基于改进的近邻传播聚类算法的 Gap 统计研究

唐 丹[1],张正军[1],王俐莉[2]
(1. 南京理工大学 理学院 统计与金融数学系,江苏 南京 210094;
2. 海军指挥学院科研部,江苏 南京 210016)

摘 要:由于 K-means 算法初始聚类中心的选取具有随机性,聚类结果可能不稳定,导致 Gap 统计估计的聚类数也可能不稳定。针对这些不足,提出一种改进的近邻传播算法-mAP。该算法考察数据的全局分布特性,不同的点赋予不同的 P 值。在 Gap 统计中用 mAP 算法代替 K-means 算法,提出基于 mAP 的 Gap 统计 mAPGap。mAP 能在较短的时间内得到较好的聚类效果,而且不需要预先设定初始聚类中心,聚类结果更稳定。实验结果表明,mAPGap 在估计聚类数的稳定性和聚类精度上都优于原 Gap。

关键词:聚类分析;近邻传播聚类;偏向参数; K-means 算法;Gap 统计
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2017)01-0182-04
doi:10.3969/j.issn.1673-629X.2017.01.041

Study on Gap Statistic Based on Modified Affinity Propagation Clustering

TANG Dan[1], ZHANG Zheng-jun[1], WANG Li-li[2]
(1. Department of Statistic and Financial Mathematics, College of Science, Nanjing University of Science and Technology, Nanjing 210094, China;
2. Scientific Research Department of Naval Command College, Nanjing 210016, China)

Abstract: Due to the randomness of choosing the initial clustering of K-means method, it may cause the instability of clustering results and then lead to that of clustering numbers which are estimated by Gap statistic. Taking consideration of those disadvantages, an modified AP clustering (mAP) is presented which utilizes the global distribution to give different P to different points. mAP method is put forward to substitute the K-means in Gap statistic named mAPGap. mAP method has more stable clustering center because the initial clustering center and numbers are not needed in advance and it can get better clustering in short time. The experimental results demonstrate mAPGap is superior to Gap in clustering stability and accuracy.

Key words: cluster analysis; affinity propagation clustering; preference; K-means algorithm; Gap statistic

0 引 言

数据集的聚类数估计是数据分析中的一项重要课题。2000 年,R. Tibshirani 等提出确定最佳聚类数的 Gap 统计量[1],采用的聚类算法是 K - means 算法,该算法需要选取初始聚类中心,通常随机选取 K 个样本点作为初始聚类中心。2013 年,刘倩基于 Gap 统计方法研究了 K - means 算法,提出了一种基于数据分布规律具有自适应特点的 DSA-K - means 算法[2]。2013 年,陆琴琴针对基于矩 Gap 统计的图像分割算法中 K - means 算法存在的缺陷,提出了 MMGSK 算法[3]。

2007 年,Frey[4] 和 Mezard M[5] 提出了属于划分聚类方法的近邻传播(Affinity Propagation, AP)算法。该算法具有如下优点:能在较短时间内得到较好的聚类效果[6];算法中类代表点是原始数据中的点,而不是数据的均值点;以误差平方和作为衡量算法优劣程度的准则函数时,算法聚类的误差平方和显著低于其他方法聚类的误差平方和。但是 AP 算法对偏向参数 P 的设定,没有考虑数据的分布结构,认为所有数据点成为类代表点的可能性相同。

文中提出了利用全局数据信息设置偏向参数 P 的改进的 AP 算法-mAP(modified Affinity Propagation),

收稿日期:2016-03-15 **修回日期**:2016-06-22 **网络出版时间**:2017-01-04
基金项目:全国统计科学研究计划重点项目(2013LZ45)
作者简介:唐 丹(1990-),女,硕士研究生,研究方向为数据分析与数据挖掘;张正军,博士,副教授,研究方向为数据分析与数据挖掘、马尔可夫链理论与方法等。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20170104.1102.086.html
万方数据

同时提出了自适应的增加步长,得到数据集的聚类数,运用 Gap 统计量估计出该数据集的合理聚类数。

1　mAP 模型

AP 是一种基于近邻信息传播的聚类算法,根据 N 个数据点之间的相似度进行聚类[7]。这些相似度组成 $N \times N$ 的相似度矩阵 $\boldsymbol{S}$。$\boldsymbol{S}$ 矩阵的对角线上的数值 $s(k,k)$ 称为偏向参数 preference,记作 P,表示数据点 k 作为类代表点的适合程度。该值越大,k 点成为类代表点的可能性也越大,同时得到的聚类数越多[8]。AP 算法中传递两种类型的消息,即归属度 $a(i,j)$ 和吸引度 $r(i,j)$,前者表示 x_i 选择 x_j 作为类代表的合适程度,后者表示 x_i 对 x_j 作为类代表点的支持程度[9]。AP 算法通过迭代过程不断更新每一个点的吸引度和归属度值,直到迭代过程收敛,类代表也随之固定,同时将其余的数据点分配到相应的聚类中[10]。

传统的 AP 算法中相似度矩阵对角线上的偏向参数 P 是相同的,一般取所有相似度的中值(median(s(:,3))),意味着数据点在开始时被选择作为类代表点的概率是相同的。但是,P 值的这种初始设置方法是不科学、不精确的,因为它忽略了数据点结构的差异对某点成为类代表点的可能性的影响。类代表点的选择与点的位置密切相关:对给定的数据集 X 和点 x_i,x_j,如果 x_i 是边缘数据点而 x_j 是中心数据点,那么从其他点到 x_i 的距离和大于到 x_j 的距离和,x_j 成为类代表的可能性要大于 x_i。

针对如上假设,文中提出基于全局数据点设置 P 的值,同时为了获得不同的聚类数,提出一种自适应设置步长获得不同聚类数的方法。具体步骤如下:

(1)对任意点 x_i,计算从其他所有点到 x_i 的相似度之和:

$$\mathrm{Ss}(i) = \sum_{j=1,j\neq i}^{n} s(x_i,x_k) \tag{1}$$

(2)标准化 Ss(i):

$$\mathrm{NormSs}(i) = \mathrm{Ss}(i) / \sum_{i=1}^{n} \mathrm{Ss}(i) \tag{2}$$

(3)对 AP 算法设置步长,获得不同聚类数。

根据上述讨论可知,当每个数据点的 P 值相同时,聚类数随 P 值的增大而增大。所以为了得到不同的聚类数,P 存在取值区间 $[P_{\min},P_{\max}]$。Wang C D[11] 通过研究得到:

$$\begin{aligned} P_{\min} &= \min_{i\neq j} s(x_i,x_j) \\ P_{\max} &= \max_{i\neq j} s(x_i,x_j) \end{aligned} \tag{3}$$

由于 AP 算法每个点的 P 值相同,可以通过下式获得:

$$P = \{P_i \mid P_i = \frac{P_{\max} - P_{\min}}{N_{\mathrm{pref}} - 1} * (i-1),\ i = 1,2,\cdots,N_{\mathrm{pref}}\} \tag{4}$$

其中,N_{pref} 是输入参数,表示设置 N_{pref} 个不同 P。

(4)对 mAP 算法设置步长,获得不同聚类数。

对 $i = 1,2,\cdots,n$(n 代表样本点个数),有:

$$\begin{aligned} P_{i\mathrm{Min}} &= P_{\min} * \mathrm{NormSs}(i) \\ P_{i\mathrm{Max}} &= P_{\max} * \mathrm{NormSs}(i) \end{aligned} \tag{5}$$

$$P(k,i) = P_{i\mathrm{Min}} + \frac{P_{i\mathrm{Max}} - P_{i\mathrm{Min}}}{N_{\mathrm{pref}} - 1} * (k-1),\ k = 1,2,\cdots,N_{\mathrm{pref}} \tag{6}$$

其中

$$\begin{aligned} P_{\min} &= \min_{i\neq j} s(x_i,x_j) * n \\ P_{\max} &= \max_{i\neq j} s(x_i,x_j) \end{aligned} \tag{7}$$

由于对每个点的 $P_{i\mathrm{Min}}$ 都是 $P_{\min}$ 乘以一个权重,所以会使得每个点的初始 P 值变小。为了确保能够取到 1 ~ max$K$ 类,这里把 $P_{\min}$ 取得更小,是用相似度矩阵元素的最小值乘以样本点个数 n。

对于连续的 k,mAP 算法得到的分类情况会大量重复。文中根据 mAP 算法的聚类结果动态计算步长,在不改变结果的基础上优化了算法的运行时间。

2　mAP 算法

(1)初始化。计算相似度矩阵 $[\boldsymbol{s}(i,k)]_{(n^2-1)\times 3}$,确定阻尼系数 $\lambda(0 < \lambda < 1)$、最大迭代次数 maxits、聚类划分连续不变次数 Convits。通常,$\lambda = 0.9$,maxits = 1 000,Convits = 100。

$$\boldsymbol{s}(i,k) = -\parallel x_i - x_k \parallel, i \neq k \tag{8}$$

(2)对于每一个 N_{pref},根据式(5) ~ (7)计算每个点的偏向参数,能得到 N_{pref} 个 $1 * n$ 数组。

(3)消息传递。

①更新吸引度矩阵 $\boldsymbol{r}(i,k)$ 和归属度矩阵 $\boldsymbol{a}(i,k)$。

$$\boldsymbol{r}(i,k) = s(i,k) - \max\{a(i,k') + s(i,k')\}\ (k' \neq k) \tag{9}$$

$$\boldsymbol{a}(i,k) = \min\{0, r(k,k) + \sum_{i's.t.\ i'\notin\{i,k\}} \max\{0, r(i',k)\}\} \tag{10}$$

$$\boldsymbol{a}(k,k) = \sum_{i's.t.\ i'\neq k} \max\{0, r(i',k)\} \tag{11}$$

②引入阻尼系数 λ,消除可能出现的震荡。

$$\begin{cases} \boldsymbol{r}_{\mathrm{new}}(i,k) = \lambda \times \boldsymbol{r}_{\mathrm{old}}(i,k) + (1-\lambda) \times \boldsymbol{r}(i,k) \\ \boldsymbol{a}_{\mathrm{new}}(i,k) = \lambda \times \boldsymbol{a}_{\mathrm{old}}(i,k) + (1-\lambda) \times \boldsymbol{a}(i,k) \end{cases} \tag{12}$$

其中,$\lambda(0 < \lambda < 1)$ 越大越能消除震荡,但会减缓算法收敛的速度。

万方数据

(4)循环执行步骤(3),直到满足终止条件。终止条件为达到最大迭代次数 maxits 或聚类划分连续 Convits 次不变。

(5)输出聚类结果。输出 N_{pref} 组类代表点及其对应的数据点划分。

3 mAPGap 模型(Gap Statistic Based on mAP)

原 Gap 统计方法的主要思想是:选择一个参考分布,将待分类数据的离散程度与由参考分布生成数据集的离散程度进行比较,以分类数为自变量,建立一个比较统计量,通过分析该统计量关于类数的变化情况确定最佳聚类数[12]。

原 Gap 统计的主要内容是:假设数据是通过 K-means 算法已被聚为 k 类:$C_1, C_2, \cdots, C_k, n_r = |C_r|$。令:

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \tag{13}$$

其中,D_r 表示聚类 r 中所有数据两两之间的距离平方之和。

再令:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r = \sum_{r=1}^{k} \sum_{i \in C_r} \| x_i - \bar{x} \|^2 \tag{14}$$

其中,W_k 表示 k 类离差程度的总和。

由此定义 Gap 统计量:

$$\mathrm{Gap}_n(k) = E_n^*(\log(w_k)) - \log(w_k) \tag{15}$$

其中,E_n^* 表示在某参考分布(文中选用均匀分布)下的期望[13]。

为了区分基于不同算法的 Gap 统计,这里把基于 K-means 算法的原 Gap 统计记作 KMGap。

由于 KMGap 统计需要预先生成大量的参考数据集,因而不适宜通过 mAP 算法实现不同类数的聚类,而且参考数据集类内离差程度是均匀分布下大量数据集类内离差程度的期望,根据 Khinchin 大数定律[14]知,使用 K-means 算法聚类能够保证结果的稳定性。鉴于以上两点,对参考数据集的聚类方法仍使用 K-means 算法。

使用 mAP 算法对样本数据集进行聚类,当选择负的欧氏距离作为两个样本点的相似度时,mAP 算法的准则函数也即是使每个样本点到其类代表点的平方距离之和最小,这与 K-means 算法的准则函数一致。因此使用 mAP 算法对样本数据集进行聚类,使用 K-means 算法对参考数据集进行聚类,再利用 Gap 统计量估计该数据集的聚类数是合理的。

mAPGap 的计算可分为以下 3 步:

(1)利用 mAP 算法,将样本数据集聚集为 k($k = 1,2,\cdots,K$)类,计算 W_k(在偏向参数不同时会出现相同的分类数 k,这里 W_k 取它们的平均值)。

(2)产生 B 个参考数据集,通常 $B = 1\,000$。利用 K-means 算法分别将这 B 个数据集聚集为 k($k = 1,2,\cdots,K$)类,计算 W_{kb}^*($b = 1,2,\cdots,B$, $k = 1,2,\cdots,K$)以及 Gap 值:

$$\mathrm{Gap}(k) = \frac{1}{B} \sum_b \log(W_{kb}^*) - \log(W_k) \tag{16}$$

(3)令 $\bar{l} = \frac{1}{B} \sum_b \log(W_{kb}^*)$, $sd_k = \left[\frac{1}{B} \sum_b (\log(W_{kb}^*) - \bar{l})^2 \right]^{\frac{1}{2}}$,则 $s_k = sd_k \sqrt{1 + \frac{1}{B}}$,使式(17)成立的最小的 k 就是寻求的最佳聚类数。

$$\mathrm{Gap}(k) \geqslant \mathrm{Gap}(k+1) - s_{k+1} \tag{17}$$

4 仿真实验与分析

文中实验采用 MATLAB7.0 开发环境,在 Windows7 操作系统的计算机上运行通过。

4.1 实验数据集

选用 UCI 标准数据集中的 Haberman、Seeds、Breast Cancer 以及人工数据集 Data1 作为测试数据集。其中,Data1 是由中心分别为(1,1),(3,3),协方差矩阵为 [1 0; 0 2] 的二维高斯分布数据组成,每类有 150 个样本点,共 300 个样本点。表 1 给出了数据集的相关信息。

表 1 实验数据集描述

数据集	样本数	维数	类簇数	各类大小
Data1	300	2	2	150,150
Haberman	306	3	2	80,226
Seeds	210	7	3	70,70,70

4.2 实验评价标准

实验采用 3 种评价指标对聚类结果进行评价:

(1)对三个数据集,分别利用 KMGap、APGap、mAPGap 重复运行 20 次,估计出最佳聚类数的正确率 p。

(2)算法运行时间 t。虽然 AP 算法复杂度为 $O(N*N*\log N)$,而 K-means 的是 $O(N*K)$,但当样本容量不是非常大时,两者时间相近。故这里对具体数据集的运行时间进行比较。

(3)聚类精度。定义为:

$$\mathrm{accuracy} = \mathrm{TC}/(\mathrm{TC} + \mathrm{FC}) \tag{18}$$

其中,TC 为正确聚类的数据数;FC 为错误聚类的数据数。

由于 AP 算法和 mAP 算法都会出现在估计的最佳聚类数相同时,具体的聚类结构不同的情况,所以对某个特定的聚类,聚类精度有不同的值。结果分析中,AP 算法和 mAP 算法的精度记录的都是最小值。当最

万方数据

小值都优于 KMGap 时,说明 APGap 算法和 mAPGap 算法聚类的精度优于 KMGap。

4.3　结果与分析

对三个数据集分别执行 KMGap、APGap、mAPGap 聚类,聚类结果及性能如图 1、表 2 所示(由于版面有限,这里只给出 Haberman 的聚类结果)。

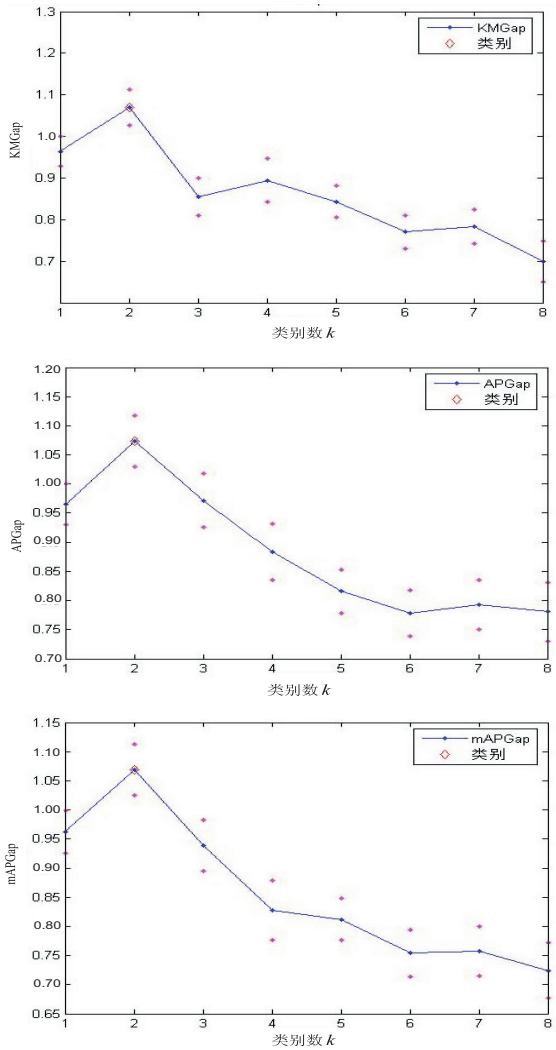

图 1　Haberman 数据集中的 KMGap、APGap、mAPGap 随类数 k 的变化曲线

表 2　KMGap、APGap、mAPGap 算法在四个数据集上的比较

数据集	算法	类数	p/%	t/s	a
Data1	KMGap	2	95	412.892	0.857 1
	APGap	2	100	613.426	0.890 5
	mAPGap	2	100	533.197	0.904 8
Haber-man	KMGap	2	70	430.372	0.490 2
	APGap	2	100	771.051	0.526 1
	mAPGap	2	100	661.109	0.532 7
Seeds	KMGap	3	90	225.975	0.83
	APGap	3	100	330.17	0.866 7
	mAPGap	3	100	303.943	0.85

实验结果表明,相比于 KMGap,APGap、mAPGap 虽然运行时间增加了一点,但二者具有很好的稳定性,同时二者均可以提高分类精度。而且 mAPGap 与 APGap 相比,在不影响稳定性和精度的情况下,减少了算法运行时间。总体来说,mAPGap 算法优势明显的原因是:利用数据结构设置每个数据点的偏向参数,减少了算法运行时间;不需要预先设定初始聚类中心和聚类数,使得聚类结果更稳定,精度更高。

5　结束语

文中提出根据数据的全局分布特性设置偏向参数 P 的 AP 算法(mAP),在 Gap 统计中,用 mAP 算法替换 K-means 算法,提出基于 mAP 的 Gap 统计量(mAPGap)。通过实验证明了 mAPGap 在聚类结果的稳定性,并且在精度上要优于原算法。

参考文献:

[1] Tibshirani R, Walther G, Hastie T. Estimatig the number of cluster in a dataset via the gap statistic[J]. Journal of the Royal Statistical Society, 2001, 63(2): 411-423.

[2] 刘　倩. 基于 GS 方法的图像分割估计数的多信息动态研究[D]. 南京:南京理工大学, 2013.

[3] 陆琴琴. 基于矩 Gap 统计的图像分割方法[D]. 南京:南京理工大学, 2014.

[4] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[5] Mezard M. Where are the examplars? [J]. Science, 2007, 315(5814): 949-951.

[6] 冯晓磊, 于洪涛. 密度不敏感的近邻传播聚类算法研究[J]. 计算机工程, 2012, 38(2): 159-162.

[7] 邢　艳, 周　勇. 基于互近邻一致性的近邻传播算法[J]. 计算机应用研究, 2012, 29(7): 2524-2526.

[8] 段丽莉. 改进的近邻传播算法及其在图像处理中的应用[D]. 西安:西安电子科技大学, 2014.

[9] 邢长征, 刘　剑. 基于近邻传播与密度相融合的进化数据流聚类算法[J]. 计算机应用, 2015, 35(7): 1927-1932.

[10] 肖　宇, 于　剑. 基于近邻传播算法的半监督聚类[J]. 软件学报, 2008, 19(11): 2803-2813.

[11] Wang C D, Lai J H, Suen C Y, et al. Multi-exemplar affinity propagation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(9): 2223-2237.

[12] 童　波. 基于 MFGS 方法图像最佳分隔数的研究[D]. 南京:南京理工大学, 2011.

[13] 黄陈蓉, 张正军, 吴慧中. 图像边缘检测的多尺度灰度 Gap 统计模型[J]. 中国图象图形学报, 2005, 10(8): 1018-1023.

[14] 冯　予, 陈　萍. 概率论与数理统计[M]. 第 2 版. 北京:国防工业出版社, 2015: 114-117.

万方数据