基于改进 K-means 算法的微平台舆情分析研究

——以 UK-means 聚类算法为例

赵立坤　吴东领　韩灿灿
（唐山职业技术学院，河北 唐山 063000）

摘　要　随着互联网的高速发展，通过微平台来表达自己的想法、情绪和态度的人也越来越多，对事件的发展存在积极和消极、正面和负面信息。微平台的开放性和隐蔽性已影响到人们意识形态的发展。因此，对舆情信息的有效、科学的分析、识别、追踪和掌控，对促进国家健康发展，维护社会稳定具有十分重要的意义。传统的 K-Means 聚类算法在进行挖掘数据时计算复杂度较高，凝聚过程不可逆，仅保证了局部最优的收敛性。通过改进的算法 K-means 聚类算法可以提高网络舆情信息聚类文本结果的识别率和数据有效性，为本领域的研究提供参考与借鉴价值，对推动精神文明建设实现高质量发展有着较为重要的使用价值。

关键词　互联网　隐蔽性　舆情　聚类算法　识别率
中图分类号：TP3　　**文献标识码**：A　　**文章编号**：1007-0745(2022)12-0109-03

在国内，网络舆情规范的法律体制相对健全，文本聚类舆情监控研究有不少，比如：北大方正技术研究院推出的方正智思舆情预警辅助决策支持系统[1]，该系统有效地解决了地方政府部门以传统的人工方式进行舆情监测的难题，但在音频、视频等多媒体信息方面还不能对不确定性数据进行挖掘，挖掘的识别率和数据效率性较低。

在国外，许多西方国家已制定了与互联网舆情相关的法律规章。

美国 TDT（Topic Detection and Tracking）系统是国外最有名的与互联网热点舆情发现与监控有关的系统，初衷只是为了研究出一些能够发现和跟踪来自数据流中重要信息和内容的算法[2]。

目前，国内外舆情分析管理方面虽然取得了较好的研究成果，Hamdan 与 Govaert 通过运用 EM 算法解决不确定性数据聚类的混合密度问题。然而，这个模型却不能任意地应用于其他聚类算法。

K-means算法是一种最经典、广泛的划分聚类算法，经常被用于网络舆情的聚类中分析中，因检测、识别不精确、抽样误差、过时数据来源等条件因素，舆情数据往往挖掘不足，导致部分舆情数据遗漏。假设实际位置是有效的，仅仅依靠记录的数据值，很多的目标可能被置于错误的数据集群中，从过时数据值中得到的数据集群有明显差异。

因此，本文提出一种基于 UK-means 聚类算法对传统的初始聚类中心选择方法进行改进，通过不确定性因素与数据挖掘相结合的算法，用于微平台的聚类中，以期能更快、更准确地对近期微平台数据进行聚类，实现热点话题识别与追踪。

1 不确定数据的分类

如图 1 所示，提出一种分类法来区分出硬聚类和模糊聚类的两种数据聚类类型。硬聚类旨在通过考虑预期的数据来提高聚类的准确性和有效性。模糊聚类表示每个数据项被赋予分配给数据簇的任意成员的概率，聚类的结果为一个“模糊”表格。

传统算法未考虑数据不确定性而导致部分数据挖掘遗漏。在数据分类和数据聚集中，通过改进 K-means 算法对聚类质心、两个目标的距离或目标与质心的距离等重要度量作重新定义和进行更深的研究[3]。

2 改进的 K-means 聚类不确定性数据

为了在聚类过程中提取数据不确定性，我们提出一种实现最小化平方误差总和的 E(SSE) 目标算法。一

★基金项目：2022 年度唐山市社会科学界联合会立项课题，课题编号：TSSKL2022-138。

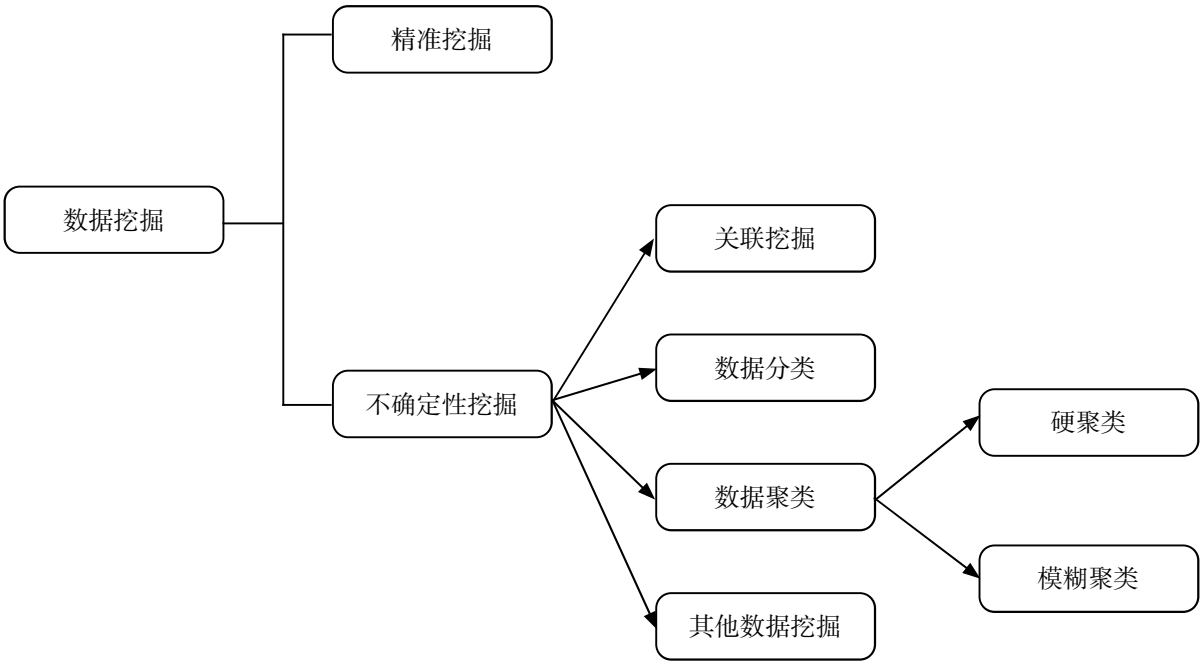

图 1 不确定性数据挖掘的一种分类

个数据对象 x_i 由一个带有不确定性概率密度 $f(x_i)$ 的不确定性区域决定。假设给定一组数据群集，期望平方误差总和计算如下：

$$E\left(\sum_{j=1}^{K}\sum_{i\in C_j}\left|c_j-x_i\right|^2\right)$$
$$=\sum_{j=1}^{K}\sum_{i\in C_j}E\left(\left|c_j-x_i\right|^2\right) \quad (1)$$
$$=\sum_{j=1}^{K}\sum_{i\in C_j}\left|c_j-x_i\right|^2 f(x_i)dx_i$$

数据集平均值如公式：

$$c_j=E\left(\frac{1}{|C_j|}\sum_{i\in C_j}x_i\right)$$
$$=\frac{1}{|C_j|}\sum_{i\in C_j}E(x_i) \quad (2)$$
$$=\frac{1}{|C_j|}\sum_{i\in C_j}\int x_i f(x_i)dx_i$$

由此，我们将提出一种 UK-means 聚类算法，来实现不确定性数据聚类。

1.Assign initial values for cluster means c1 to cK

2.repeat

3.for i = 1 to n do

4.Assign each data point xito cluster Cj where E(||cj-xi||)is the minimum.

5.end for

6.for j = 1 to K do

7.Recalculate cluster mean cj of cluster Cj

8.end for

9.until convergence

10.return C

通过 UK-means 基于数据不确定性模型计算预期的距离和数据集质心，收敛性可按照不同的标准来定义。如果收敛性依赖于下平方误差，公式（1）中 E(SSE) 替代 SSE。在第 4 步中采用代数方法来确定 $E(||c_j-x_i||)$，采用数值积分法确定线，圆等几何图形不确定性区域和不确定性概率密度。鉴于此，获得的 E(||cj-xi||2) 用来替代 $E(||c_j-x_i||)$。

3 实验

3.1 线性移动不确定性数据聚类

UK-means 算法适用于任意一个不确定性区域和概率密度函数。为了证明方法的可行性，我们假设在一个质心 C=(z,q) 和一个数据对象 x 被指定在一个线性不确定的均匀分布的区域中。线性不确定性线段的终结点为（a,b）和（c,d），则参数 δ 表示的线性方程式为 (a+(c-a)t,b+(d-b)t)，其中 t 取值范围属于 [0,1]。$f(t)$ 表示不确定性概率密度函数。

不确定性线段的距离公式为：

$$D=\sqrt{(c-a)^2-(d-b)^2} \quad (3)$$

表 1 实验结果

D	2.5	5	7.5	10	20	50
ARI(UK-means)	0.73	0.69	0.65	0.63	0.51	0.31
ARI(K-means)	0.70	0.63	0.57	0.52	0.35	0.12
改进	0.03	0.06	0.08	0.11	0.16	0.19
改进百分比	4.8%	10 %	13.8%	20.8%	44.3%	155.8%

由此，可以得到：

$$E\left(\|c-x\|^2\right)=\int_0^1 f(t)\left(D^2t^2+Bt+C\right)dt \quad (4)$$

其中 B=2[(c-a)(a-z)+(d-b)(b-q)]

$C=(z-a)^2+(q-b)^2$

函数 f(t) 是均匀分布时，且 f(t)=1 时，计算公式如下：

$$E(\text{线性不确定性与质心的偏差}^2)=\frac{D^2}{3}+\frac{B}{2}+C \quad (5)$$

公式（4）、（5）计算为均匀分布的线性移动不确定性的平方距离。当概率密度函数不是均匀分布时（如，高斯分布），采样技术用来估计取值 $E(\|c_j-x_i\|)$。

3.2 UK-means 算法的评估实验

为了评估 UK-means 算法的可行性，我们采用 100×100 的二维空间所组成的一组随机数据点作为记录。对于每个数据点根据单向线性不确定性模型为其随机产生不确定性。根据记录和不确定性模拟记录中的原始位置的偏移来表示目标的真实位置。对于每个数据点位置记录在案，目标可能的移动距离由随机产生一个数据来决定。计算和比较以下数据集的聚类输出结果：

（1）记录（传统 K-means）

（2）记录 + 不确定性（改进 UK-means）

（3）真实值（传统 K-means）

为核实 UK-means 算法产生的数据群集接近真实数据中数据群集，采用调整相似度的兰德指数（ARI）进行比较聚类结果 [4]，计算两个数据群集之间的相似度来对聚类结果进行评估。ARI 取值范围为 [-1,1]，值越大意味着聚类结果与真实情况越近似。

通过（2）与（3）数据群集间的 ARI 指数和（1）与（3）数据群集间的 ARI 指数比较，在不同的参数组合下，允许 K-means 算法（（1）和（3））和 UK-means 算法（2）在一直运行至迭代次数达到 10000 次或群集中的所有目标在两次连续迭代中没有发生任何变化时结束，n=1000 和 K=20 时，从表 1 可以看出 D 值的不同实验结果。

研究表明：当不确定性程度增加时，UK-means 算法改进度就越高。当群集的个数非常小时，目标的个数和群集的个数对 UK-means 算法的作用基本无影响。从表 1 记录数据中可以看到 UK-means 算法中兰德指数（ARI）的调整近似度始终高于传统的 K-means 算法。因此，UK-means 算法得到的数据群集更接近于从真实世界的数据群集。

4 结语

传统数据挖掘算法无法挖掘固有的不确定性，产生的挖掘结果与真实世界的数据不相符。在本论文中，提出了在不确定性数据挖掘领域研究的一个分类方法，提高网络舆情信息聚类结果的识别率、有效性，实现热点话题识别与追踪，从而准确高效地管理互联网信息 [5]，防患于未然，对推动精神文明建设实现高质量发展有着较为重要的使用价值和应用价值。

参考文献：

[1] 衣波 . 网络舆情信息的话题发现和追踪技术的研究与应用 [D]. 广州 : 广东工业大学 ,2013.

[2] 陈力铭 , 叶朱荪 , 张峰 , 等 . 一种数据聚类方法及装置 CN201810723419.8[P]. 深圳软通动力科技有限公司 :INVENTION_GRANT,2018-07-02.

[3] 卢修配 , 齐向伟 , 艾斯卡尔 . 维吾尔文网络舆情研究现状及几个关键问题 [J]. 新疆师范大学学报 (自然科学版), 2012,31(02):86-88.

[4] 同 [2].

[5] 刘德鹏 . 互联网舆情监控分析系统的研究与实现 [D]. 成都 : 电子科技大学 ,2011.