Hans Journal of Data Mining 数据挖掘, 2017, 7(3), 70-76
Published Online July 2017 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2017.73008

Hans汉斯

Data Clustering Based on Random Walk

Wei Cui[1], Xun Xia[1], Yulu Sun[2*]

[1]Luzhou Vocational and Technical College, Luzhou Sichuan
[2]College of Electronic & Information Engineering, Sichuan University, Chengdu Sichuan
Email: *sunylcn@163.com

Received: Jun. 28th, 2017; accepted: Jul. 17th, 2017; published: Jul. 20th, 2017

Abstract

In order to realize the clustering analysis of large data volume and complex types of data, the random walk algorithm maps the data set into graphs, each data represents node, and uses a weighting function to represent the relationship between data and data. The similarity criterion indicates the weight between two data in the data set. In the random walk algorithm, the weight of the weight represents the random walker from the non-seed point for the first time to reach a seed point of preference. Finally, cluster analysis is realized according to the maximum transition probability. The results show that the random walk algorithm can achieve clustering in the clustering analysis of numerical data.

Keywords

Clustering Analysis, Random Walk Algorithm, Weighting Function

基于随机游走的数据聚类

崔　伟[1]，夏　汛[1]，孙瑜鲁[2*]

[1]泸州职业技术学院，四川 泸州
[2]四川大学电子信息学院，四川 成都
Email: *sunylcn@163.com

收稿日期：2017年6月28日；录用日期：2017年7月17日；发布日期：2017年7月20日

摘　要

为了实现大数据量、复杂类型数据的聚类分析，本文运用随机游走算法是将数据集合映射为图，各个数据表示节点，用一个加权函数表示数据与数据之间的关系，该加权函数能根据相似性准则表示数据集中

*通讯作者。

文章引用: 崔伟, 夏汛, 孙瑜鲁. 基于随机游走的数据聚类[J]. 数据挖掘, 2017, 7(3): 70-76.
https://doi.org/10.12677/hjdm.2017.73008

两个数据间的权重。在随机游走算法中，权重的大小代表了随机游走者从非种子点第一次到达某一种子点的偏好。最后根据最大转移概率实现聚类分析。结果表明随机游走算法在数值型数据的聚类分析中能够实现聚类。

关键词

聚类分析，随机游走，权重函数

Copyright © 2017 by authors and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

聚类是按照某个特定准则把已知数据集分成不同的类，同类的数据对象间相似度尽可能大，不同类的数据对象间的相似度尽可能小。聚类分析作为数据挖掘技术中的重要组成部分，目前在许多领域都得到了广泛的研究和应用如模式识别[1]、数据分析[2]、图像处理[3]、市场研究[4]、Web 文档分类[5]等。聚类算法的选择取决于数据的类型及其聚类的目的。根据其基本思想可分为划分、层次、密度、基于网格的方法以及基于模型的方法。

基于划分的主要思想是：首先给定簇数目，然后对数据集采用迭代重定位方法实现划分，划分质量取决于初始种子和聚类标准。*K*-means 算法[6]从数据集中任意选择 *k* 个对象作为初始种子，以最短距离为准则将数据进行分类，该方法以均值表示类中心易受奇异数据的影响，为了抑制异常数据对聚类的影响，*K*-medoids 算法[7]以类个体表示聚类中心。上述算法因采用平方误差作为收敛条件，聚类结果为局部最优解，对此提出了调和均值(KHM)算法[8]的评价函数。基于划分方法的聚类需要先验知识，即事先指定数据类别个数，为了避免个数对聚类结果的影响提出了层次聚类算法如 AGNES 算法[9]，其主要思想是首先将每个数据对象作为初始簇，然后对这些单元类逐层根据距离最近原则进行聚合，使单元簇越来越大直至满足所要求的簇数目为止。AGNES 算法比较简单，但可伸缩性较差。为此提出了 DIANA 算法[10]，该算法将给定数据看作一个大的簇，在每一步迭代过程中将上层簇根据簇的直径或平均相异度分解为更小的簇，直至满足终止条件。传统层次聚类方法聚类过程中会遇到合并或分裂点选择的困难，因此 Guha 等人提出了改进的层次聚类算法 CURE 算法[11]。该方法用具有代表性的若干点代表一个聚类，避免了用所有点或单个质心代表一个簇的传统方法，使其能够识别具有复杂形状和不同大小的聚类，从而对孤立点的处理更加健壮。大多数聚类算法在进行聚类时只估计点与点之间的相似度，这种局部算法很容易出现错误。因此 ROCK 算法[12]在 CURE 算法基础上根据成对点的邻域情况进行聚类比只关注相似度的聚类方法更加鲁棒。基于层次和划分的聚类算法对凸形的聚类簇效果较好，而数字点阵图由于形状变化较大，聚类效果较差且运行时间较长。

为了弥补上述聚类算法的不足，本文利用随机游走算法进行分析。随机游走在聚类分析中的应用首先选择聚类中心，以初始聚类中心为中心，逐个输入样本；利用随机游走算法得到各个初始聚类中心到达输入样本的概率，以最大转移概率为原则将样本归入聚类中心所属的那一类。同时利用均值计算该类的重心，以该重心作为新的聚类中心再输入下一个样本，直到所有数据被分类。

2. 随机游走

随机游走算法将给定数据集合看作固定数目的节点和边的离散对象，然后将数据聚类分析问题转化

为无向加权图来进行求解。对数据集进行统一的定义，首先将数据集映射成一个无向加权图$G=(V,E)$，它由代表数据值的节点$v_i \in V$和表示数据与其相邻数据间关系的边界$e \in E \subseteq V \times V$组成。$e_{ij}$表示连接两个顶点$V_i$和$V_j$的边，每条边被赋予一定的权值，记为$w_{ij}$，表示两个顶点之间的相似或差异程度。顶点$v_i$的度定义为$d_i = \sum w_{ij}$，它等于所有与结点$V_i$相关联的边的权值的和；此外，假设$w_{ij} > 0$且$w_{ij} = w_{ji}$。由于随机游走算法是一种人工交互式的算法，因此用户需要预先根据数据性质设置k个种子点(标记点)，然后为每个未被标记的数据节点分配一个k维向量，来表示一个未被标记点到达所有种子点的随机游走过程，每一维向量表示从每个未标记点出发，第一次到达k个种子点的概率。k个概率中最大的值为未标记点所属的类标签，通过该方法具有相似性的数据就可归为一类，从而根据不同类别之间的差异实现数据聚类。

利用径向基核函数定义数据间的相似度即：

$$w_{i,j} = \exp\left(-\frac{\left\|q_i - q_j\right\|^2}{k}\right) \tag{1}$$

其中，k表示聚类数目，q_i, q_j表示数据集中任意两个数据。

3. 随机游走求解

在一定的边界条件下，随机游走转移概率的求解问题与联合狄利克雷求解问题的解相似。因此，可以通过求解联合狄利克雷问题的解来实现随机游走算法求解。在给定区域Ω上狄利克雷积分形式为：

$$D[u] = \frac{1}{2}\int_{\Omega} |\nabla u|^2 \, \mathrm{d}\Omega \tag{2}$$

随机游走从一个非标记点到标记点的概率等于该标记点在边界条件下的狄利克雷函数，求解的问题即在某个边界条件下求解拉普拉斯函数如：

$$\nabla^2 u = \frac{\partial^2 u}{\partial i^2} + \frac{\partial^2 u}{\partial j^2} = 0 \tag{3}$$

组合拉普拉斯矩阵在映射图中定义如下所示：

$$L_{ij} = \begin{cases} d_i & i = j \\ -w_{ij} & v_i \text{与} v_j \text{为相邻接点} \\ 0 & \text{其它} \end{cases} \tag{4}$$

拉普拉斯L_{ij}的值由节点v_i与v_j共同决定，该矩阵是满足边界条件下的对称正定矩阵。d_i为节点v_i的度，定义为$d_i = \sum_{j=1}^{n} w_{ij}$表示$w$第$i$行所有元素之和。

图G的$m \times n$条边即顶点间的关联矩阵为：

$$A_{e_{ij}v_k} = \begin{cases} +1 & i = k \\ -1 & j = k \\ 0 & \text{其它} \end{cases} \tag{5}$$

由上式可知，关联矩阵由边e_{ij}和节点v_k决定，图中e_{ij}为任意方向，A为联合梯度算子，A^T为联合散度算子。

我们构造一个大小为$m \times m$的对角阵C，其对角线上的值为映射图边上的权值即：

$$C_{e_{ij}e_{ks}}=\begin{cases}w(e_{ij}) & i=k,j=s\\ 0 & \text{其它}\end{cases} \tag{6}$$

如果连续，联合梯度算子和联合散度算子之积可以表示各向同性的联合拉普拉斯矩阵即：$L=A^TA$。在映射图中，矩阵C可看作向量上一个加权内积大小的度量，此情况下，当$C=I$时，$L=A^TCA$可简化为$L=A^TA$。因此，调和函数求解问题可通过上述定义解决即：在固定标记点值已知情况下，非标记点到标记点的概率值可求。于是式(2)可转化为：

$$D[x]=\frac{1}{2}(Ax)^T C(Ax)=\frac{1}{2}\sum_{e_{ij}\in E}w_{ij}\left(x_i-x_j\right)^2 \tag{7}$$

式中，L为联合的拉普拉斯矩阵，x为图中数据的概率值，$D[x]$的最小值可通过联合调和函数x求得，映射图中的所有节点可分为未标记点集合V_U和标记点集合集V_M且$V_M\cup V_U=V$，$V_M\cap V_U=\varphi$。将拉普拉斯矩阵按标记点和未标记点排列得：

$$D[x_U]=\frac{1}{2}\left[X_M^T X_U^T\right]\begin{bmatrix}L_M & B\\ B^T & L_U\end{bmatrix}\begin{bmatrix}x_M\\ x_U\end{bmatrix} \tag{8}$$

拉普拉斯矩阵分解为：

$$L=\begin{bmatrix}L_M & B\\ B^T & L_U\end{bmatrix} \tag{9}$$

其中，X_B,X_U分别为标记点和非标记点的随机游走概率值，对$D[x_U]$求X_U的微分得：

$$L_U x_U=-B^T x_M \tag{10}$$

令x_i^s表示未标记点到达标记点为s的概率，定义一个表示所有标记点集合的函数：$Q(v_j)=s,\forall v_j\in V_M$且$0<s\le k$，($k$为种子点数目)。为所有$v_j\in V_M$的点定义一个矩阵：

$$m_j^s=\begin{cases}1 & Q(v_j)=s\\ 0 & Q(v_j)\neq s\end{cases} \tag{11}$$

因此，通过求解：$L_U x^s=-B^T m^s$得到到达单个标记点的概率；

通过$L_U X=-B^T M$求得到所有种子点的概率，其中，k个列矢量x^s组成X，k个列矢量m^s组成M。因为对任意未被标记节点来说，它到所有种子点的概率之和为1，即：

$$\sum_s x_i^s=1\quad \forall v_i\in V \tag{12}$$

因此，对于k个标记种子点来说，计算$k-1$组方程，求可得出$k-1$个概率值。

在获得每个结点v_i第一次到达k个种子点的概率后，逐个比较大小，以最大转移概率$\max_s\left(x_i^s\right)$实现聚类。

4. 实验结果与分析

本文基于随机游走的数据聚类算法流程如图1所示。

为了验证本文算法，首先对来自UCI数字点阵图进行聚类分析，从中随机抽取数字值为0~9中的数据集，如图2所示，每个图形都是有32×32的点阵构成，这样每个点阵图可以看作是R^{1024}空间上的数据点，聚类数目和样本属性已知，聚类结果用信息检索领域的常用评价指标准确率P、召回率R和F测度衡量。使用*K*-means算法和模糊聚类算法和本文算法进行对比，得到的结果如表1所示。

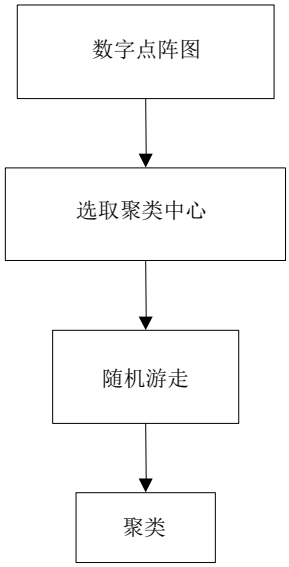

Figure 1. Flow chart of algorithm
图 1. 算法流程图

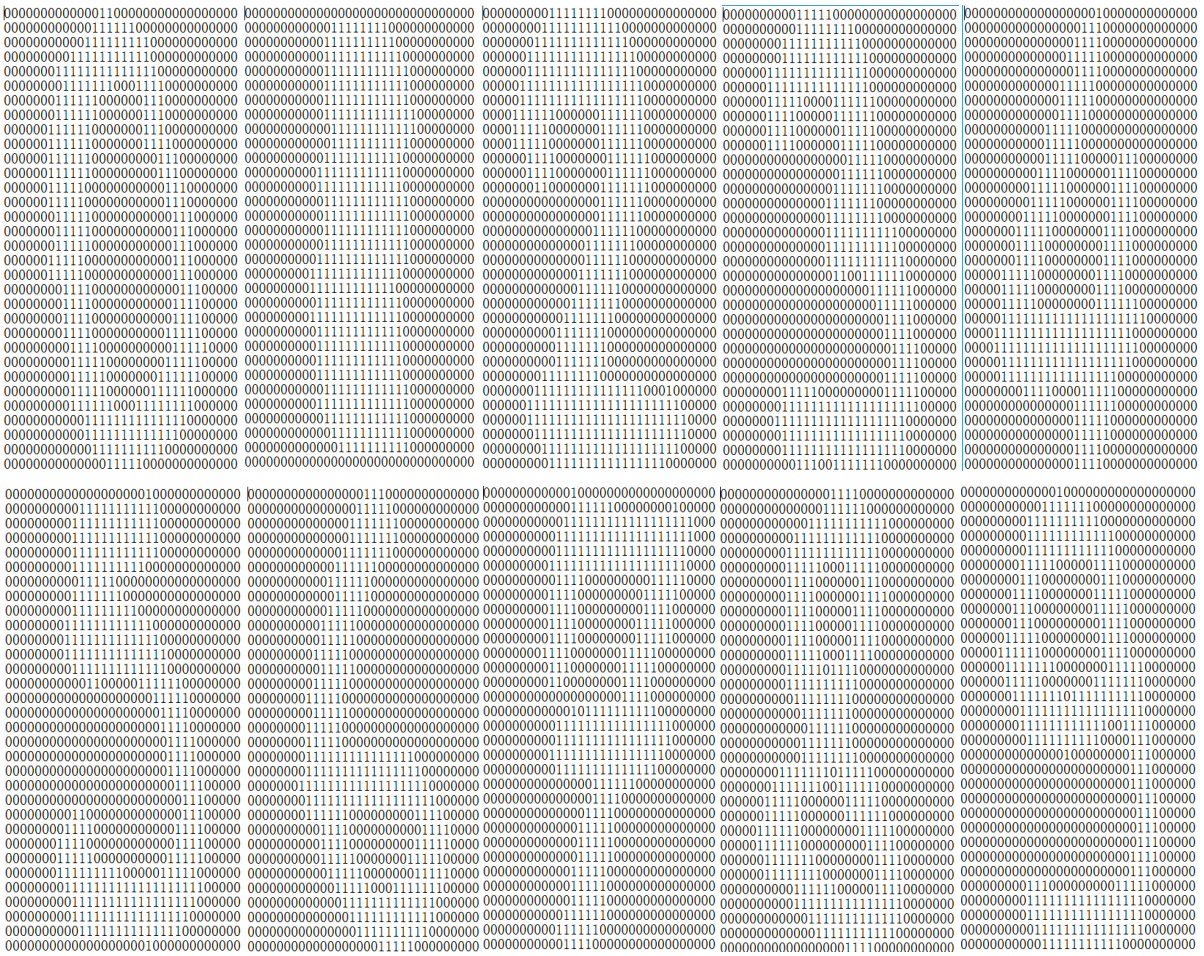

Figure 2. Digital bitmap
图 2. 数字点阵图

Table 1. Evaluation score of digital bitmap
表 1. 数字点阵图的测评指标

数字值	聚类算法	测评分数			运行时间
		P (准确率)	R (召回率)	F 测度	(s)
0	*K*-means	0.993	0.944	0.968	3.511
	模糊聚类	0.860	0.958	0.906	3.956
	本文算法	0.915	0.921	0.918	5.724
1	*K*-means	0.869	0.971	0.917	3.402
	模糊聚类	0.928	0.969	0.948	3.055
	本文算法	0.936	0.992	0.963	5.127
2	*K*-means	0.867	0.940	0.902	3.865
	模糊聚类	0.920	0.982	0.950	4.213
	本文算法	0.964	0.991	0.977	5.624
3	*K*-means	0.915	0.921	0.918	3.966
	模糊聚类	0.993	0.944	0.968	4.452
	本文算法	0.993	0.953	0.973	6.258
4	*K*-means	0.986	0.850	0.913	3.687
	模糊聚类	0.937	0.947	0.942	3.925
	本文算法	0.961	0.997	0.979	5.685
5	*K*-means	0.885	0.968	0.925	3.257
	模糊聚类	0.993	0.953	0.973	4.384
	本文算法	0.978	0.982	0.980	6.258
6	*K*-means	0.860	0.958	0.906	3.139
	模糊聚类	0.869	0.971	0.917	3.839
	本文算法	0.954	0.897	0.924	6.284
7	*K*-means	0.894	0.943	0.918	3.264
	模糊聚类	0.867	0.940	0.902	4.025
	本文算法	0.946	0.983	0.964	5.575
8	*K*-means	0.915	0.921	0.918	3.586
	模糊聚类	0.937	0.947	0.942	3.621
	本文算法	0.950	0.978	0.964	6.362
9	*K*-means	0.885	0.968	0.925	4.208
	模糊聚类	0.894	0.943	0.918	4.803
	本文算法	0.954	0.897	0.924	6.424

从表 1 中可知，*K*-means 算法、模糊聚类和本文算法对从数据集任意选取的 0~9 十个点阵图数字的准确率、召回率和 F 测度差异较小，其 F 测度均在 0.9 以上，但本文算法运行时间低于 *K*-means 算法、模糊聚类算法，由此可知随机游走算法能够应用于聚类分析且耗时较短。为了更直观的评价以上方法，根据表 1 得到求得 *K*-means 算法、模糊聚类和本文算法的 F 测度的均值分别为：0.921、0.934、0.957。由此可以看出利用随机游走对数字点阵图进行聚类分析，与 *K*-means 算法、模糊聚类相比具有一定的鲁棒性。

5. 结语

本文根据随机游走算法原理，将数据集转换为矩阵图的形式然后利用随机游走算法求得样本到各个聚类中心的概率，根据最大概率原则划分样本所属类别，有效实现了随机游走算法在聚类分析中的应用。但由于本文只是针对数字点阵图进行了聚类分析，而没有对更为复杂的数据集进行分析，所以还有待进一步研究和改进。

基金项目

川大-泸州战略合作科技项目(2015CDLZ-S12)。

参考文献 (References)

[1] 黄震华, 向阳, 张波, 等. 一种进行 *K*-Means 聚类的有效方法[J]. 模式识别与人工智能, 2010, 23(4): 516-521.

[2] 汤效琴, 戴汝源. 数据挖掘中聚类分析的技术方法[J]. 微计算机信息, 2003(1): 3-4.

[3] 张鑫, 赵丞. 层次聚类算法在图象处理中的应用[J]. 计算机光盘软件与应用, 2011(11): 23-23.

[4] 黄劲松, 赵平. 聚类分析在品牌市场定位研究中的应用[J]. 数理统计与管理, 2005, 24(1): 21-26.

[5] 王自强, 钱旭. 基于流形学习和 SVM 的 Web 文档分类 4.681 算法[J]. 计算机工程, 2009, 35(15): 38-40.

[6] 吴夙慧, 成颖, 郑彦宁, 等. *K*-Means 算法研究综述[J]. 现代图书情报技术, 2011, 27(5): 28-35.

[7] Park, H.S. and Jun, C.H. (2009) A Simple and Fast Algorithm for *K*-Medoids Clustering. *Expert Systems with Applications*, **36**, 3336-3341. https://doi.org/10.1016/j.eswa.2008.01.039

[8] Gungor, Z. and Unler, A. (2008) *K*-Harmonic Means Data Clustering with Tabu-Search Method. *Applied Mathematical Modelling*, **32**, 1115-1125. https://doi.org/10.1016/j.apm.2007.03.011

[9] Barbakh, W.A., Wu, Y. and Fyfe, C. (2009) Non-Standard Parameter Adaptation for Exploratory Data Analysis. Springer, Berlin Heidelberg, 7-28.

[10] 姚明海. 基于连续性原理的聚类算法研究[D]: [硕士学位论文]. 长春: 东北师范大学, 2010.

[11] 魏桂英, 郑玄轩. 层次聚类方法的 CURE 算法研究[J]. 科技和产业, 2005, 5(11): 22-24.

[12] 王荣, 王飞戈, 吴坤芳. 基于改进 ROCK 算法的个性化推荐系统研究[J]. 河南科学, 2011, 29(11): 1346-1349.

Hans汉斯

期刊投稿者将享受如下服务：

1. 投稿前咨询服务 (QQ、微信、邮箱皆可)
2. 为您匹配最合适的期刊
3. 24 小时以内解答您的所有疑问
4. 友好的在线投稿界面
5. 专业的同行评审
6. 知网检索
7. 全网络覆盖式推广您的研究

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：hjdm@hanspub.org