第34卷 第16期 计 算 机 工 程 2008年8月
Vol.34 No.16 Computer Engineering August 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)16—0224—03 文献标识码：A 中图分类号：TP301.6

基于随机游走模型和KL-divergence的聚类算法

何会民
(邯郸学院计算机系，邯郸 056005)

摘 要：聚类分析在数据挖掘领域有着广泛的应用，该文提出一个聚类新思路，它不需要任何参数的假设，只基于数据两两之间的相似性。该方法假设数据点之间存在随机游走关系，根据数据相似性构造随机游走过程的转移矩阵，当随机游走过程进入收敛期后，t 阶转移矩阵揭示了数据点的分布。用迭代方法寻找最小的KL-divergence来对这些分布聚类。该方法具有严谨的概率理论基础，避免了传统算法需要参数假设、限于局部最优等不足。实验表明，该算法具有较优的聚类效果。

关键词：聚类；随机游走；KL散度

Clustering Algorithm Based on Random Walk Model and KL-divergence

HE Hui-min
(Computer Science Department, Handan College, Handan 056005)

【Abstract】 Clustering analysis is broadly applied in data mining. This paper presents a new idea in clustering based on pair-wise similarities, and assumes no parametric statistical model. Similarities are transformed to a Markov random walk probability matrix. It is assumed the dataset is under a Markov random walk process. When the process is going into convergence, the *t*-step transform matrix indicates the distribution of the dataset. It uses iterative algorithm to cluster these data with the goal of decreasing KL-divergence. This method has a solid foundation of probability theory, which can avoid some insufficiency of the traditional algorithms. The experiment shows the algorithm can achieve better results than K-means and mixture models.

【Key words】 clustering; random walk; KL-divergence

1 概述

聚类分析[1]是基于数据间的相似性将数据点分到不同簇中的一类算法，它属于无监督学习的一种。在很多领域有着广泛的应用，比如数据挖掘、图像分割、模式识别等。在很多聚类问题中，由于很少有关于数据的先验知识，因此聚类算法必须利用数据间的关系来估计其分布结构以便对数据做出较优的划分。

1.1 问题定义

给定一个数据集 $\boldsymbol{X}=\{x_n\}_{n=1}^N$，聚类算法找出原始数据集 $\boldsymbol{X}$ 的一个划分 $\{\boldsymbol{X}_k\}_{k=1}^K$，使得 $\cup_{k=1}^K \boldsymbol{X}_k=\boldsymbol{X}$，并且 $\boldsymbol{X}_k \cap \boldsymbol{X}_l=\phi$ $(k\neq l)$。这里 $\boldsymbol{X}_k$ 就是聚类算法找出来的簇。

目前，聚类分析算法主要分为两大类：一是基于层次的，如单连接、全连接等；二是基于划分的，如K-means[2]、混合模型[3]等。这些方法存在一些不足，如在混合模型中，EM算法用来计算混合概率密度，但是这需要关于模型分布及其参数的假设；另外由于EM只是找到局部最优，这样就需要算法运行多遍，取不同的起始点以找到一个全局较优的结果。K-means也存在相似的问题。本文主要研究第二类方法，笔者提出了一个基于随机游走模型和KL-divergence的聚类 算法。

1.2 随机游走模型

将数据点看成是一个全连通图上的节点，将两节点之间的距离转换为随机游走过程中两状态之间的转移概率，这样可以得到一个 $N\times N$ 的转移矩阵。假设每个节点上都有一个可运动的粒子，根据刚才的转移矩阵，这些粒子就可以在各个节点上随机游走，根据不可约非周期马尔科夫过程的极限性质，最后各个粒子将会收敛到一个平稳分布[4]。这样，在粒子游走进入收敛期后，就可以根据这N个粒子的分布对其进行聚类。对于这N个粒子的分布，尽管K-means可以用来对其聚类，但是它不能很好地反映粒子间的关系，因为欧式距离并不能较好地解释概率分布之间的关系。而信息论中的KL-divergence[5]就可以很好地表示分布间的关系，因此本文的算法采用该指标。

1.3 KL-divergence

在信息论中，KL-divergence是度量2个分布 P 和 Q 之间差异性的指标。通常 P 代表观测数据的分布；Q 代表一个模型，或者一个假设的分布等。对于离散分布的情形，P 和 Q 之间的KL-divergence定义为：

$$D_{KL}(P\|Q)=\sum_i P(i)\lg\frac{P(i)}{Q(i)}$$

2 算法描述

2.1 相似性度量

假设数据集 X 中的所有点都是来自某个度量空间(metric space)，引入符号 $\boldsymbol{I}_k$ 指示第 k 个簇中的所有数据点。这样聚类结果就是数据集 X 的 k 个划分 $\boldsymbol{I}=\{\boldsymbol{I}_k\}_{k=1}^K$。$X$ 中任意2个数

作者简介：何会民(1967－)，男，副教授，主研方向：数据挖掘，人工智能，图像处理

收稿日期：2007-09-26 **E-mail：**bhj33@163.com

据点 x_i 和 x_j 之间的相似性可以由一个非负单调递减的函数 $d(x_i,x_j)$ 来计算。相似度越大，则它们越相邻。该相似性函数可根据实际问题而定，如对于数据点是集合的情形，可以利用集合差来定义相似度 $d(x_i,x_j)=1-\frac{\|(x_i-x_j)\cup(x_j-x_i)\|}{\|x_i\|+\|x_j\|}$。对于连续的情形，可以利用高斯核函数定义其相似度为 $d(x_i,x_j)=\exp(-\frac{\|x_i-x_j\|^2}{2\sigma^2})$。对所有数据点计算其两两之间的相似度就可以得到一个 $N\times N$ 的矩阵 $\boldsymbol{D}$，其中元素 $D_{ij}=d(x_i,x_j)$。于是，一个粒子从点 x_i 移动到点 x_j 的概率定义为 $P_{ij}=\frac{D_{ij}}{\sum_{j=1}^{N}D_{ij}}$。从该式中可以看出，当 $\|x_i-x_j\|$ 减小时 P_{ij} 增大，也就是说 2 个点越相似，它们之间的转移概率也越大，越有可能被聚在一类。

2.2 马尔科夫链

在时刻 t 某粒子从 x_i 开始移动到下一时刻的点 x_j 的概率 $P_{ij}=P(x_j(t+1)\,|\,x_i(t))$。在 2.1 节利用距离来近似表示转移概率得到转移矩阵后，就可以将数据点看成是一个全连通图上的节点，想象每个节点上都有一个粒子开始运动，它们按转移概率随机游走到其他节点。这样，某个粒子 x_i 在 t 步转移之后的概率分布就是 t 阶转移矩阵的第 i 行 P_i^t。于是知道某个起始点在 x_i 的粒子在经过 t 步转移之后到达节点 x_j 的概率为 $P(x_j(t)\,|\,x_i(0))=P_{ij}^t$。根据马尔科夫过程的性质知道它将收敛到一个平稳分布 π，这样就可以避免初始分布的不确定性，从而使得聚类更能体现数据之间的真实关系。

2.3 聚类算法

2.3.1 确定游走阶数 t

由于马尔科夫过程最终将收敛到平稳分布，它与初始分布无关。从信息论的角度来说，初始点的分布信息完全丢失了。这样引入互信息[6]来衡量随机游走过程中的信息损失量。2 个随机变量 X 和 Y 的互信息定义为：$I(X;Y)=\sum_{x\in X,y\in Y}p(x,y)\lg(\frac{p(x,y)}{p(x)p(y)})$。于是，在马尔科夫过程中，某个初始点 $X(0)=\{x_j(0)\}$ 和它随机游走到t时刻的点 $X(t)=\{x_i(t)\}$ 的互信息为

$$I(t)=I(X(0);X(t))=\sum_j p_j\sum_i P_{ji}^t\lg\frac{P_{ji}^t}{p_i^t}=\sum_j p_j D_{KL}[P_{ji}^t\,\|\,p_i^t]$$

其中，p_j 是节点 x_j 的先验分布，假设其为均匀分布 $p_j=\frac{1}{N}$；$p_i^t=\sum_j p_{ji}^t p_j$ 是节点 x_i 在时刻 t 的概率；D_{KL} 代表 KL divergence。知道根据马尔科夫过程的收敛性，当 $t\to\infty$ 时，p_{ji}^t 将收敛到平稳分布 π。也就是互信息 $I(t)$ 递减到 0。这样定义一个阈值 $\varepsilon(0<\varepsilon<1)$，当 $I(t)<\varepsilon$ 时，马尔科夫过程进入收敛期，其 t 阶转移矩阵反应了数据间的本质关系，揭示了其稳定的簇结构，于是就可以用下面的 K 原型簇算法对其进行聚类。

2.3.2 K 原型簇算法

算法的目标是要找到 K 个簇的分布 $\{\boldsymbol{Q}_K\}_{k=1}^K$，称它为原型簇，及一个划分 $\boldsymbol{I}$ 使得目标函数 $J(\boldsymbol{Q},\boldsymbol{I})=\sum_{k=1}^{K}\sum_{m\in I_k}D_{KL}(\boldsymbol{p}_m^t\,\|\,\boldsymbol{Q}_k)=\sum_{k=1}^{K}\sum_{m\in I_k}\sum_{n=1}^{N}\boldsymbol{P}_{mn}^t\ln\frac{\boldsymbol{P}_{mn}^t}{\boldsymbol{Q}_{kn}}$ 的值最小。因为让各个簇的 KL-divergence 最小反应了簇内距离最小化的原则。当 $\boldsymbol{Q}^{(\text{old})}$ 和 $\boldsymbol{I}^{(\text{old})}$ 已知时，可以通过迭代的方法找到更优的新值 $\boldsymbol{Q}^{(\text{new})}$，$\boldsymbol{I}^{(\text{new})}$ 使 $J(\boldsymbol{Q}^{(\text{new})},\boldsymbol{I}^{(\text{new})})<J(\boldsymbol{Q}^{(\text{old})},\boldsymbol{I}^{(\text{old})})$。算法如下：

算法 1 K 原型簇聚类算法

输入 t 阶转移矩阵 $\boldsymbol{p}^t$，簇数目 K。

输出 原始数据集 $\{x_1,x_2,\cdots,x_n\}$ 的一个划分，以及 $K\times N$ 的原型簇矩阵 $\boldsymbol{Q}$，矩阵的行 $\boldsymbol{Q}_k$ 代表簇 k 的分布。

算法过程：

(1)初始化 $\boldsymbol{Q}^{(\text{old})}$。

(2)为簇 k=1,2,⋯,K 寻找新的划分以减小目标函数 $J(\boldsymbol{Q},\boldsymbol{I})$，对每个节点 x_m，寻找与其最相似的簇 k，将其赋给簇 k。即 $\boldsymbol{I}_k^{(\text{new})}=\{m:k=\arg\min_k D_{KL}(\boldsymbol{p}_m^t\,\|\,\boldsymbol{Q}_k^{(\text{old})})\}$。

(3)更新原型簇矩阵：对 k=1,2,⋯,K，令 $\boldsymbol{Q}_k^{(\text{new})}=\frac{1}{|\boldsymbol{I}_k^{(\text{new})}|}\sum_{m\in\boldsymbol{I}_k^{(\text{new})}}\boldsymbol{p}_m^t$。

(4)如果 $J(\boldsymbol{Q}^{(\text{old})},\boldsymbol{I}^{(\text{old})})\geq J(\boldsymbol{Q}^{(\text{new})},\boldsymbol{I}^{(\text{new})})$，令 $\boldsymbol{Q}^{(\text{old})}=\boldsymbol{Q}^{(\text{new})}$，转(2)；否则，算法停止。

2.3.3 $\boldsymbol{Q}$ 的初始化

2.3.2 节算法 1 中的(1)需要初始化 $\boldsymbol{Q}^{(\text{old})}$，它可以初始化为任意的 K 个分布，但没有涵盖更多关于数据的初始信息，这样会导致算法 1 收敛较慢。因此，需要初始化 $\boldsymbol{Q}^{(\text{old})}$ 使它涵盖尽可能多的数据信息并且使簇之间离得尽可能远；这样可以取{ $\boldsymbol{P}_n^t,1\leq n\leq N$ }的均值作为第一个簇的原型，其他簇的原型选{ $\boldsymbol{P}_n^t,1\leq n\leq N$ }中的某个向量 $\boldsymbol{P}_m^t$ 使它到与它最近的原型簇的距离最大。算法如下：

算法 2 初始化 $\boldsymbol{Q}$

输入 转移矩阵 $\boldsymbol{P}^t$，聚类的簇数目 K

输出 初始化的原型簇 $\boldsymbol{Q}$

算法过程：

(1) $\boldsymbol{Q}_1=\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{P}_n^t$

(2)for k=2,3,⋯,K

$z=\arg\max_n\min_{j=1,2,\cdots,k-1}D_{KL}(\boldsymbol{P}_n^t\,\|\,\boldsymbol{Q}_j)$，$\boldsymbol{Q}_k=\boldsymbol{P}_z^t$

(3)end

综上，给定一个 t 阶转移矩阵和簇数目 K，算法 1 和算法 2 就可以通过简单的迭代过程将数据集分成 K 个簇。考虑算法执行的时间代价：因为 KL-divergence 对于变量数目为 N 的分布的计算代价为 $O(N)$，因此算法 1 的执行代价为 $O(cKN^2)$，c 是算法 1 的迭代次数(通常 $c<<N$)；算法 2 的执行代价为 $O(KN^2)$；因此算法总的时间代价为 $O(N^2)$。可见它的时间复杂度不高于现有算法。

3 实验

对UCI[7]的手写数字点阵图数据进行聚类，从中随机抽取总大小为 1 000 的含有数字 2, 5, 6, 9 的数据集。如图 1 所示，每个数字图形都是由 32×32 的点阵构成，这样每个点阵图可以看作是 $R^{1\,024}$ 空间上的一个数据点，这里簇数目、样本属性都已知，聚类结果以误分到其他簇的样本比例来衡量。使用sourceforge提供的weka中的K-means和混合模型算法来与本文的算法进行对比，考虑均匀分布(每个数字对应的样本容量为 250)和高斯分布(数字 2, 5 各 400 个样本；数字 6, 9 各 100 个样本)的情形得到聚类的准确率比较分别如图 2 和 图 3 所示。

图 1　数字点阵图

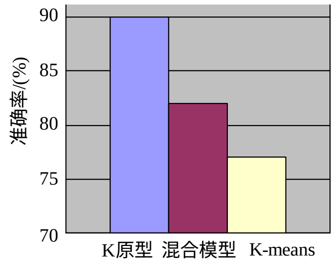

图 2　均匀分布下算法准确率比较

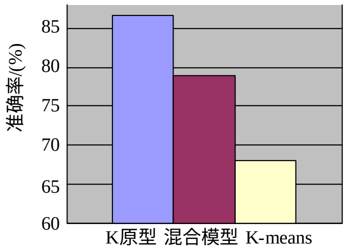

图 3　高斯分布下算法准确率比较

在聚类的结果中，K-means 和混合模型容易混淆含有数字 2, 5 的情形；而 K 原型簇算法则比较健壮。

4　结束语

本文提出了一个非参数聚类的新方法，在将来的工作中，笔者将进一步考虑特征值理论以寻找更高效的方法，并考虑改进聚类的度量指标。

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] MacQueen J B. Some Methods for Classification and Analysis of Multivariate Observations[C]//Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, USA: [s. n.], 1967: 281-297.

[3] Vlassis N, Likas A. A Greedy EM Algorithm for Gaussian Mixture Learning[J]. Neural Processing Letters, 2002, 15(1): 77-87.

[4] Norris J R. Markov Chains[M]. Cambridge, UK: Cambridge University Press, 1997: 40-46.

[5] Kullback S, Leibler R A. On Information and Sufficiency[J]. Annals of Mathematical Statistics, 1951, 22(1): 79-86.

[6] Tishby N, Slonim N. Data Clustering by Markovian Relaxation and the Information Bottleneck Method[C]//Proc. of the Advances in Neural Information Processing Systems. Denver, CO, USA: [s. n.], 2000: 640-646.

[7] Asuncion A. UCI Machine Learning Repository[EB/OL]. (2007-05-11). http://www.ics.uci.edu/$\sim$mlearn/MLRepository.html.

(上接第 223 页)

受到环境噪声的影响；图 5(c)采用基于最小错误率的贝叶斯决策方法，相对于前一种方法，它在更大程度上保留了车辆运动信息，但是由于推理的结果在很大程度上依赖于先验概率，当接触到实际问题时，可以发现使错误率最小并不一定是一个普遍适用的最佳选择，如把前景错分为背景，或相反方向的错误，图中体现为车辆周围冗余的信息，而且当车辆运动速度较慢的时候，这种通过统计同一位置像素颜色特征进行决策的方法就很可能把车辆当做背景，造成误检测；图 5(d)是采用 BSFTG 检测方法的效果，在图中保留了较为完整准确的车辆运动信息，并能更好地抵制环境噪声影响，车辆信息较为清晰，对运动速度慢的车辆也有很好的检测效果。图 5(e)为采用新算法获得的车辆检测结果。

5　结束语

本文提出一种改进的适合于复杂背景下视频车辆检测的方法 BSFTG，与传统方法不同之处在于它兼顾视频序列时间和空间上的相关性，充分利用时空信息，通过隔帧对称差分结合背景补偿的方式弥补了两者相互间的不足，保证运动目标信息的完整性和准确性，提高了车辆检测效率。在阈值处理方面则综合考虑像素的灰度信息和空间分布信息，采用改进的二维阈值法替代传统一维方法，提高了车辆目标分割的准确度，并结合改进的遗传算法加快寻优求解的速率，使系统能够满足实时性的要求，具有更好的抗噪能力与检测效率。

参考文献

[1] Foresti G L. Object Recognition and Tracking for Remote Video Surveillance[J]. IEEE Transactions on Circuits and Systems for Video Technology, 1999, 9(7): 1045-1062.

[2] Abutaleb A S. Automatic Thresholding of Gray-level Picture Using Two-dimensional Entropies[J]. Pattern Recognition, 1989, 47(1): 22-32.

[3] Otsu N. A Threshold Selection Method from Gray-level Histograms[J]. IEEE Trans. on SMC, 1979, 9(1): 62-69.

[4] Ding Lifen, Chen Zhiwu, Chen Zhenfeng. Simulation and Research on the Control Parameters of Genetic Algorithm[J]. Science & Technology Information, 2007, 36(1): 618-621.

[5] Li Liyuan, Huang Weimin, Irene Y. Foreground Object Detection from Videos Containing Complex Background[C]//Proceedings of the 11th ACM International Conference on Multimedia. New York, USA: ACM Press, 2003: 2-8.