文章编号:1001 - 9081(2010)03 - 0810 - 03

全最小一乘准则下的 LGA 新算法

曹慧荣,方　杰
(廊坊师范学院 数学与信息科学学院,河北 廊坊 065000)
(huirongcao@126.com)

摘　要:为克服传统提取数据集中线性结构的 LGA 对噪声数据比较敏感的缺陷,提出了两种基于稳健的全最小一乘准则下的 LGA 新算法。首先证明了全最小一乘准则下数据集最优划分的存在性,并据此给出一种有限步终止算法。其次为提高计算速度,根据 k-means 算法、全最小一乘准则和重抽样方法给出另一种快速收敛算法。通过与传统的 LGA 和基于 Trimmed k-means 思想的稳健 LGA 的比较,仿真结果表明提出的算法具有较好的稳健性,可以在离群数据较多的情形下,同时找出数据集合中的所有强线性结构。

关键词:LGA;全最小一乘;稳健性;聚类分析
中图分类号: TP301.6　　**文献标志码**:A

New linear grouping algorithm based on total least absolute deviation criteria

CAO Hui-rong, FANG Jie
(*College of Mathematics and Information Science, Langfang Normal College, Langfang Hebei* 065000, *China*)

Abstract: To overcome the drawback that the traditional Linear Grouping Algorithm (LGA), when extracting linear structures, is sensitive to outliers in datasets, a new finite step according to existence of optimal linear grouping in data set and a new algorithm based on k-means clustering, total least absolute deviation and resampling were proposed, which detected several different linear relations at once to minimize the total orthogonal distances from n given points to its nearest hyperplanes. Finally, by comparison with linear grouping algorithm and robust linear grouping algorithm based on impartial trimmed k-means, the proposed algorithms are more robust and can detect all strong linear structures in datasets including a lot of outliers.

Key words: Linear Grouping Algorithm (LGA); total least absolute deviation; robustness; cluster analysis

0　引言

聚类分析是一种无监督的模式识别问题,已经被广泛地应用到许多领域中,如数据挖掘、计算机视觉、机器学习和空间数据库技术等领域。常见的聚类算法有 k-means 聚类和谱系聚类算法等[1-2],但这些算法有两点不足。一是算法有效性对数据的空间分布有较强的依赖性。k-means 聚类对于特征空间呈超球体的情况聚类效果较好,而对于呈任意形状簇分布的情况则聚类效果较差。二是算法是以点作为每一类的中心,而当类中心为超平面时,算法不适用[3]1289。

LGA(Linear Grouping Algorithm)[3]1287是基于全最小二乘准则的线性聚类方法,可以有效地在数据集中提取不同的线性结构。LGA 在动物异速生长问题[3]1297-1300、地震断层面的确定[4]等方面取得了初步应用。实验发现,该算法可以对具有线性结构的数据集进行有效聚类,但同时对数据噪声点和异常点较敏感,算法的鲁棒性较差。为此,Garcia-Escudero 等人[5]301针对数据集存在噪声和异常点情形,使用 Trimmed k-means 方法增强了聚类的鲁棒性,但这样做有两方面的不足:一是剔除异常点后获得的模型会受到一定程度的影响;二是异常点恰好在某些方面确实反映了某些特殊的信息,不应该随意剔除。因此,利用 LGA 对有噪声数据或异常点的数据进行建模和参数估计时,应选择稳健的准则以减少噪声数据或异常点的影响,得到符合实际的模型。

本文给出一种全最小一乘准则下的 LGA 新算法,将 LGA 中的全最小二乘准则改为全最小一乘准则,利用全最小一乘准则准则下拟合超平面的性质,提出了一种有限步终止的全局最优算法和类似于 k-means 聚类算法的局部最优算法。通过计算机仿真发现,新算法能够有效地克服噪声数据和异常数据的影响,较文献[3]中的 LGA 和文献[5]中的 RLGA 方法更稳健,进一步验证了新算法的有效性。

1　LGA

LGA 对数据集进行聚类分析时,首先随机地对数据集进行初始分类,每一类利用正交最小二乘回归给出拟合的超平面,然后重新将每个数据分配给离它最近的超平面,从而得到新的分类,再由主成分分析方法给出各个新类的超平面,不断重复上述过程,直到超平面的变化很小或不变为止。因此,LGA 可以用来挖掘数据集中不同线性结构模式。由 LGA 给出的低维超平面描述了不同类的数据特征,并且揭示数据集的不同子类中的线性关系,从而克服了传统聚类分析方法的不足。LGA 使用迭代优化算法收敛于局部最优值,因此使用重抽样技术选择大量的迭代初始值从而增加算法的性能,对未知数据有效地聚类。

LGA 用超平面替代传统 k-means 聚类算法中的中心点作

收稿日期:2009-09-23;**修回日期**:2009-11-27。　**基金项目**:河北省教育厅自然科学研究项目(Z2009138)。
作者简介:曹慧荣(1975-),女,山东嘉祥人,讲师,硕士,主要研究方向:数据挖掘、模式识别;　方杰(1981-),男,河北廊坊人,助教,主要研究方向:计算机网络。

为聚类原型。

设 $Z=\{z_i\,|z_i=(z_{i1},z_{i2},\cdots,z_{ip})^{\mathrm{T}},i=1,2,\cdots,n\}\subset\mathbf{R}^p$ 为 n 个数据点构成的数据集,$H=\{h_j\,|\boldsymbol{\alpha}_j^{\mathrm{T}}z+\beta_j=0,\|\alpha_j\|=1,j=1,2,\cdots,k\}$ 为 $\mathbf{R}^p$ 中的一组超平面,LGA的目标是找到一组平面,使得每个数据点到距离其最近的超平面的距离平方之和最小,使目标函数:

$$S_{\mathrm{TLS}}=\sum_{i=1}^{n}\min_{1\leqslant j\leqslant k}d^2(z_i,h_j)\tag{1}$$

取得最小值,其中 $d(z_i,h_j)=\|\boldsymbol{\alpha}_j^{\mathrm{T}}z_i+\beta_j\|$ 为数据点 z_i 到超平面 h_j 的欧式距离。

由于 k-means 聚类算法是 NP-hard 问题[6-7],寻找全局最优解比较困难,因而 LGA 也使用类似于 k-means 聚类的迭代优化收敛于一个局部最小值。对于每一个子类,LGA 使用正交回归拟合超平面。为加快算法的收敛速度,使用重抽样技术选择大量的迭代初始值。LGA 按照求式(1) 最小给出每个数据点所属的子类和每一子类的中心 H。

2　全最小一乘准则下的 LGA 新算法

传统的 LGA 是基于全最小二乘准则的,而最小二乘稳健性不如最小一乘,因而将目标函数改为:

$$S_{\mathrm{TLAD}}=\sum_{i=1}^{n}\min_{1\leqslant j\leqslant k}d(z_i,h_j)\tag{2}$$

即可得到稳健的全最小一乘准则下的 LGA 新算法。

由于全最小一乘准则下目标函数不可微,因而本质上是一个不可微优化问题,求解比较困难。幸运的是,利用全最小一乘准则拟合超平面有一些好的性质可以帮助计算,下面给出基于这些性质的有限步终止的全局最优算法和类似于 k-means 聚类算法的局部最优算法。

2.1　有限步终止算法

如果超平面个数 $k=1$,则由文献[8] 中定理 3.2 可得如下引理。

引理 1　由 n 个互异数据点按全加权最小一乘准则确定的最优拟合超平面一定通过该数据集中的 p 个互异点。

当超平面个数 $s\geqslant 2$ 时,可得如下结论。

定理 1　n 个互异点可以确定 C_n^p 个超平面,k 个最优超平面必位于这些超平面中。

证明　因为任意 p 个互异点可以确定唯一超平面,所以 n 个互异点可以确定 C_n^p 个超平面。

寻找 k 个最优超平面本质上是按加权全最小一乘准则将 n 个数据点划分为 k 个互斥(互不相容) 子集,因为所求 k 个最优超平面为每一子集的最优中心,由引理 1 可知这些中心必通过每一子集中的 p 个互异点,所以 k 个最优超平面必位于这些直线中。　证毕。

因为 k 个最优超平面必位于这些超平面中,只需要在 $T_n=C_n^p$ 个超平面中穷举 $C_{T_n}^k$ 种情形即可得到最优的 k 个超平面,其计算量为 $O(n^{pk})$。

有限步终止算法的步骤如下。

步骤 1　由 n 个互异点确定 $T=C_n^p$ 个超平面。设点 $z_{i_1},z_{i_2},\cdots,z_{i_p}$ 为数据集中 p 个互异点,且都在超平面 $h_t:\boldsymbol{\alpha}_t^{\mathrm{T}}z+\beta_t=0$ 上,则将点代入超平面方程中,得到一个 p 元恰定线性方程组,又点是互异的,所以可以确定一个超平面 h_t。

步骤 2　从步骤 1 中得到的超平面中任选 k 个超平面,计算相应的 $S_t=\sum_{i=1}^{n}\min_{1\leqslant j\leqslant k}d(z_i,h_{tj})$,其中 $d(z_i,h_{tj})$ 为第 i 个点到第 t 组超平面中第 j 个超平面的距离,$t=1,2,\cdots,C_T^k$。

步骤 3　求出 $\min_{1\leqslant t\leqslant C_T^k}S_t$,与其对应的 k 个超平面即为 k 个最优超平面。

使用该算法一定可以得到 k 个最优超平面的所有精确解,但是随着数据点数量 n 的增大,计算量会迅速增大,因此需要研究数据量 n 较大时的有效算法。

2.2　快速收敛算法

下面给出一种类似于 k-means 算法的收敛于局部最小的线性聚类算法。

快速收敛算法的步骤如下。

步骤 1　随机产生初始 k 个超平面。随机选取互不相同的 p 个点,确定一个超平面,重复该过程,直到产生 k 个互不相同的超平面为止,作为初始超平面集合。

步骤 2　初始化子类。计算每个点到 k 个超平面的欧式距离,按照每一个点分配给距离最小的超平面的原则进行初始分类。再由引理 1 对每一子类拟合超平面,共重新确定 k 个超平面。

步骤 3　迭代改进。对步骤 2 中得到的 k 个超平面,计算每个点到 k 个超平面的欧式距离。按照步骤2 中原则进行分类并重新确定 k 个超平面。如果能使得模型(2) 中 S_{TLAD} 减小,则更新 k 个超平面,否则按步骤 1 随机产生 k 个超平面,转步骤 2。这种重抽样方法可以减少迭代次数,加快算法的收敛速度。

步骤 4　中止准则。重复步骤 2 ~ 3 一定次数或迭代一定次数后 S_{TLAD} 没有减小则停止迭代,选择使得模型(2) 中 S_{TLAD} 最小的 k 个超平面作为最优超平面。

快速收敛算法中,有几点需要说明。

1) 超平面数目 k 的确定。超平面数目 k 可以通过增加一个超平面后得到的 S_{TLAD} 能否较增加前明显减小来自动给定,也可以根据 GAP 统计量[3]1292-1294 确定。

2) 确定每一子类的超平面中心的方法。当空间维数 p 较小时,如 $p=2,3$,使用文献[8] 中的方法可以较快确定超平面;而当 n 和 p 都较大时,按文献[8] 中的方法确定超平面计算量很大,可以使用随机组合优化算法[9] 较快地确定。

3) 重抽样方法。本算法步骤3 中当 S_{TLAD} 没有减小时再随机产生初始 k 个超平面进行迭代是一种很好的策略,可以有效地避免死循环,加快收敛速度。

3　计算机仿真

为便于观察,考查 2 维空间情形下的聚类效果,并与文献[3]中的 LGA 和文献[5]中基于 Trimmed k-means 思想的稳健 RLGA 方法比较。根据快速收敛算法,我们使用 Matlab 7.6编写了相关程序,模拟了线性结构数据和噪声数据。LGA 和 RLGA 采用文献[10]的作者提供的 R 包在统计软件 R2.8.1下运行,并将结果导入 Matlab 中后绘出图 3、4。

观测数据由两部分组成:一部分是线性结构数据,共 300 个点,分布在 3 条线段附近,噪声标准差 $\sigma=0.01$,在图 1 中分别以“○”、“□”和“×”标出;另一部分是离群数据,为在

(0,1) × (0,1)区域中均匀分布的 300 个点,以“·”标出,离群数据占所有数据的 50%。

使用快速收敛算法得到的结果见图 2,可以看出,该算法在离群数据较多(占 50%)的情形下,能够有效地发现数据集中的强线性结构。

图 3 为按模型式(1),即全最小二乘准则[3]1290-1291得到的聚类结果,可以看出传统的 LGA 方法无法挖掘数据集中的强线性结构。图 4 为按修匀(Trimmed)全最小二乘准则[5]304得到的聚类结果,可以看出基于 Trimmed k-means 思想的稳健 RLGA 方法也不能正确挖掘数据集中的所有强线性结构。

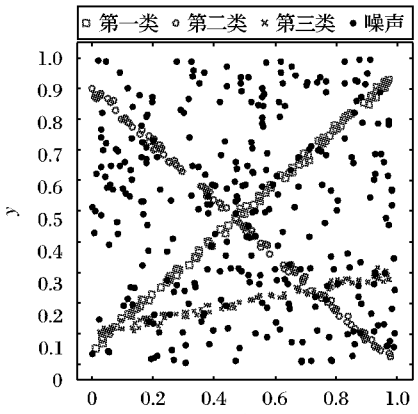

图 1 点的分布

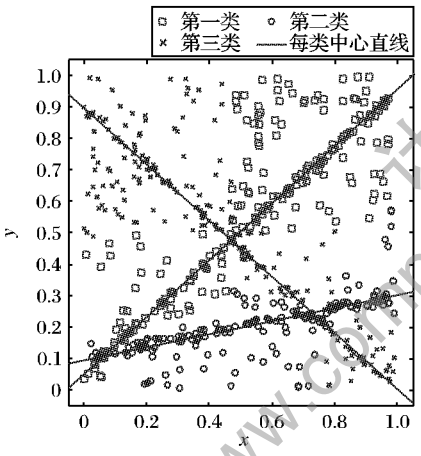

图 2 快速收敛算法得到的结果

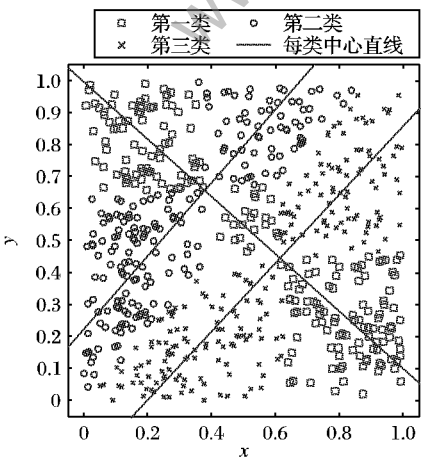

图 3 LGA 得到的结果

4 结语

本文讨论了全最小一乘准则下 LGA 的鲁棒估计问题,提出了一种新的鲁棒算法。实验表明,与全最小二乘准则下的 LGA 和 RLGA 相比,该算法具有很强的鲁棒性,能够处理含有高比例离群数据的多结构混杂观测样本集。

此外,算法在提取线性结构的同时,可以给出每一结构的超平面中心参数。全最小一乘准则还可以换成其他的稳健估计准则,如 T-型估计。

快速收敛算法仍是基于 k-means 聚类算法思想,比以点为核心情形收敛速度要慢且可能陷入局部极小值,因而提高收敛的速度和准确性也是要研究的问题。

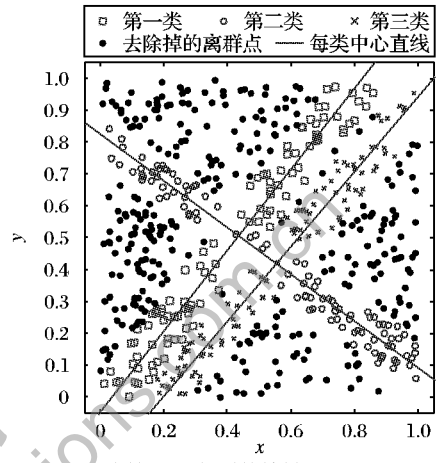

图 4 稳健 LGA 得到的结果(α=0.50)

参考文献:

[1] JAIN A K, MURTY M N, FLYNN P J. Data clustering: A review [J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[2] XU RUI, DONALD WUNSCH I I. Survey of clustering algorithms [J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678.

[3] van AELST S, WANG X, ZAMAR R H, *et al*. Linear grouping using orthogonal regression [J]. Computational Statistics and Data Analysis, 2006, 50(5): 1287-1312.

[4] OUILLON G, DUCORBIER C, SORNETTE D. Automatic reconstruction of fault networks from seismicity catalogs: Three-dimensional optimal anisotropic dynamic clustering [J]. Journal of Geophysical Research, 2008, 113(B1): B01306.1-B01306.15.

[5] GARCIA-ESCUDERO L A, GORDALIZA A, SAN MARTIN R, *et al*. Robust linear clustering [J]. Journal of the Royal Statistical Society, 2009, 71(1): 301-318.

[6] MAHAJAN M, NIMBHORKAR P, VARADARAJAN R K. The planar k-means problem is NP-hard [C]// Proceedings of the 3rd Annual Workshop on Algorithms and Computation WALCOM, LNCS 5431. Berlin: Springer-Verlag, 2009: 274-285.

[7] ALOISE D, DESHPANDE A, HANSEN P, *et al*. NP-hardness of Euclidean sum-of-squares clustering [J]. Machine Learning, 2009, 75(2): 245-248.

[8] 梁怡,吴可法.一类非光滑最优化问题的有限步解法[C]//中国运筹学会第六届学术交流会论文集.香港: Global-Link 出版公司,2000:801-811.

[9] 王福昌,曹慧荣,安霞.基于门限接受算法的正交最小一乘回归新算法[J].数学的实践与认识,2009,39(20):122-128.

[10] HARRINGTON J. Tools for linear grouping analysis (LGA) [EB/OL]. (2009-04-17) [2009-09-19]. http://cran.r-project.org/web/packages/lga/lga.pdf.