一种改进的 K-means 聚类方法在惯导系统中的应用*

党宏涛[1],杜祖良[1,2],于湘涛[2],王常虹[1],曲雪云[2]

(1 哈尔滨工业大学空间控制与惯性技术研究中心,哈尔滨 150001;2 北京自动化控制设备研究所,北京 100074)

摘 要:为了提高高维数据聚类精度,提出了一种基于数据分布规律的 K-means 聚类方法。通过 K-means 聚类粗略寻找高维数据分布规律,构造不同的自适应因子对聚类数据进行综合 K-means 聚类精度校正。将所提出方法应用于平台惯导系统标定数据聚类中,计算结果表明该方法可以很好的对加速度计标定数据进行聚类和评价,具有较好的实际应用价值。

关键词:高维数据;分布规律;K-means 聚类;惯导系统

中图分类号:TJ760.2 **文献标志码**:A

The Application of An Improved K-means Clustering Algorithm in Inertial Navigation System

DANG Hongtao[1], DU Zuliang[1,2], YU Xiangtao[2], WANG Changhong[1], QU Xueyun[2]

(1 Space Control and Inertial Technology Research Center, Harbin Institute of Technology, Harbin 150001, China;
2 Beijing Institute of Automatic Control Equipment, Beijing 100074, China)

Abstract: To improve the precision of high-dimensional data cluster, an improved K-means clustering algorithm based on data distribution was proposed. The distribution of high-dimensional data was found by the K-means clustering method. The clustering accuracy of the data was corrected by the clustering factor. The proposed method is applied in the platform inertial navigation system (INS); the results show that the method is good for calibration data clustering and evaluation.

Keywords: high-dimensional data; distribution; K-means clustering; INS

0 引言

聚类分析是数据挖掘的一个重要组成部分,近年来随着大量科研和商业数据的涌现,K-means 聚类算法得到了广泛的应用[1-2]。但由于高维空间数据分布稀疏性,基于经典的 K-means 算法很难达到满意的聚类效果,因此许多研究者分别从近似度函数[3]、降维技术[4]、遗传算法[5]、模糊加权[6]等对 K-means 算法进行了研究,在聚类准则、K 值确定、计算效率方面取得了较好的效果,但在聚类精度上仍需进一步研究。

基于不同类不同维数据具有不同的分布特征,文中提出一种新的自适应 K-means 高维聚类算法。UCI 数据是验证聚类算法的典型数据库,采用 UCI 标准数据进行仿真验证,结果表明该算法具有很高的聚类精度。将所提出方法应用于加速度标定数据聚类中,计算结果表明该方法具有很好的聚类效果,而且对于发现加速度标定数据分布规律及其对其重复性进行评价具有很好的应用价值。

1 基于数据分布规律的改进 K-means 聚类算法

1.1 算法思想

在一定条件下(比如各类中心距离较大),实际工程中大多数测试数据具有类似正态分布的特性,对于高维数据,由于每一类样本的每一维数据分布规律不同,因此,样本总体在每一维上会呈现出不同程度的可分性。利用数据的这一特征可进行自适应 K-means 聚类校正。首先,对聚类样本各维数据进行方差计算,对方差最大的维数进行单维数据 K-means 聚类,并将单维数据聚类转换到总体样本 K-means 聚类;其次,计算经 K-means 聚类后的方差,构造自适应 K-means 聚类算法继续进行聚类,为防止过聚类;最后,用每类每维数据方差构造自适应因子,通过

* **收稿日期**:2011-05-16
基金项目:"十一五"惯性技术预研项目(51309030103)资助
作者简介:党宏涛(1976-)男,甘肃正宁人,博士研究生,研究方向:惯性技术及数据挖掘。

自适应 K-means 聚类计算,有效抑制了过聚类现象,提高了聚类精度。

1.2　算法步骤

对于 m 个 n 维待聚类数据样本 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$,设定聚类类数为 K。

1)计算样本 $\boldsymbol{X}$ 的 n 维数据的方差,对方差最大的第 $c(1\leqslant c\leqslant n)$维数据进行单维 K-means 聚类;

2)将单维聚类结果对应到 n 维数据样本中,得到初始聚类结果 $\boldsymbol{S}_k^1$,聚类中心 $\boldsymbol{U}_k^1$;

3)利用 $\boldsymbol{U}_k^1$ 对聚类样本 $\boldsymbol{S}_k^1$ 进行高维 K-means 聚类,得到聚类结果 $\boldsymbol{S}_k^2$,聚类中心 $\boldsymbol{U}_k^2$;

4)计算 $\boldsymbol{S}_k^2$ 各类数据的每一维数据方差 $\boldsymbol{V}$ 和自适因子 $\boldsymbol{Q}^1$,$\boldsymbol{S}_k^2$ 的方差矩阵为:

$$\boldsymbol{V}=\begin{bmatrix}\sigma_{1,1}^2 & \cdots & \sigma_{1,n}^2\\ \vdots & & \vdots\\ \sigma_{k,1}^2 & \cdots & \sigma_{k,n}^2\end{bmatrix} \tag{1}$$

$$S_j=\sum_{i=1}^{k}\sigma_{i,j}^2 \tag{2}$$

$$Q_j^1=1/\ln(k_0+s_j^2) \tag{3}$$

其中:S_j 为第 j 维数据的方差累计和,Q_j^1 为第 j 维数据的聚类自适应因子,k_0 为先验系数。

5)采用自适应因子 $\boldsymbol{Q}^1$ 对 $\boldsymbol{S}_k^2$ 进行自适应 K-means 聚类

$$\Delta X_{i,k}=\sum_{j=1}^{n}(X_{i,j}-U_{k,j})^2\times Q_j^1 \quad (1\leqslant k\leqslant K) \tag{4}$$

$\Delta X_{i,k}$ 为数据 X_{ij} 对第 k 类中心的加权平方和,按照 $\Delta X_{i,k}$ 最小分配原则进行聚类,每次迭代需重新计算自适应因子 $\boldsymbol{Q}^1$,得到新的聚类 $\boldsymbol{S}_k^3$,聚类中心 $\boldsymbol{U}_k^3$;

6)通过计算 $\boldsymbol{S}_k^3$ 方差矩阵构造自适应因子 $\boldsymbol{Q}^2$

$$Q_j^2=Q_j^1/\sigma_{k,j}^2=\frac{1}{\sigma_k^j\ln(k_0+s_j^2)} \tag{5}$$

Q_j^2 为第 j 维数据的聚类自适应因子;

7)采用自适应因子 $\boldsymbol{Q}^2$ 对 $\boldsymbol{S}_k^3$ 进行自适应 K-means 聚类

$$\Delta X_{i,k}=\sum_{j=1}^{n}(X_{i,j}-U_{k,j})^2\times Q_j^2 \tag{6}$$

按照 $\Delta X_{i,k}$ 最小分配原则进行自适应 K-means 聚类,每次迭代需重新计算自适应因子 Q_j^2,得到新的聚类 $\boldsymbol{S}_k^4$,聚类中心 $\boldsymbol{U}_k^4$;

8)聚类结束,最终聚类结果为 $\boldsymbol{S}_k^4$,聚类中心 $\boldsymbol{U}_k^4$。

2　改进聚类算法的验证

便于验证算法的有效性和直观性,取 UCI 标准测试数据 Iris 数据集,共有 150 个样本,分三类:Iris Setosa(第 1～50 条记录)、Iris Versicolour(第 51～100 条记录)、Iris Virginica(第 101～150 条记录),样本由 4 个维数构成,代表该植物的 4 种属性,sepal length、sepal width、petal length 及 petal width。采用先验系数取 $k_0=1.0$,经过文中算法聚类后,其三类数据的聚类结果如图 1 所示。

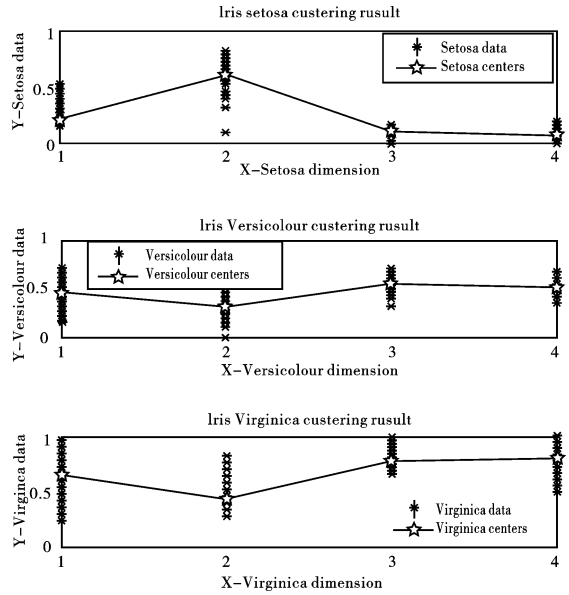


图 1　Iris 数据聚类后效果图

图 1 选取 Iris 方差最大的单维数据经过 3 次聚类,满足收敛条件,有效提高了聚类效率,错分数为 6 个;然后对数据进行 $\boldsymbol{Q}^1$ 自适应全维聚类计算 2 次满足收敛条件,错分数 6 个;最后采用 $\boldsymbol{Q}^2$ 自适应聚类,计算 3 次,错分数为 5 个,最终错分率为 96.66%。比文献[6]算法高出 4 个百分点。

3　改进 K-means 聚类方法在平台惯导系统中的应用

为了保证惯性系统在长期贮存后的使用精度,一般需要对惯性系统进行定期标定,通过对这些长期标定数据进行重复性分析,对评判惯导系统的精度和工艺改进具有重要意义。

以某型号两个批次平台惯导加速度计标定数据为例,选取标定数据均值、极差、方差作为评价参数的重复性指标,得到聚类样本为 96 组三维数据样本,经过离群数据处理和归一化后,其数据分布和聚类中心如图 2 所示。

设定聚类类数为 2,先验系数 $k_0=1.0$,采用改进 K-means 聚类算法,得到结果如图 3 所示。

惯导系统聚类前后的方差对比如表 1 所示。

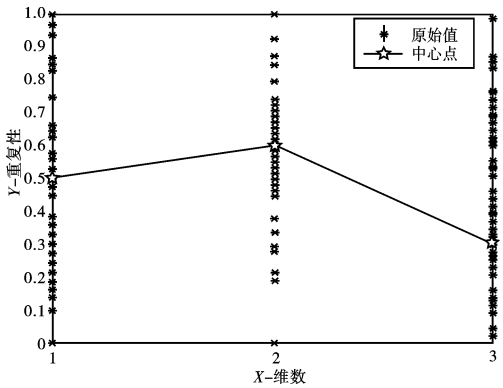
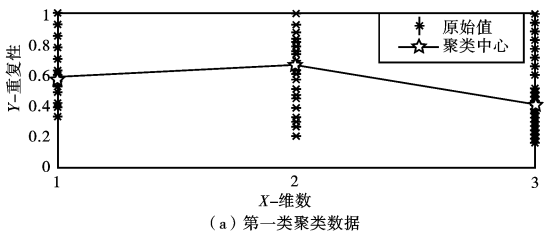
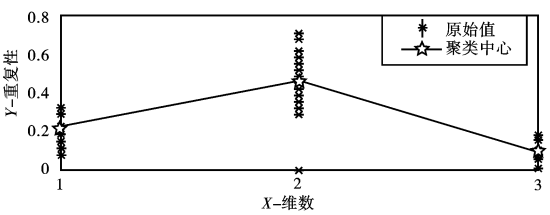


图 2 加速度计重复性评判数据



(a) 第一类聚类数据



(b) 第二类聚类数据

图 3 加速度计重复性评判数据聚类后效果图

表 1 加速度计重复性评判数据聚类方差对比

	第一维	第二维	第三维
聚类前方差	0.0486	0.0313	0.0498
第一类聚类方差	0.0265	0.0255	0.0392
第二类聚类方差	0.0060	0.0178	0.0015

由图 3 聚类数据检查对比，第一类为 70 套产品，第二类为 26 套产品，发现第一类聚类结果中有 27 套第一批次产品，其余 43 套为第二批次产品，第二类聚类结果中有 7 套第一批次产品，其余 19 套为第二批次产品；从表 1 的聚类方差来看，聚类后的方差较聚类前均有明显下降，其中第二类方差更小，说明第二类数据分布更集中，聚类效果很好；从图 3 中聚类中心来看，第二类产品的重复性要好于第一类产品重复性，如果把第二类产品认为是质量最优产品，第一类是次优产品，那么第一批次产品的优秀率为25.93%，第二批次产品的优秀率为 34.88%，从而验证了随着生产经验的丰富或者生产工艺的提高，产品质量也会得到提高，这一结论从总体样本中无法看到，通过改进的 K-means 聚类算法，可以把重复性趋于相同分布规律的产品更好的寻找出来。

4 结论

基于不同类别不同维数数据分布特征提出的自适应 K-means 聚类算法，通过构造自适应因子，在通用 K-means 聚类基础上进行自适应 K-means 聚类，有效提高了聚类精度。对 UCI 标准数据进行仿真验证，结果表明该算法具有很高的聚类精度。最后将所提方法应用于平台惯导系统标定数据重复性聚类中，计算结果表明该方法可以很好的对加速度计标定数据的重复性进行聚类，同时该方法对惯导系统其它标定参数进行聚类和评价也具有很好的应用价值。

参考文献：

[1] Guojun Gan, Chaoqun Ma, Chaoqun Ma. Data clustering: Theory, algorithms, and applications[**M**]. Society for Industrial and Applied Mathematics, 2007.

[2] Sergios T, Konstantinos K. Pattern recognition[**M**]. 3rd ed. [S. l]: Academic Press, 2007.

[3] Gonzalez J Rojas, H Ortega. A new clustering technique for function approximation[**J**]. IEEE Transactions on Neural Networks, 2002, 13 (1): 132—142.

[4] Zhao Y, Karypis G. Hierarchical clustering algorithms for document datasets[**J**]. Data Mining and Knowledge Discovery, 2005, 10(2): 141—168.

[5] Hao-jun Sun, Lang-huan Xiong. Genetic algorithm-based high-dimensional data clustering technique[**C**]// Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009.

[6] 曲福恒，马驷良，胡雅婷. 一种基于核模糊聚类算法[**J**]. 吉林大学学报，2008，46(6)：1137—1141.