ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 3,May 2016,pp. 599－605
DOI:10.16337/j.1004-9037.2016.03.021

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于主动学习的微博聚类分析

朱　丽　陆建峰

(南京理工大学计算机科学与工程学院,南京,210094)

摘　要：K-Means聚类算法由于无法准确确定初始化聚类中心,容易造成聚类结果准确率低下。对微博数据聚类时,可能会导致无法正确反映兴趣热点。本文设计了基于主动学习的聚类算法,在确定初始聚类中心过程中应用Min-Max主动学习策略,使得算法每次在很小数量的查询后都会提供数据点供用户进行初始中心点确认,并在K-Means算法中重新计算聚类中心时设置其权重值,从而减少迭代的数量,提高聚类结果的准确率,并将这一算法运用于微博聚类分析,得出微博热门话题。

关键词：主动学习;K-Means;微博

中图分类号：TP3　　**文献标志码**：A

Clustering Analysis of Micro Blogs Based on Active Learning

Zhu Li, Lu Jianfeng

(Department of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China)

Abstract: The K-Means clustering algorithm can not determine the initial clustering centers, which results in low accuracy and inability to reflect the interesting hotspots. Here, algorithm based on clustering is proposed through applying Min-Max active learning strategy to ask the user for identifying the seed points. Several points are provided in small quantities of query for users to confirm the initial centers, and the weight is set in the recalculation of K-Means centers, which reduces the number of iterations and improves the accuracy of clustering results. Moreover, the hot topics are obtained by applying this algorithm to the micro-blog clustering analysis.

Key words: active learning; K-Means; micro blogs

引　言

随着互联网技术[1],尤其是为用户提供交互功能的Web 2.0技术的日益成熟与发展,社交媒体服务网站(如Facebook、人人校内网、Twitter和新浪微博等)为人们的社交生活提供了极大便利。腾讯微博中一条用户状态会限定最多280个字符,不仅更适合现代社会快速生活节奏的需要,也更方便互联网用户通过移动通信终端(如手机、平板电脑)上传和分享感兴趣的信息,同时也方便了商家进行网络微博营销等等一系列操作。所以对微博进行分析具有非常重要的价值。如今微博的信息数据特征是计算机网络和通信等学科领域近年来关注的主要问题[2],很多微博研究基于这些微博数据进行分析[3-5]。主要的研究问题如下:(1)基于微博用户的研究:即对用户的行为特征及用户的影响力进行研究。(2)基于微博

收稿日期:2014-06-05;**修订日期**:2014-06-30

用户关系的研究：即对用户关系网络的基本属性、关系网络生成和演进、微博人员关系以及微博用户人际关系特点等方面进行研究。(3)基于微博内容的研究：即对微博消息内容特点、消息活跃时间特点和微博热点话题特点等方面进行研究。(4)基于微博消息传播的研究：即对微博消息传播特点、微博消息传播影响力等的一系列研究[6]。本文研究在大量微博文本集上的热门话题发现问题，聚类分析是个重要的分析手段，它可以将一组数据按照本身内在规律较合理地分为几类，缩小了全凭主观判断所造成的误差，使数据分析结果更具客观性，且它的应用可以完成人工所不能完成的工作。如果按传统人为观察的话，会带来工作量太大和主观色彩太浓两个弊端，并且还需要丰富的专业知识，否则结果可能无法正确反映数据特点。K-Means 算法[7]是数据挖掘中用于知识发现常用的聚类算法之一。它是一种基于距离的聚类算法，将研究对象的空间距离指标按照相似性准则划分到若干子集中，即两个对象的距离越近，其相似度就越大，从而相同子集中各元素间差别最小，而不同子集中各元素差别最大。设计算法时[8]，先由用户确定聚类的类别数 k，并随机选择 k 个对象，每个对象是一个种子，代表一个簇(类)的均值或中心，对剩余的每个对象，根据其与各簇中心的距离将它赋给最近的簇。然后重新计算每个簇内对象的平均值形成新的聚类中心，这个过程重复进行，直到满足一定条件，如标准测度函数收敛为止。

但 K-Means 存在两个主要缺陷[8]：(1)k 事先给定，而 k 非常难以选定。很多时候，事先并不知道给定数据应该分成几类才最合适；(2)K-Means 需要随机选取种子作为聚类初始中心，这可能会对聚类结果产生较大影响，不同的随机种子会得到完全不同的聚类结果。另外，由于微博属于短文本数据，同一个词在不同短文中出现的概率较小[9]，所以采用传统的 K-Means 聚类算法在度量相似度时往往不太理想。本文主要解决确定初始中心和相似度度量问题，用于分析微博数据得到热点信息，快速了解微博动态。主要解决方案是采用主动学习策略，要求用户选择潜在的有趣数据作为标签，而这种高效的活性种子选择算法依赖于能覆盖整个数据集的 Min-Max 方法(Min-Max approach, MMA)，之后将这些初始中心点赋上权重值参与 K-Means 聚类。

1 基于主动学习的聚类算法

1.1 主动学习策略

主动学习[10]作为一种新的机器学习方法，其目标是发现训练集中高信息量的样本。这里使用 Min-Max 方法来选择一组相互远离的点，确定一套合适的种子，这种基于种子的 K-Means 算法在文献[11]中提过。文献[12]中展示了 Min-Max 方法，就是从数据集 X 构建标签集 Y，Y 彼此远离并且能很好地覆盖数据集。MMA 的思路如下：首先从数据集 X 中随机选择一个起点 y_1，然后从 X 中选择其他点进入 Y，必须满足它们到已经包含在 Y 中的点的最小距离最大。因此，当 t 个点已经在 Y 中，要从 X 中选择第 $t+1$ 个点时，过程为

$$y_{t+1}=\arg\max_{x\in X}(\min_{y\in Y}d(x,y)) \tag{1}$$

式中：$d(\cdot,\cdot)$定义了空间中对象间的距离(如欧几里得距离等)。主动学习中 MMA 的基本思想就是选择距离被选点最远的点，即在每个迭代中，根据前面用户的答案选择标签不确定性最大的点[13]。图 1 将 MMA 应用于本文数据，可看到 MMA 能覆盖整个数据集，所以可用于 K-Means 中初始化聚类中心[14]。图中“*”表示挑选出的 7 个询问点。因此，用 MMA 编写主动学习系统，即由 MMA 提供的点作为候选点向用户提出询问，如果满足

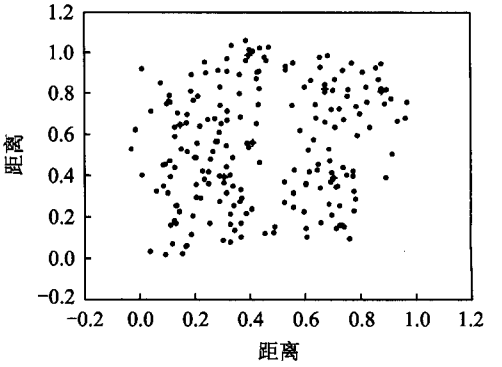

图 1 将 MMA 用于含有 5 类的数据集

Fig. 1 MMA application to the data set containing 5 clusters

用户的条件,就可以将它列为种子点,为后面的 K-Means 聚类做准备。主动选种过程表示为 1 个循环,在每次迭代中,算法从候选集 Candidate_Set(根据 MMA 得到)中选择一个点 u,并询问用户它可否作为种子点。当种子选择完成或候选集中所有的点都检索过(Candidate_Set 为空)时循环停止。算法步骤见流程图 2。

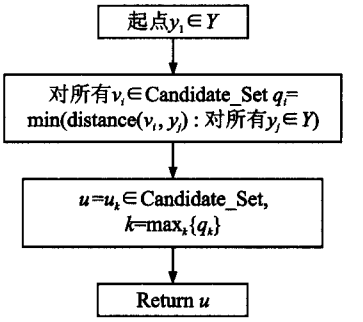

图 2　Min_Max_Approach (Y, Candidate_Set) 算法

Fig. 2　Min_Max_Approach (Y, Candidate_Set) algorithm

1.2　方法扩展

虽然 MMA 能很好地覆盖数据集且使 K-means 算法收敛更快[15],但它不能保证所选点都靠近每组中心,所以在应用 MMA 之前,通过最近邻居图(K-nearest neighbor graph, K-NNG)方法初始化数据集,使数据集中在中心附近的高密集区,可以采用文献[14]中局部密度评分(Local density score,LDS)来估计数据集中每一个点的密集程度。K-NNG 被定义为一个加权无向图,每个顶点代表 1 个数据点,最多拥有 k 个到最近邻居的边。在一对点 x_i 和 x_j 间创建一个边当且仅当 x_i 和 x_j 在最近 k 个邻居集中有彼此。x_i 和 x_j 之间边的权重(相似度)$\omega(x_i, x_j)$ 定义为两点共有的最近邻居的数量,则

$$\omega(x_i, x_j) = |\ \mathrm{NN}(x_i) \cap \mathrm{NN}(x_j)\ | \tag{2}$$

式中:NN(·)为某点的最近 k 邻居点集。一个点的 LDS 定义为

$$\mathrm{LDS}(x_i) = \frac{\sum_{q \in \mathrm{NN}(x_i)} \omega(x_i, q)}{k} \tag{3}$$

一个点的 LDS 就是距离它所有最近邻居的平均距离。LDS 值越大表示 x_i 和它的邻居的关联越大,x_i 属于密集区;LDS 值越小意味着 x_i 属于稀疏区域或集群之间的过渡区。在算法中,只对在密集地区的数据感兴趣,即被选点的 LDS 要高于阈值 ε,候选集根据式(4)选择,则

$$\mathrm{Candidate\_Set} = \{p \in X : \mathrm{LDS}(p) \geqslant \varepsilon\} \tag{4}$$

图 3 表明将 K-NNG 用到图 1 数据上得到的候选集 Candidate_Set,发现候选点分布在密集区靠近集群中心。整个算法使用了两个参数:最近邻居数 k 和阈值 ε。大量实验显示 ε 可设置在区间$[\frac{k}{2}-2, \frac{k}{2}+2]$之间,而 k 不能适用所有数据集,它依赖于数据集的大小和结构。图中用"+"标记的点表示分布在密集区域的候选点 Candidate_Set,其中 $k=34$,$\varepsilon=16$。

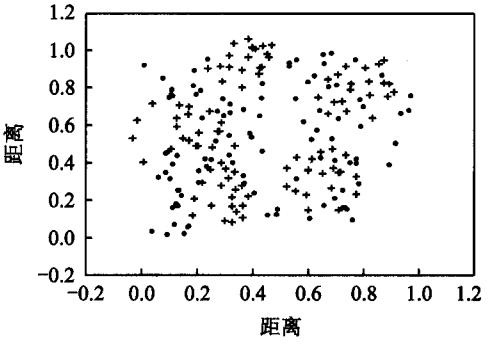

图 3　将 K-NNG 用于图 1 中有 5 个聚类中心的数据集

Fig. 3　K-NNG application to the data set of 5 clusters in Fig. 1

1.3　K-Means 算法改进

通过 MMA 方法确定初始聚类中心以后,在以后的聚类迭代中可能会产生偏移,造成初始点作用弱化的问题。为解决这个问题,本文方案是对 K-Means 每次重新计算聚类中心这一步进行改进,就是每次迭代中为初始聚类中心加上权重值。这里权重值选为已计算过的这一点的 LDS 与最近邻居数 k 的乘积,那么这一类的聚类中心就可理解为它的重心[16],算法流程图见图 4。

2 实验结果与分析

首先基于微博第三方应用程序接口获取某段时间内的 3 000 条微博，这些微博各有侧重，包括娱乐、淘宝、新闻、星座和语录 5 个大家感兴趣的大方面，对这些微博进行聚类分析，得出 5 类热门点。本文实验所采用的软件平台有盘古分词中文分词系列组件、SQL Server 2005 及 Visual Studio 2012，将这些微博数据进行简单的预处理，提取出每条微博有用信息、删掉一些表情符和火星文等，再使用盘古分词利用字典匹配方法，过滤掉英文、标点符号及停用词，同时将微博进行分词处理，如表 1 所示。再将分好词的微博录入 SQL 数据库，然后采用“.net”搭建数据挖掘平台，通过 C# 语言设计聚类算法，进行 3 组对比试验。从表 1 可以看到，经过分词分析：(1)微博中出现的所有标点符号已经被过滤。(2)大部分停用词，比如“有”、“的”等已经被过滤。(3)微博中出现的所有英文以及网址英文均已经被过滤。由于需要计算文本相似度，优先计算词频，表 2 列出了一部分热词及频率。

经过筛选，选择了 1 000 个热词作为关键词，设定为文本特征向量 $\boldsymbol{V}_0$，表示出每条微博的 0-1 特征向量，然后通过主动学习算法从候选集中挑选点来征询用户，直至最后选出感兴趣的种子(见表 3，序号是每条数据的编号)。通过 K-Means 聚类，将这 3 000 条热门微博分成 5 类，可以合理调节阈值，得到每类核心词观测分类情况见表 4。

为了验证算法实用性，按照选择种子的顺序确定 5 个标签：娱乐、淘宝、新闻、星座和语录，人工将这 3 000 条微博标上标签，将算法得到的标签结果与人为标定的结果比对，计算出准确率。对这 3 种方法做了对比试验，不采用主动学习时直接 K-Means 聚类准确率在 40%～50%，主动学习 K-Means 算法以及主动学习改进 K-Means 在选取相同种子点情况下对比实验结果见表 5 和图 5。实验结果表明，主动学习结合 K-Means 准确率在 50%～60%，结合改进 K-Means 准确率在 70%～80%，即主动挑选种子点能提高聚类算法准确性，能解决 K-Means 局部最优的问题。且结合改进 K-Means 后能解决种子漂移问题，更有利于解决实际问题。这种通过主动学习算法先确定种子，然后再通过 K-Means 算法聚类，并且改进 K-Means 重新选取中心的方法，对结果有明显的改善。

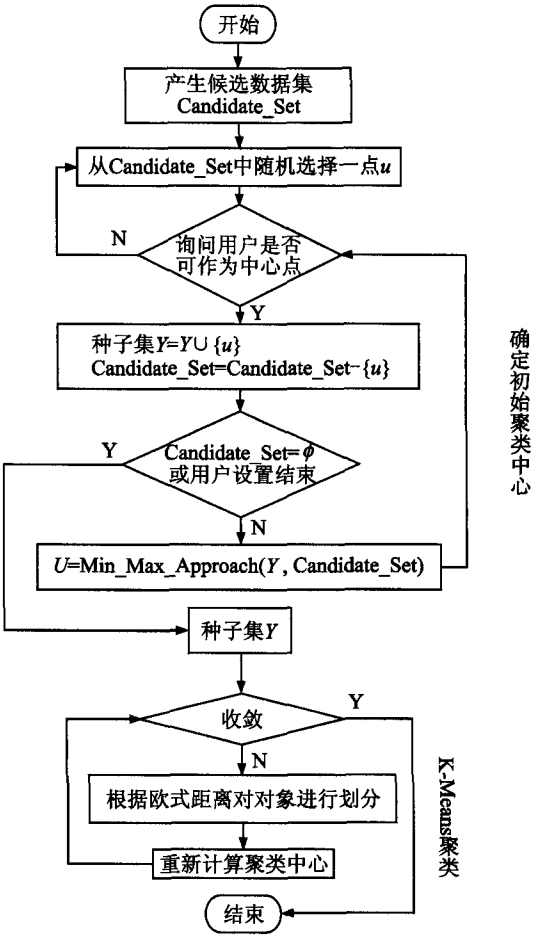

图 4 基于主动学习策略的 K-Means 算法流程图

Fig. 4 Algorithm flow chart by K-Means based on active learning

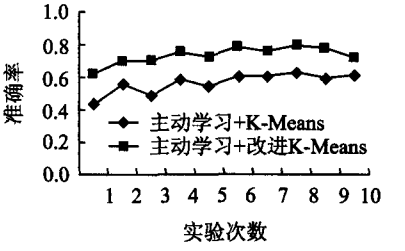

图 5 主动学习结合 K-Means 与主动学习结合改进 K-Means 准确率对比实验

Fig. 5 Comparative experiments between active learning with K-Means and with improved K-Means

表 1 微博分词结果示例

Tab. 1 Examples of segmentation results of micro blogs

微 博	分词结果
#中国好岳父# 马伊琍爸爸追问周刊 放过我们孩子放过文章马伊琍,人家父母都不介意,你们瞎搅和什么	中国 好 岳父 马伊琍 爸爸 追问 周刊 放过 我们 孩子 放过 文章 马伊琍 人家 父母不介意 你们 瞎 搅合
—句歌词证明你听过张国荣#我在:http://htjfk.html	— 句 歌词 证明 你 听 过 张国荣

表 2 热词及其使用频率

Tab. 2 Hot words and their frequency

分享	301	青春	97	平安	75	白羊座	142	来自	84	头发	64
链接	288	瘦	97	脱毛	73	连衣裙	118	麦当劳	78	白羊座	64
新款	280	鞋	96	短裤	73	时尚	117	心理	78	毛孔	64
韩版	267	淘宝	96	双子座	73	语录	113	凉鞋	78	小米	64
销量	261	正品	96	新闻	70	夏天	112	视频	77	张宏民	64
修身	201	签到	94	爸爸	69	夏装	112	星座	77	学校	64
手机	180	代购	93	消息	69	腿	98	健康	76	高考	63
纯棉	173	售出	93	腾讯	66	欧美	97	新品	76	李瑞英	63
休闲	170	明星	89	男人	66	长裙	97	面膜	76	青春	63
网友	163	肌肤	89	减肥	64	瘦	97	女人	75	黄海波	62

表 3 通过 Min-Max 用户选择的种子

Tab. 3 Seed selected by user with Min-Max

序号	微博
88	腾讯 新闻 时至今日 咎由自取 任何人 无关 不易 干 珍惜
135	月 销量 男装 春装 新款 男士 牛仔裤 男 韩 版 潮 码 牛仔 裤子
411	小米 赞 米粉 首发 百万 手机 预约 免费 送 小米 3 红 Note 配件
966	十二 星座 今年 相亲 成功率 白羊座 天秤座 成功率 金牛座 狮子座 双子座 水瓶座 巨蟹座 双鱼座 处女座
688	语录 理想 丰满 现实 骨感 时刻 记得 感恩 人生 路上 帮助 人生 低处 好处 方向 努力 向上 丑 结婚 美 单身 依靠 最大 依赖 叫醒 闹钟 梦想 生活 玩 心跳

表 4 主动学习 K-Means 聚类结果

Tab. 4 K-Means clustering result by active learning

类别	聚类结果核心词
1	家 第三者 文章 珍惜 不易 陈朝华 帅 教授 马 祝福 举报 媒体 道歉 变形计 嚣张 人品 原则 马云 黄海波 明星 爸爸 来自 金秀贤
2	代购 正品 瘦 春装 长款 秘籍 腿 搭配 宽松 新品 运动 牛仔 男 欧美 背带 个性 单鞋 宝贝 衬衫 铅笔 外套 短裤 赞 销量 韩 潮 地址 时尚
3	小米 米粉 配件 预热 手机 推荐 业务 像素 全球 科技
4	白羊座 处女座 星座 五行 网友 运势 幸运 速配 四月 性格
5	语录 心灵 希望 坚强 思考 男人 行走 女人 岁 婚姻 爱情 经济 发表 共享 值得 遗憾 文艺 身边 精彩 直觉 喜欢

表 5 主动学习结合 K-Means 与主动学习结合改进 K-Means 准确率对比实验

Tab. 5 Comparative experiments between active learning with K-Means and active learning with improved K-Means

方法	1	2	3	4	5	6	7	8	9	10
主动学习结合 K-Means	0.431	0.555	0.489	0.589	0.535	0.602	0.598	0.621	0.594	0.615
主动学习结合改进 K-Means	0.612	0.701	0.698	0.756	0.725	0.789	0.758	0.801	0.775	0.723

3 结束语

本文以 K-Means 聚类算法为基础，针对在确定聚类初始中心时随机选取种子，可能会对聚类结果产生较大影响，不同的随机种子会得到完全不同的聚类结果问题，采取了主动学习的策略。另外，具体应用于微博聚类时，因为微博属于短文本数据，同一个词在不同短文中出现的概率较小，所以为中心点设置了权重值。本文主要解决确定初始中心和相似度度量两个问题，用于分析微博数据得到热点信息，快速了解微博动态。K-Means 算法采用合理的主动学习策略，通过征询用户确定初始中心，提高了聚类的速度和准确性，实验结果证明此算法有效可行，但也面临着两个问题：(1)通过主动学习挑选种子点询问用户，虽然大大缩小了挑选范围，提高了精确度，但是最后一步人为确定，造成不可避免的主观性。(2)如何确定聚类数目，使得算法更加客观合理，这些还有待研究。

参考文献：

[1] 李涛. 数据挖掘的应用与实践：大数据时代的案例分析[M]. 厦门：厦门大学出版社，2013：95-120.
Li Tao. Application and practice of data mining: Analysis of the age of big data case[M]. Xiamen: Xiamen University Publishing House, 2013:95-120.

[2] 王晶，朱珂，汪斌强. 基于信息数据分析的微博研究综述[J]. 计算机应用，2012，32(7)：2027-2029，2037.
Wang Jing, ZhuKe, Wang Binqiang. Survey on microblog research based on information data analysis[J]. Journal of Computer Application, 2012, 32(7):2027-2029,2037.

[3] 张剑锋，夏云庆，姚建民. 微博文本处理研究综述[J]. 中文信息学报，2012，26(4)：21-27，42.
Zhang Jianfeng, Xia Yunqing, Yao Jianmin. A review towards microtext processing[J]. Journal of Chinese Information Processing, 2012,26(4):21-27,42.

[4] Turney P, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transaction on Information Systems, 2003,21(4):315-346.

[5] Locke B, Martin J. Named entity recognition: Adapting to microblogging[D]. Colorado: University of Colorado, 2009.

[6] Shen Y, Tian C, Li S, et al. The grand information flows in micro-blog[J]. Journal of Information & Computational Science, 2009,6(2):683-690.

[7] Han J, Kamber M. Data mining, southeast asia edition: Concepts and techniques[M]. San Francisco: Morgan Kaufmann, 2006:55-99.

[8] 扬善林，李永森，胡笑旋，等. k-means 算法中的 k 值优化问题研究[J]. 系统工程理论与实践，2006，26(2)：97-102.
Yang Shanlin, Li Yongsen, Hu Xiaoxuan, et al. Optimization study on k value of K-means algorithm[J]. Systems Engineering-Theory & Practice, 2006,26(2):97-102.

[9] Liu Zitao. Short text feature selection for micro-blog mining[C]// Proceeding of International Conference on Computational Intelligence and Software Engineering. Wuhan: IEEE, 2010:1-4.

[10] 刘康，钱旭，王自强. 主动学习算法综述[J]. 计算机工程与应用，2012，48(34)：1-4.

Liu Kang, Qian Xu, Wang Ziqiang. Survey on active learning algorithms[J]. Computer Engineering & Application, 2012,48(34):1-4.

[11] Basu S, Banerjee A, Mooney R J. Semi-supervised clustering by seeding[C]// Proceeding of the Nineteenth International Conference on Machine Learning. Morgan Kacfmann, 2002:27-34.

[12] Mallapragada P K, Jin R, Jain A K. Active query selection for semi-supervised clustering: Pattern recognition[C]// 2008. ICPR, 2008. 19th International Conference on IEEE. [S. l.]: IEEE, 2008:1-4.

[13] Hasan M A, Chaoji V, Salem S, et al. Robust partitional clustering by outlier and density insensitive seeding[J]. Pattern Recognition Letters, 2009,30(11):994-1002.

[14] Le D D, Satoh S. Unsupervised face annotation by mining the Web[C]// Data Mining Eighth IEEE International Conference on IEEE. [S. l.]: IEEE, 2008:383-392.

[15] Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying density-based local outliers[J]. ACM Sigmod Record, 2000, 29(2):93-104.

[16] Du Q, Faber V, Gunzburger M. Centroidal voronoi tessellations: Applications and algorithms[J]. SIAM Review, 1999,41(4):637-676.

作者简介:

朱丽(1990-),女,硕士研究生,研究方向:数据挖掘,E-mail:shang8jie @ 126.com。

陆建峰(1969-),男,教授,博士生导师,研究方向:数据挖掘,图像处理,智能机器人。