K-Means 聚类中序列模式和批量模式的比较研究*)

徐 君 黄亚楼 李 飞
(南开大学信息技术科学学院 天津300071)

摘 要 数据挖掘中的聚类分析对发现数据中隐含的类别和分布有着重要的应用。传统的K-Means聚类算法在给出簇数目的条件下能够对数据进行较好的聚类,算法采用批量模式进行学习,在每一趟数据扫描结束后更新簇中心。序列模式是另外一种学习方式,它每扫描一条记录就更新簇中心。本文提出并实现了基于序列模式的K-Means算法,并与采用批量模式的K-Means算法进行了比较。

关键词 数据挖掘,聚类,序列模式,批量模式

Research on Comparing the Sequential Learning with Batch Learning for K-Means

XU Jun HUANG Ya-Lou LI Fei
(College of Information Science & Technology, NanKai University, Tianjin 300071)

Abstract Clustering, in data mining, is useful for discovering groups and identifying interesting distributions underlying in the data. Classical K-Means algorithm can give a good result when given the cluster number. It uses batch mode to adjust the centers of clusters at the end of each epoch. Sequential mode is another method which updates the centers when each record is scanned. In this paper a K-Means algorithm employing sequential mode is proposed, implemented and compared with algorithm employing batch mode.

Keywords Data mining, Clustering, Sequential mode, Batch mode

1 引言

聚类分析是数据挖掘领域中一个重要的研究课题,它依据数据对象的特点和对象之间的关系把它们分组,其目标是使得分在一个组内的对象具有较大的相似性,而分在不同组中的对象具有较高相异性[1]。到目前为止,大部分的聚类算法都致力于寻找一个对数据对象集的分割,不同的簇是不重叠的,每一个对象最多只能属于一个类。

聚类作为统计学的一个分支,已经被广泛地研究了许多年,大部分工作研究基于距离的聚类分析。基于K-Means、K-Medoids和其他一些方法的聚类分析工具已经被加入到许多统计分析软件包中。在数据挖掘领域,聚类的研究工作主要集中在为大型数据库的有效和实际的聚类分析寻找适当的方法,提高聚类方法的可伸缩性,聚类算法对复杂分布数据和类别型数据聚类的有效性,高维数据聚类技术等方面。

K-Means是应用最广泛的聚类算法之一,它在用户指定类别数的情况下对数据进行聚类。算法尝试找出使平方误差函数值最小的划分,通过反复递推计算出簇中心,把对象指定到不同的簇中去。可伸缩性是K-Means面临的主要问题之一,在需要处理的记录数目很多且目标簇分布比较复杂的时候,算法需要递推多次才能够收敛,这极大地影响了K-Means算法的效率。

从机器学习的角度来看,学习模式可以分为批量模式和序列模式两种。批量模式在扫描完一趟数据后才进行一次参数调整,而序列学习则在每扫描一条记录后就对参数进行调整。这两种学习模式各有优缺点:批量学习能够保证每一次的参数调整都使得目标函数递减,并且最终收敛,但收敛速度和局部极值问题制约了其应用范围;序列模式使得算法在参数空间的搜索具有更大的随机性,并且收敛速度比较快,但随机性也使得在理论上证明它的收敛性比较困难,而且算法受记录输入顺序的影响。经验表明,在实际应用中采用序列模式的学习算法要优于采用批量模式的算法[2]。

本文借鉴了序列学习模式的思想提出传统K-Means算法的一种改进,希望能够克服传统K-Means算法在效率和稳定性方面的不足。

2 基于批量模式的K-Means算法

指定类别数为C,对样本集合进行聚类,聚类的结果由C个簇中心来表示。基于给定的聚类目标函数(或者说是聚类效果判别准则),算法采用迭代更新的方法,每一次迭代过程都是向目标函数值递减的方向进行,最终的聚类结果使目标函数达到一个极小值点,取得较优的聚类效果。根据聚类结果的表达方式又可分为硬K-Means(HCM)算法、模糊K-Means算法(FCM)[5]和概率K-Means算法(PCM)[6]。

设聚类的样本集为:$X=\{\vec{x}_i \mid \vec{x}_i \in R^p, i=1,2,\cdots,N\}$,得到的$C$个簇中心为$\vec{z}_1,\vec{z}_2,\cdots,\vec{z}_c$。令$w_j(j=1,2,\cdots,C)$表示聚类的$C$个簇,则:

$$\vec{z}_j=\frac{1}{n_j}\sum_{\vec{x}\in w_j}\vec{x} \tag{1}$$

定义目标函数:

$$J=\sum_{j=1}^{C}\sum_{i=1}^{n_j} d_{ij}(\vec{x}_i,\vec{z}_j) \tag{2}$$

其中n_j表示w_j类包含的样本个数,一般采用欧氏距离$d_{ij}(\vec{x}_i,\vec{z}_j)=\sqrt{(\vec{x}_i-\vec{z}_j)^T(\vec{x}_i-\vec{z}_j)}$作为样本间的距离,适合于类

*)本文得到教育部重点科学技术研究项目(02038)、天津自然科学基金项目(023600611)资助。

内样本数据为超球形分布的情况。目标函数 J 为每个样本数据点到相应簇中心的距离平方和，即聚类的最小均方误差。

传统的 K-Means 聚类算法(HCM)流程如下：

(1)随机指定 C 个样本点 $\vec{z}_1^1,\vec{z}_2^1,\cdots,\vec{z}_C^1$ 为初始簇中心；

(2)按照距离最近的原则，对样本集合进行分配，确定每个样本的类属关系；

(3)使用式(1)，计算新的簇中心 $\vec{z}_1^k,\vec{z}_2^k,\cdots,\vec{z}_C^k$($k$ 表示迭代次数)；

(4)重复执行(2)～(3)，直到簇中心稳定为止。

K-Means 算法的时间复杂度为 $O(IPNC)$，其中 P 是数据的维度，N 是数据集合 X 的大小，C 是指定的类别数，I 是算法重复执行步骤(2)(3)的次数。一般来说，对于小的数据集合，在有限的几次重复之后，算法能够很快地收敛；与 N 相比较，C 和 P 都要小得多，因此在这种情况下 K-Means 的执行时间是 N 的线性阶函数，算法表现出非常好的可伸缩性。

但重复次数 I 是和数据的数目、数据点的分布以及初始簇中心的选取相关的，随着数据点数目以及数据分布复杂度的增加，I 也迅速地增加，再考虑到初始点选取的影响，此时 I 就成为了不可以忽略的因素，它的迅速增大使得 K-Means 在处理大数据集合时显得非常的不适应，极大地影响了 K-Means 的应用范围。如何提高 K-Means 在处理大数据集合时的效率成为了一个重要的研究方向。

解决办法之一是选取合适的初始簇中心。传统的 K-Means 算法随机地选取初始簇中心，虽然操作起来简单，但是算法效率和结果的正确性都不能得到保证。文[4]提出了利用遗传算法对初始中心进行优化，在很大程度上提高了算法的收敛速度，减少了结果对初始中心的依赖性。

改变算法的学习模式是另外一种解决方法。把传统 K-Means 算法的批量模式改成序列模式，以提高算法的效率和稳定性。文[3]中有对这种思想的简单描述。

本文下节在文[3]基础上深入研究基于序列模式的 K-Means 算法，尔后依照实验结果对采用序列模式的 K-Means 算法和传统采用批量模式的 K-Means 算法进行比较。

3 基于序列模式的 K-Means 聚类算法

K-Means 算法的执行过程可以看成是在参数空间中进行搜索以达到一个最佳的目标，这个搜索的过程是通过对参数的逐步调节实现的：根据输入记录逐步修改聚类模型的参数(簇中心坐标)，直到目标函数(平方误差函数)值达到预定的目标。在实际应用中，参数的调节方式有序列模式和批量模式两种[2]。

在批量模式下，只有当所有的记录都被扫描一遍之后，算法才对模型的参数进行更新。设输入样本集合为 $X=\{\vec{x}_i|\vec{x}_i\in R^P,i=1,2,\cdots,N\}$，算法扫描 $\vec{x}_1,\vec{x}_2,\cdots,\vec{x}_N$，在扫描的过程中模型的参数保持不变，直到一趟扫描完毕之后，才对参数进行一次更新，最终调节值可以看成是所有单个输入数据调节值的一个算术平均。

序列模式又称在线模式或者随机模式，在这种模式下，每扫描一条记录后立即对模型参数进行更新，这样使得在一趟数据扫描中对参数进行多次调节。序列模式所带来的问题是算法执行结果对记录输入顺序的有很强的依赖性。为了避免这个问题，在每一趟数据扫描之前，记录的输入顺序被随机地打乱，使得这种依赖性降到最低程度。

在基于序列模式的 K-Means 算法中，每扫描到一个记录点 $\vec{x}$，设样本 $\vec{x}$ 点原属于簇 w_i，w_i 中当前含有 n_i 个样本，中心为 $\vec{z}_i$；此时簇 w_j 的中心 $\vec{z}_j$ 距离 $\vec{x}$ 最近，则 $\vec{x}$ 被分配给簇 w_j，则对两个簇的参数按照如下规则进行更新：

$$\vec{z}_i=\frac{n_i\cdot\vec{z}_i-\vec{x}}{n_i-1},n_i=n_i-1 \tag{3}$$

$$\vec{z}_j=\frac{n_j\cdot\vec{z}_j+\vec{x}}{n_j+1},n_j=n_j+1 \tag{4}$$

在整个算法的执行过程中，每扫描到数据 $\vec{x}_k$，会出现如下三种情况之一：

1. 第一次扫描到数据 $\vec{x}_k$(此时 $\vec{x}_k$ 没有所属的簇)。$\vec{x}_k$ 距离的簇 w_i 的中心 $\vec{z}_i$ 最近，则按照(4)更新簇 w_i；设置 $\vec{x}_k$ 的所属簇为 w_i。

2. 扫描到数据点 $\vec{x}_k$，$\vec{x}_k$ 原来属于簇 w_i，当前 $\vec{x}_k$ 与簇 w_j 的中心 $\vec{z}_j$ 最近，则 $\vec{x}_k$ 被分配给簇 w_j，按照(3)更新簇 w_j，按照(4)更新簇 w_j；改变 $\vec{x}_k$ 所属簇为 w_j。

3. 扫描到数据点 $\vec{x}_k$，$\vec{x}_k$ 原来属于簇 w_j，当前 $\vec{x}_k$ 仍然与簇 w_j 的中心 $\vec{z}_j$ 最近，则不做任何改变，继续扫描下一个数据点。

结束条件：当扫描完一遍数据后，没有任何记录所属的簇发生改变，则算法收敛到了一个极值点，算法退出。

基于序列更新的 K-Means 聚类算法流程如下：

(1)随机指定 C 个样本点 $\vec{z}_1^1,\vec{z}_2^1,\cdots,\vec{z}_C^1$ 为初始簇中心，并把这 C 个点的所属簇分别标记为 $w_1,w_2,\cdots,w_C$，其余的样本点标记为无效。

(2) 按照距离最近的原则，逐个扫描样本点 $\vec{x}_k$，$k=1,2,\cdots,N$，按照情况1、2、3之一更新 $\vec{x}_k$ 的所属类别和对应的簇参数。

(3) 扫描完一遍样本记录，打乱样本点的输入顺序。

(4) 重复执行(2)～(3)，直到样本的分配不再发生变化为止。

与批量模式相比，序列模式在一趟扫描的过程中需要多次对参数进行更新，并且在每一趟扫描之前要打乱记录的输入顺序以避免扫描顺序对结果产生影响，这样需要占用更多的处理时间；但是序列模式有更快的收敛速度，并且它的参数更新方法使得在参数空间中搜索具有更大的随机性，算法不容易陷入某一个局部极值。

基于序列模式的 K-Means 算法和基于批量模式的 K-Means 算法的结束条件是一致的。当中心稳定下来后，所有数据点的分配不再发生变化；同时，在所有数据点的所属类别在一趟扫描中都没有发生变化后，中心自然也就稳定不再变化。此时算法收敛到了一个极值点。

4 实验分析

在这一节，通过实验结果比较采用两种模式的 K-Means 算法在执行效率和聚类结果上的差异。

实验采用的数据是随机产生的二维的数据集合，它的分布如图1所示，一共5个类别，中心分别为(0,0)，(3,3)，(−3,3)，(3,−3)和(−3,−3)，每一个类别的每一维都服从正态分布，方差为1。一共10000条数据，每一个类别2000条。

我们在相同的平台下(Win2000，VC6.0)分别实现基于批量模式和序列模式的 K-Means 算法，并详细记录下了它们对以上数据集合进行聚类的重要中间数据(如簇中心移动、目标函数值变化、扫描记录趟数、运行时间等)和聚类结果，基于这些数据和结果对两种模式进行了详细的对比分析。

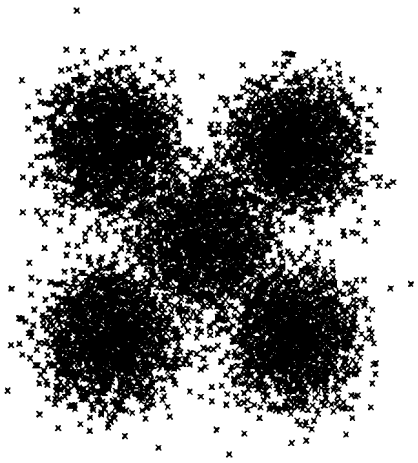


图1 实验数据集

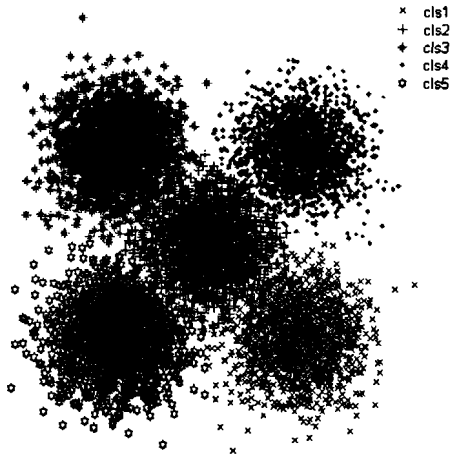


图3 序列模式 K-Means 执行结果

聚类结果 由上一节分析可知，基于批量模式和序列模式的算法在结束条件上是一致的，因此它们的聚类结果也是一致的，实验验证了这一点。图2和图3分别是应用两种模式的 K-Means 的算法在图1所示的数据集合上执行的结果，簇数目设置为5。显然两种模式对以上的实验数据都能够得到比较理想的结果。

算法效率 一般说来，序列模式总是显示出它在效率上的优势，对于大问题它比批量模式具有更快的收敛速度[2]。实

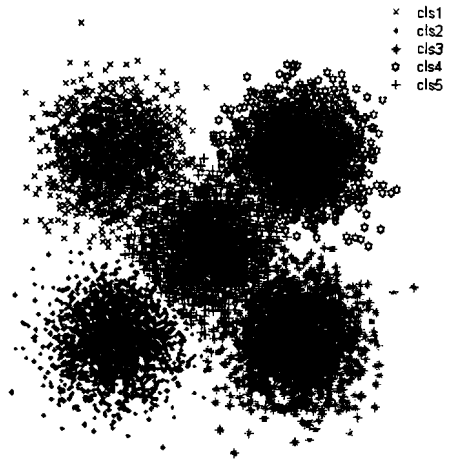


图2 批量模式 K-Means 执行结果

验表明，基于序列模式的 K-Means 算法对数据的扫描趟数远小于传统的基于批量模式的 K-Means 算法，虽然在每一趟扫描中需要多耗费一些时间(用于更新簇信息、打乱数据输入顺序等)，但是从总体来看，基于序列模式的算法在时间上要优于基于批量模式的算法。

在对以上数据集合进行的20次重复实验中，传统的 K-Means 算法平均扫描数据趟数为13.8趟，平均耗费时间为878.75ms。采用序列模式的 K-Means 算法平均扫描数据趟数为5.30趟，平均耗费时间为382.05ms。在其中一次实验中，传统的 K-Means 算法扫描了数据13趟，耗时811ms，表1显示了每趟扫描结束时目标函数值和改变类别的记录数目。基于序列模式的 K-Means 算法扫描数据5趟，耗时361ms，每一趟扫描结束时目标函数值和改变类别的记录数目如表2所示。

图4显示了采用批量模式的 K-Means 算法的一个簇中心的移动情况，横坐标表示中心的移动次数(扫描数据的趟数)，纵坐标表示中心的坐标值，$d1$和$d2$两条折线分别表示中心的 x 坐标和 y 坐标随着数据扫描趟数的增加时坐标值变化的情况。图5是基于序列模式 K-Means 算法的一个簇中心的移动情况，此时的横坐标不再表示数据的扫描趟数，由于平均一个簇的数据数目为2000个，因此在中心移动了2000次左右算法完成了第一趟扫描，从图中可以看出在第一趟扫描结束后簇中心已经很接近最终的簇中心。

表1 批量模式 K-Means 每趟扫描结束时目标函数值和改变类别的记录数目

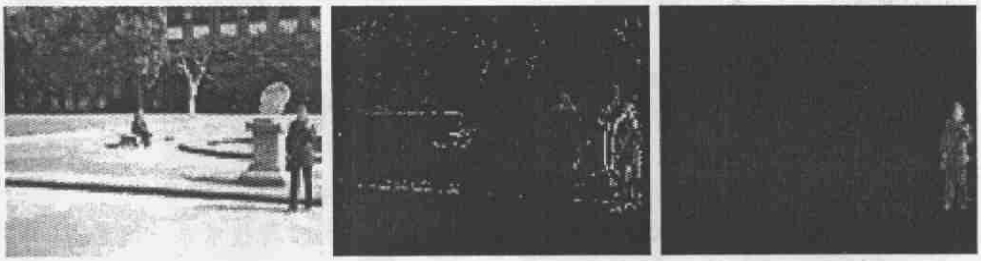
趟数	1	2	3	4	5	6	7	8	9	10	11
目标函数值	60221.507	56907.74	53949.21	48659.57	33714.66	20422.21	19096.99	19027.49	19024.01	19024.01	19024.01
改变类别数	10000	1005	982	1225	1670	1266	345	69	17	1	1

表2 序列模式 K-Means 每趟扫描结束时目标函数值和改变类别的记录数目

趟数	1	2	3	4	5
目标函数值	19053.59	19024.27	19023.98	19023.97	19023.96
改变类别数	10000	253	54	6	1

综合实验结果可以看出：基于批量模式的 K-Means 算法在前几趟扫描中，簇中心的变化相当剧烈，一趟扫描中要改变类别的记录数目很多，目标函数值也比其最终收敛的极值要大很多；而基于序列更新的算法在一趟数据扫描之后，目标函数很快就收敛到了极值周围，簇中心也基本稳定下来并且非常接近最终的结果，在后面的几趟扫描数据只是对中心做出了微小的调节，需要改变类别的数据点个数很少，并且这些数据都是处于类边缘的噪声数据，忽略这些噪声的类别分配可以进一步改善算法的效率，这使得基于序列模式的 K-Means 算法具有更大的改进空间。可以设定一个噪声比例的阈值 T，$0 \leqslant T < 1$。当在某一轮循环中，改变类别的数据数目与总数目的比小于时，算法结束，这样在不显著牺牲聚类质量的前提下可以较大地改善了算法的效率。

(下转第193页)



(a) 原始帧　　(b) 初始运动模板　　(c) 分类得到的运动对象

图9　第1365帧分割过程，运动对象在靠近视频边界的位置

图10　校园视频动态拼接结果，仅仅背景部分，参考帧第1155帧，大小636×274

图11　校园视频动态拼接结果，放置上参考帧的运动对象

结论　本文采用多重分层叠代算法，有效地估计全局运动参数，并得到初始运动模板，并且利用了图像的颜色和纹理特征和视频运动信息，采用结合空间合并和时间运动合并的有效方法以及鲁棒的且可以得到精确分割边缘的区域分类方法，一方面剔除运动目标附近的背景边缘，又避免摭挡问题对估计全局运动参数精度的影响，为拼接图的精确性得到了有力的保证。而且在图像合成时我们解决了拼接图可能产生模糊或某些区域不连续等问题。最后实现了动态视频图像序列高质量的全景图像拼接。

参考文献

1　Nicolas H. New Methods for Dynamic Mosaicking. IEEE Trans. Image Processing, 2001, 10(8): 1239～1251

2　Candocia F M. Synthesizing a Panoramic Scene with a Common Exposure via the Simultaneous Registration of Images. In: Proc. of the 15th Florida Conf. on Recent Advances in Robotics (FCRAR 2002), Miami, FL., May 2002

3　Can A, Stewart C V, Roysam B. Robust Hierarchical Algorithm for Constructing a Mosaic from Images of the Curved Human Retina. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Fort Collins, Colorado, June 1999. 286～292

4　Zhao T, Nevatia R. Stochastic Human Segmentation from a Static Camera. In: Proc of IEEE Workshop on Motion and Video Computing (WMVC'02), Orlando, Florida, 2002. 9～14

5　Lu Peizhong, Wu Lide. Double Hierarchical Algorithm for Video Mosaics. Lecture Notes in Computer Sciences 2195, Springer, Oct. 2001. 206～213

6　Mech R, Wollborn M. A Noise Robust Method for 2D Shape Estimation of Moving Objects in Video Sequences Considering a Moving Camera. Signal Processing, 1998, 66(2): 203～217

7　Yu W, Fritts J, Sun F. A Hierarchical Image Segmentation Algorithm. In: IEEE Intl. Conf. on Multimedia and Expo, Lausanne, Switzerland, Aug. 2002

(上接第158页)

结束语　本文设计实现了传统基于批量模式的 K-Means 算法和基于序列模式的 K-Means 算法，并通过实验对两种算法的结果和效率进行了详细的比较分析，实验结果表明序列模式在效率上要优于传统 K-Means 采用的批量模式。

序列模式的不足之处在于它对记录输入顺序的依赖性很强，虽然在每一趟扫描开始之前打乱数据的输入顺序可以部分解决这个问题，但是随之而来的另外一个问题是当数据量很大，需要存储在外部存储器上的时候，这种做法的有效性受到了很大的限制。

参考文献

1　Han Jiawei, Micheline Kamber 著，范明等译. 数据挖掘：概念与技术. 机械工业出版社，2001

2　Haykin S. Neural Networks: A Comprehensive Foundation, 2nd Ed. 1999, Prentice-Hall: Upper Sadle River, New Jersey

3　An Introduction to Cluster Analysis for Data Mining, 2000. http://www.cs.umn.edu/~han/dmclass/

4　李飞，薛彬，黄亚楼. 初始中心优化的 K-Means 聚类算法. 计算机科学，2002，29(7)：94～96

5　Baraldi A, Blonda P. A Survey of Fuzzy Clustering Algorithms for Pattern Recognition. Parts I and II. IEEE Trans. on Systems, Man and Cybernetics, 1999, 29: 778～785, 786～801

6　Krishnapuram R, Keller J M. A Possibilistic Approach to Clustering. IEEE Transactions on Fuzzy Systems, 1993, 1: 98～110