	0.21;0.14
DS2	2000;100	50	0.90;0.64
DS3	5000;200	100	1.64;1.34

表 2 合成数据集不同算法聚类结果

Tab. 2 Clustering results of different algorithms on synthetic datasets

度量标准	数据集	MCN	GMM	Verify2	IFCM
Macro-F1	DS1	0.9665 ± 0.015	0.5595 ± 0.000	0.4238 ± 0.001	0.5595 ± 0.000
	DS2	0.9523 ± 0.023	0.5333 ± 0.000	0.4129 ± 0.004	0.5333 ± 0.000
	DS3	0.9351 ± 0.034	0.5143 ± 0.000	0.4470 ± 0.010	0.5272 ± 0.000
NMI	DS1	0.9335 ± 0.059	0.0000 ± 0.000	0.0057 ± 0.001	0.0000 ± 0.000
	DS2	0.9150 ± 0.073	0.0000 ± 0.000	0.0144 ± 0.003	0.0000 ± 0.000
	DS3	0.8954 ± 0.089	0.0000 ± 0.000	0.0318 ± 0.005	0.0000 ± 0.000

注:粗体部分为各数据集上最高的指标值。

不同算法在合成数据上的运行时间如表 3 所示。表 3 中,本文 MCN 算法的运行时间低于对比算法 GMM、Verify2 和 IFCM。Verify2 的运行时间远高于 GMM 和 MCN 算法,一个主要原因是 Verify2 采用了谱聚类方法,涉及到矩阵特征值计算等,当样本数量和数据维度较大时,其算法运行时间较长。

表 3 不同算法在合成数据上的运行时间

Tab. 3 Running time of different algorithms on synthetic datasets /s

数据集	MCN	GMM	Verify2	IFCM
DS1	0.0016	0.0041	1.0082	0.0165
DS2	0.0092	0.0131	1.2898	0.0811
DS3	0.0420	0.0633	5.5024	0.3501

3.3 实际数据实验结果

实验使用的实际数据来自聚类分析常用的 UCI Machine Learning Repository (<http://Archive.ics.uci.edu/ml/datasets.html>)。选用了四个实际数据集:Breast Cancer Wisconsin(简写为 BCW)、Wine、ForestType 和 Ionosphere,数据集主要参数如表 4 所示。其中,BCW 为乳腺癌诊断数据,包含 241 个恶性样本和 458 个良性样本;Wine 是相关研究常用的不平衡数据集,其普通品质酒类的样本数较多,而品质较好和品质较差

表 2 显示不同算法在合成数据集上取得的聚类结果。如表所示,本文 MCN 算法的聚类精度和 NMI 值都优于对比算法,表明 MCN 能更好地聚类此型非均匀数据。GMM 算法在三个合成数据集上的 NMI 值均为 0,这是因为 GMM 算法将数据中的所有样本都划分到同一个簇中,侧面反映了基于经典高斯模型的方法并不能有效处理非均匀数据。在两个指标上,IFCM 算法与 GMM 接近。Verify2 的聚类精度最低,但与 GMM 和 IFCM 算法相比,其 NMI 值有一定的提升,表明基于样本抽样的方法对非均匀数据聚类效果的改善有限。

表 2 合成数据集不同算法聚类结果

Tab. 2 Clustering results of different algorithms on synthetic datasets

的样本数量则较少;ForestType 是森林遥感数据,包含三种不同的森林类型和一类空地,其中 Sugi forest 类的样本数量较多;Ionosphere 为电离层雷达波数据,其中具有某种特定结构的样本数量较多。这四个数据集中,不同簇类的样本数有较大差异,且样本分布(方差)也存在差异,是典型的非均匀数据。实验将基于 BCW、Wine 数据集验证各种算法在低维数据上的性能,在 ForestType、Ionosphere 上对比算法在较高维度数据上的表现。本文 MCN 算法与对比算法在四个实际数据上的聚类结果如表 5 所示。表 5 显示,MCN 算法在 Wine 数据上的两项指标稍低于 IFCM 算法,但在其他数据集上的聚类精度和 NMI 值都明显优于对比算法,表明 MCN 算法可以有效聚类实际应用中的非均匀数据。

表 4 实际数据集参数

Tab. 4 Parameters of real-world datasets

数据集	簇大小	维度	方差
BCW	241;458	9	0.13;0.83
Wine	59;71;48	13	0.20;0.38;0.26
ForestType	158;86;83;195	27	0.23;0.19;0.58;0.20
Ionosphere	225;126	34	1.55;3.56

表 5 算法在实际数据集上的聚类结果

Tab. 5 Clustering results of different algorithms on real-world datasets

度量标准	数据集	MCN	GMM	Verify2	IFCM
Macro-F1	BCW	0.9131 ± 0.007	0.6520 ± 0.000	0.5139 ± 0.002	0.6346 ± 0.000
	Wine	0.9103 ± 0.009	0.5129 ± 0.002	0.4076 ± 0.001	0.9279 ± 0.001
	ForestType	0.6951 ± 0.010	0.3927 ± 0.000	0.3094 ± 0.001	0.5535 ± 0.004
	Ionosphere	0.7055 ± 0.001	0.6548 ± 0.000	0.5186 ± 0.001	0.6399 ± 0.000
NMI	BCW	0.6515 ± 0.042	0.0071 ± 0.000	0.0074 ± 0.002	0.1628 ± 0.001
	Wine	0.7725 ± 0.018	0.0509 ± 0.007	0.0126 ± 0.000	0.7762 ± 0.002
	ForestType	0.4856 ± 0.011	0.0064 ± 0.001	0.0113 ± 0.000	0.3416 ± 0.004
	Ionosphere	0.2163 ± 0.003	0.0005 ± 0.000	0.0042 ± 0.000	0.0452 ± 0.001

注:粗体部分为各数据集上最高的指标值。



如前所述,本文提出的 MCN 算法可以实现非均匀数据的子空间聚类,实现途径是在聚类过程中自动地赋予每个特征以不同的权重。下面以 Wine 数据集为例,从 MCN 算法的一次聚类结果中提取特征权重信息,作进一步分析。图 3 显示该数据集中三个簇(分别记为 Cluster1、Cluster2 和 Cluster3)各自的 13 个特征(分别命名为 A1,A2,⋯,A13)的权重分布。

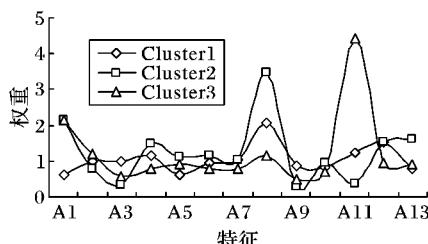


图 3 Wine 数据中三个簇的特征权重分布

Fig. 3 Distribution of feature weights of three clusters in dataset Wine

如图 3 所示,不同簇的特征权重分布并不相同。例如,对于 Cluster3,MCN 算法赋予 A11(指“酒的色调”)较大的权重,这表明“色调”对识别 Cluster3 有重要的作用;而特征 A8(一种称为“Nonflavonoid phenols”的酚类化学物质)对 Cluster2 中酒的品质有较大影响。以上结果表明,MCN 算法可以有效识别特征对于不同簇类有差别的贡献度,从而提高了实际应用中非均匀数据聚类的性能。

4 结语

针对 K-means 型算法的“均匀效应”问题,本文提出了 MCN 算法。首先分析了经典 K-means 算法隐含使用的概率模型,它是基于有关参数为常数这一假设的高斯混合模型,此简化模型并不能很好地刻画非均匀数据簇之间样本数量和密度都有较大差异的特点。接着,从概率模型角度入手,结合软子空间聚类技术定义了一种非均匀数据簇的概率模型,并推导出了相应的聚类优化目标函数。最后给出了 MCN 的算法过程。在合成数据和实际数据上的实验结果表明,与 GMM、Verify2、IFCM 等算法相比,MCN 算法在多数情况下都可以取得较大的聚类性能提升,从而验证了本文所提算法的有效性。

在大数据时代如何结合大数据处理工具分析非均匀数据是一项有意义的工作,因此下一步将结合分布式 Spark 平台进一步研究非均匀数据聚类新方法。

参考文献(References)

- [1] 韩家炜,坎伯 M,裴健. 数据挖掘:概念与技术[M]. 3 版. 范明, 孟小峰,译. 北京: 机械工业出版社, 2012: 288. (HAN J W, KAMER M, PEI J. Data Mining: Concepts and Techniques [M]. 3rd ed. FAN M, MENG X F, translated. Beijing: China Machine Press, 2012: 288.)
- [2] BERKHIN P. A survey of clustering data mining techniques [M]// KOGAN J, NICHOLAS C, TEBOLLE M. Grouping Multidimensional Data. Berlin: Springer, 2002: 25–71.
- [3] AGGARWAL C C, REDDY C K. Data Clustering: Algorithms and Applications [M]. Boca Raton: Chapman and Hall/CRC, 2013: 3–15.
- [4] HARTIGAN J A, WONG M A. Algorithm AS 136: a K-means clustering algorithm [J]. Journal of the Royal Statistical Society, 1979, 28(1): 100–108.
- [5] JAIN A K. Data clustering: 50 years beyond K-means [J]. Pattern Recognition Letters, 2010, 31(8): 651–666.
- [6] XIONG H, WU J, CHEN J. K-means clustering versus validation measures: a data-distribution perspective [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2): 318–331.
- [7] HE H, GARCIA E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263–1284.
- [8] KUMAR N S, RAO K N, GOVARDHAN A, et al. Undersampled K-means approach for handling imbalanced distributed data [J]. Progress in Artificial Intelligence, 2014, 3(1): 29–38.
- [9] KUMAR C N S, RAO K N, GOVARDHAN A. An empirical comparative study of novel clustering algorithms for class imbalance learning [C]// Proceedings of the 2nd International Conference on Computer and Communication Technologies, AISC 380. Berlin: Springer, 2016: 181–191.
- [10] 刘云. 不平衡数据的模糊聚类算法研究及在宏基因组重叠群分类中的应用 [D]. 长春: 吉林大学, 2016: 15–48. (LIU Y. Research of fuzzy clustering method on imbalanced dataset and its application in metagenomic contigs binning [D]. Changchun: Jilin University, 2016: 15–48.)
- [11] LIANG J, BAI L, DANG C, et al. The K-means-type algorithms versus imbalanced data distributions [J]. IEEE Transactions on Fuzzy Systems, 2012, 20(4): 728–745.
- [12] 程铃钫, 杨天鹏, 陈黎飞. 不平衡数据的软子空间聚类算法 [J]. 计算机应用, 2017, 37(10): 2952–2957. (CHENG L F, YANG T P, CHEN L F. Soft subspace clustering algorithm for imbalanced data [J]. Journal of Computer Applications, 2017, 37(10): 2952–2957.)
- [13] CHEN L, JIANG Q, WANG S. A probability model for projective clustering on high dimensional data [C]// ICDM 2008: Proceedings of the 8th IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2008: 755–760.
- [14] VIDAL R. Subspace clustering [J]. IEEE Signal Processing Magazine, 2011, 28(2): 52–68.
- [15] XU L, JORDAN M I. On convergence properties of the EM algorithm for Gaussian mixtures [J]. Neural Computation, 1996, 8(1): 129–151.
- [16] 李航. 统计学习方法 [M]. 北京: 清华大学出版社, 2012: 162–165. (LI H. Statistical Learning Method [M]. Beijing: Tsinghua University Press, 2012: 162–165.)
- [17] TASKAR B, SEGAL E, KOLLER D. Probabilistic classification and clustering in relational data [C]// IJCAI 2001: Proceedings of the 17th International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2001, 2: 870–876.
- [18] 朱杰, 陈黎飞. 类属数据的贝叶斯聚类算法 [J]. 计算机应用, 2017, 37(4): 1026–1031. (ZHU J, CHEN L F. Bayesian clustering algorithm for categorical data [J]. Journal of Computer Applications, 2017, 37(4): 1026–1031.)

(下转第 3029 页)



先策略触发概率模型,并以西门子 2070 信号机为载体,利用公交优先硬件在环仿真进行触发概率计算与分析,探索公交优先策略参数及控制逻辑优化方法。红灯早断策略的触发概率远高于绿灯延长策略,但需从交叉口的整体运行效率出发进行控制参数调节;绿灯延长策略控制参数敏感性较低,最佳优化方法是通过优化固定信号配时参数让尽可能多的公交车辆集中在优先相位的时长范围内到达,增加公交优先申请落在优先相位的数量。本文未考虑信号机接受公交申请的概率及交叉口信号配时的相位目标,后续研究可分析这些因素对触发概率的影响。

参考文献(References)

- [1] 王保菊. 城市道路公交优先信号控制方法及仿真研究[D]. 北京: 北京工业大学, 2015: 1–2. (WANG B J. Research on transit signal priority control method and simulation in urban road [D]. Beijing: Beijing University of Technology, 2015: 1–2.)
- [2] 杜鹏程. 交叉口公交信号优先控制策略及优化模型研究[D]. 长春: 吉林大学, 2013: 13–37. (DU P C. Study on intersection transit signal priority strategy and signal timing optimization model [D]. Changchun: Jilin University, 2013: 13–37.)
- [3] 高奎红. 城市道路交叉口公交优先信号协调控制及其仿真研究[D]. 北京: 北京交通大学, 2010: 8–23. (GAO K H. Research on bus priority signal coordination control and simulation urban road intersections [D]. Beijing: Beijing Jiaotong University, 2010: 8–23.)
- [4] BAKER R J, COLLURA J, DALE J J, et al. An overview of transit signal priority[R]. Washington, DC: The Intelligent Transportation Society of America, 2002: 24–31.
- [5] SUNKARI S R, BEASLEY P S, URBANIK T, et al. Model to evaluate the impacts of bus priority on signalized intersections[J]. Transportation Research Record, 1995, 1494: 117–123.
- [6] SKABARDONIS A. Control strategies for transit priority[J]. Transportation Research Record, 2000, 1727: 20–26.
- [7] LUDWICK Jr, J C. Bus priority system: simulation and analysis, UMTA-VA-06-0026-76-1 [R]. Washington, DC: Urban Mass Transportation Administration, 1976: 36–46.
- [8] 徐洪峰, 李克平, 郑明. 基于逻辑规则的单点公交优先控制策略[J]. 中国公路学报, 2008, 21(5): 96–102. (XU H F, LI K P, ZHENG M M. Isolated transit signal priority control strategy based on logic rule [J]. China Journal of Highway and Transport, 2008, 21(5): 96–102.)
- [9] 别一鸣, 宋现敏, 朱慧, 等. 无公交专用道下的单点公交优先控
- 制[J]. 交通信息与安全, 2009, 27(增刊1): 36–40. (BIE Y M, SONG X M, ZHU H, et al. Bus priority signal control for signalized intersection without bus lane [J]. Computer and Communications, 2009, 27(S1): 36–40.)
- [10] 李赫楠. 单个交叉口公交信号优先控制方法研究[D]. 长春: 吉林大学, 2008: 23–27. (LI H N. Research on methods of transit priority signal control of isolated intersection [D]. Changchun: Jilin University, 2008: 23–27.)
- [11] 李凤, 王殿海, 杨希锐. 单点公交被动优先下信号配时方法研究[J]. 交通信息与安全, 2009, 27(3): 48–52. (LI F, WANG D H, YANG X R. Signal timing method for transit passive priority at an isolated intersection [J]. Computer and Communications, 2009, 27(3): 48–52.)
- [12] WADJAS Y, FURTH P G. Transit signal priority along arterials using advanced detection [J]. Transportation Research Record, 2003, 1856: 220–230.
- [13] KENNY L. Transit headway control through conditional signal priority: a micro-simulation based approach using reinforcement learning [D]. Toronto: University of Toronto, 2003: 25–32.
- [14] WU J, HOUNSELL N. Bus priority using pre-signals[J]. Transportation Research, Part A: Policy and Practice, 1998, 32(8): 563–583.
- [15] 邵俊. 公共汽车交通专用道(路)系统设计与评价方法研究[D]. 上海: 同济大学, 2000: 20–50. (SHAO J. Design and evaluation methodologies for exclusive bus lane system [D]. Shanghai: Tongji University, 2000: 20–50.)
- [16] 连培昆, 李振龙, 荣建, 等. 基于 VISSIM 微观交通仿真软件的导流岛机非冲突元胞自动机模型[J]. 计算机应用, 2016, 36(6): 1745–1750. (LIAN P K, LI Z L, RONG J, et al. Cellular automaton model of vehicle-bicycle conflict at channelized islands based on VISSIM microscopic traffic simulation software [J]. Journal of Computer Applications, 2016, 36(6): 1745–1750.)

This work is partially supported by the National Natural Science Foundation of China (51578028).

HUANG Hainan, born in 1983, Ph. D. candidate, lecturer. His research interests include public transport planning and optimization.

LI Xiaofeng, born in 1990, Ph. D. candidate. His research interests include urban traffic planning and optimization.

LIAN Peikun, born in 1985, Ph. D., lecturer. His research interests include urban traffic management and control.

RONG Jian, born in 1972, Ph. D., professor. His research interests include highway capacity, traffic simulation.

(上接第 2849 页)

- [19] LI X, CHEN Z, YANG F. Exploring of clustering algorithm on class-imbalanced data [C]// Proceedings of the 8th International Conference on Computer Science and Education. Piscataway, NJ: IEEE, 2013: 89–93.
- [20] STREHL A, GHOSH J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions [J]. Journal of Machine Learning Research, 2003, 3(3): 583–617.

Foundation of China (61672157), the Innovation Team Project of Fujian Normal University (IRTL1704).

YANG Tianpeng, born in 1991, M. S. candidate. His research interests include data mining.

CHEN Lifei, born in 1972, Ph. D., professor. His research interests include statistical machine learning, data mining, pattern recognition.