引用格式:刘荣凯,孙忠林. PCA-KDKM 算法及其在微博舆情中的应用[J]. 山东科技大学学报(自然科学版),2018,37(6):84-92.

LIU Rongkai, SUN Zhonglin. PCA-KDKM algorithm and its application in Weibo public opinion[J]. Journal of Shandong University of Science and Technology (Natural Science), 2018, 37(6):84-92.

PCA-KDKM 算法及其在微博舆情中的应用

刘荣凯,孙忠林

(山东科技大学 计算机科学与工程学院,山东 青岛 266590)

摘 要:针对 K-means 算法因随机选取聚类中心而易造成聚类结果不稳定的问题,提出 PCA-KDKM 算法。该算法使用主成分分析法对数据集的属性降维,提取主属性;利用 k′dist 曲线自动获取 k 值;计算平缓曲线上所含数据对象的均值并选取其中一值,作为首个初始聚类中心;利用基于密度和最大最小距离的算法思想进行聚类;结合类间距离和类内聚类提出聚类质量评价函数。将该算法与 K-means、KNE-KM、QMC-KM、CFSFDP-KM 在 UCI 数据集上进行聚类比较,结果表明该算法聚类结果稳定,聚类准确率高。将 PCA-KDKM 算法应用在微博舆情分析中,抓取不同类别的数万条数据进行聚类分析。实验结果表明,PCA-KDKM 算法在微博舆情分析中有更高的准确性和稳定性,有利于及时发现热点舆情。

关键词:K-means 算法;k′dist 曲线图;聚类;质量评价函数;准确率;PCA-KDKM 算法;微博舆情

中图分类号:TP311 **文献标志码**:A **文章编号**:1672-3767(2018)06-0084-09

DOI:10.16452/j.cnki.sdkjzk.2018.06.010

PCA-KDKM Algorithm and Its Application in Weibo Public Opinion

LIU Rongkai, SUN Zhonglin

(College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)

Abstract: To solve the problem of unstable clustering results of the K-means algorithm due to its random selection of cluster centers, this paper proposed the PCA-KDKM algorithm. Principal component analysis was used to reduce the dimension of the data set and extract the main attribute. The k′dist curve was used to obtain the k value automatically. The average value of the data objects contained in the flat curve was calculated and one was selected as the first initial clustering center. Clustering was made based on density and minimum distance algorithm ideas. clustering quality evaluation functions were proposed by combining the distance between classes and intra-class clustering. The clustering results of the algorithm and K-means, KNE-KM, QMC-KM, CFSFDP-KM algorithm on UCI data sets were compared and the results show that the proposed algorithm has stable clustering results and high clustering accuracy. The PCA-KDKM algorithm was then applied to the Weibo public opinion analysis to capture tens of thousands of data in different categories for cluster analysis. The experimental results show that the PCA-KDKM algorithm has higher accuracy and stability in the Weibo public opinion analysis, which is conducive to timely detection of hot public opinion.

收稿日期:2018-04-08

基金项目:国家自然科学基金项目(61702305);山东省高等学校科技计划项目(J16LN08)

作者简介:刘荣凯(1991—),男,山东泰安人,硕士研究生,主要从事计算机软件与理论研究. E-mail:1033128066@qq.com

孙忠林(1962—),男,山东蓬莱人,教授,主要从事模式识别、数据库系统、系统集成及安全工程方面的系统及预测研究,本文通信作者. E-mail:zhonglinsun@163.com

Key words: K-means algorithm; k′dist graph; clustering; quality evaluation function; accuracy; PCA-KDKM algorithm; Weibo public opinion

聚类分析是一种无监督的学习[1]。聚类是根据数据对象集合中各数据对象的特征,将具有相似特征的数据对象划分为一个类别,使得同类型相似性尽量高,不同类型相似性尽量低[2]。

K-means[3]算法具有理论可靠和计算快速的优点,在实践中被广泛应用[4-6]。随着研究的深入,算法的一些缺点逐渐暴露出来。存在的主要问题有:①在数据对象分类中,将平方误差和作为聚类质量好坏的评价指标,运算量大,效率低;②算法在运行时需要事先确定初始中心的个数 k;③对初始中心的选择敏感,由于算法随机的选择 k 个初始中心,造成聚类的结果不稳定。

许多学者从算法初始 k 个聚类中心的选择、数据对象相似度的计算、聚类质量评价函数这三方面对算法进行改进,使算法聚类效果得到了一定改善。郑丹等[5]基于密度选取初始聚类中心,选取各主要密度水平曲线上的第一个点作为 k 个初始聚类中心;冯波等[6]利用对象的距离矩阵来生成最小生成树,然后将树修剪成 k 个分支,将 k 个分支中所包含点的均值作为初始中心进行聚类;石文峰等[7]提出了一种用个体轮廓系数作为聚类有效性评估参数的改进算法;周炜奔等[8]提出基于数据密度分布特征,利用均衡函数自动生成最优簇数 k 值的算法;王宏杰等[9]结合初始中心优化和特征加权,改进 K-Means 聚类算法;陈小雪等[10]提出了一种基于萤火虫优化的加权 K-means 算法;王赛芳等[11]提出了基于点密度的初始聚类中心选择方法;李敏等[12]提出密度峰值优化初始中心的 K-means 算法;庄瑞格等[13]提出基于拟蒙特卡洛的 K-means 聚类算法;熊开玲等[14]提出基于核密度估计的 K-means 聚类优化算法;张雪凤等[15]通过调整 K-means 算法运行中数据对象被重新分配时的分配策略,在再分类过程中,将数据对象分配到最小加权距离中心点所在的簇中,以提高聚类质量;赵将等[16]提出了一种改进 K-means 聚类的推荐方法 IKC(Improved K-means Clustering Recommendation Method)。以上研究仍存在聚类 k 值需指定、聚类质量评价函数不合理等问题,导致聚类结果不稳定,准确率低。

针对以上问题,本文提出 PCA-KDKM 算法:基于 PCA 进行数据集的降维,提取出主属性以加速聚类。借助 k′dist[17]曲线图来确定聚类簇数 k 值,取第 $i(i \leqslant k)$段平缓曲线上所包含点的均值作为第一初始聚类中心。用基于密度和最小距离的算法思想,在剩下的数据对象中选取 $k-1$ 个初始聚类中心。再利用传统的 K-means 算法,把集合中剩下的数据对象分到最近的聚类中心所在的簇,计算聚类质量评价函数的值来评价本次的聚类效果。重复 k 次,取聚类评价函数值最大的一组作为最后聚类结果。实验证明:PCA-KDKM 算法在 UCI 数据集上得到的聚类结果比传统 K-means 算法聚类结果准确度高,聚类结果更稳定。

参数 k 在 k′dist 图和 K-means 算法中没有直接联系,为避免混淆,本节以下 K-means 算法中的 k 仍然沿用 k,而将 k′dist 图中 k 改为 k'。

1　$k'dist$ 图

点 p 到离它最近的第 k'(研究表明,$k'>4$ 的 k′dist 曲线变化非常小,几乎与 $k'=4$ 的 k′dist 曲线完全相同)个点的距离为该点 p 的 k′dist 值[18]。在同一簇中更改 k' 值不会导致 k′dist 值的显著变化,该点的 k′dist 半径至少包含 $k'+1$ 个点。k′dist 值按升序排列,然后用曲线连接起所有的点,即可绘出 k′dist 曲线图。

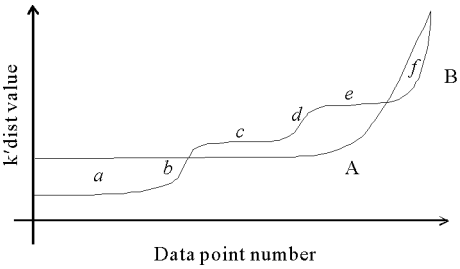

图 1　k′dist 曲线

Fig. 1　k′dist graph

图 1 中两条不同的曲线 A 和 B 分别对应两个不同的数据对象集合的 k′dist 曲线图。其中曲线 A 开始部分比较平缓,说明 A 数据集大部分数据相似性很高。曲线最后部分的 k′dist 值快速增加,表明 A 数据集只有少量点的 k′dist 值之间有很大的差异,这部分通常是边界点或噪声点。曲线 B 包含三部分平缓处 a、c、e,说明这三部分数据分别处于三个主要密度水平上。曲线 B 上

b、d 处的 k′dist 值增加迅速,表明这两部分点是边界点。其中 b 部分的数据连接 a 和 c 两个主要的水平密度,而 d 部分数据点连接 c 和 e 这两个主要密度,f 处 k′dist 迅速增加表明这部分点是噪声点(一般情况下,处于低密度区域的数据对象被称为噪声点[19])。

2 PCA-KDKM 算法

2.1 PCA-KDKM 算法相关定义

假设数据对象集合 U 含有 N 个样本数据,样本数据有 m 个属性,数据对象 i 表示为 $i=(i_1,i_2\cdots i_m)$。

定义 1 点密度。对于数据对象集合 U 中任一对象 i,以 i 为球心,以某一正数 r 为半径的圆球中所包括的数据对象 j 的数量即为数据对象 i 的点密度,记作 $Dens(i)$,

$$Dens(i)=|Dist(i,j)\leqslant r,j\in U|。\tag{1}$$

定义 2 两个 m 维对象 i,j 的欧式距离公式为:

$$d(i,j)=\sqrt{(i_1-j_1)^2+\cdots+(i_m-j_m)^2}。\tag{2}$$

其中:i_1,$i_2\cdots i_m$ 和 j_1,$j_2\cdots j_m$ 分别表示数据对象 i,j 的各维数据;$d(i,j)$ 表示 i 和 j 的距离。

定义 3 高密度对象。在集合 U 中,如果一个对象 i,在 ε 邻域内至少有 $pts_{\min}$ 个对象,则称该对象 i 为高密度对象。

定义 4 聚类质量评价函数[20]

$$S=\frac{1}{N}\sum_{i=1}^{N}S_i,\tag{3}$$

其中

$$S_i=\frac{b(i)-a(i)}{\max[a(i),b(i)]},\tag{4}$$

$$a(i)=\frac{1}{n_c-1}\sum_{i,j\in c_c,i\neq j}d(i,j),\tag{5}$$

$$b(i)=\min_{p,p=}\left[\frac{1}{n_p}\sum_{j\in c_p,i\in c_c}d(i,j)\right]。\tag{6}$$

N 表示集合中数据对象的个数,假设样本 i 被分类到 C 类,$a(i)$ 表示样本 i 和 C 类中其他数据对象之间的平均距离。$b(i)$ 表示样本 i 与其他非 C 类的各个类中所有样本平均距离的最小值。

PCA-KDKM 算法的聚类质量评价函数 S 的值域介于 −1 和 1 之间。结合类内距离与类间距离这两个因素来评估数据对象分类的合理性。S 越接近 1,说明该数据对象的类内平均距离远远小于类间的平均距离,该数据对象分类越合理;S 越接近 −1,说明数据对象类内平均距离远远大于类间的平均距离,分类不合理,应当取消。

2.2 主成分分析算法

主成分分析算法的思想是:当数据对象集合中的每个对象数据有多维属性信息,且数据对象的各维属性之间具有一定的相关性时,首先要对数据对象进行降维,去除无关属性。数据对象降维,具体指使用主成分分析方法对各维数据进行分析,提取并合成数据对象中无相关性的主要属性,实现降维。主要属性能够代表数据对象集中的原有属性且所包含的属性互不相关。使用主成分分析法能够有效降低数据对象集合中的无关属性,大幅提高聚类运算效率及聚类准确率。

标准化变换公式:

$$z_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j},i=1,2,\cdots,p;j=1,2,\cdots,m。\tag{7}$$

其中,$\bar{x}_j=\dfrac{\sum_{i=1}^{p}x_{ij}}{p}$,$s_j=\dfrac{\sum_{i=1}^{p}(x_{ij}-\bar{x}_j)}{p-1}$。$x_{ij}$ 表示矩阵中的每个样本值,z_{ij} 表示标准化矩阵的值,$\bar{x}_j$ 表示列均

值，$\bar{s}_j$ 表示列标准差。

$$\boldsymbol{R} = [r_{ij}]_m \times m = \frac{\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z}}{p-1} \tag{8}$$

其中，$r_{ij} = \dfrac{\sum z_{kj} \cdot z_{kj}}{p-1}, i, j = 1, 2 \cdots m$；$\boldsymbol{R}$ 为相关系数矩阵；$\boldsymbol{x}$ 为随机向量；m 为维数；$\boldsymbol{z}^{\mathrm{T}}$ 表示标准化矩阵 $\boldsymbol{z}$ 的转置。

2.3　PCA-KDKM 算法步骤：

输入：包含 N 个对象的数据对象集合 $U_{P\times m}$，邻域半径以及包含对象的最小数目 $pts_{\min}$。

输出：根据数据对象特征分成的 k 个类别。

算法步骤：

1）$U_{P\times m}$ 按公式(7)、公式(8)计算相关系数矩阵 $\boldsymbol{R}$；

2）根据 $|\boldsymbol{R}-\lambda \boldsymbol{I}_m| = 0$ 计算 m 个特征值 λ_i；

3）由贡献率 $\dfrac{\lambda_i}{(\sum_{i=1}^{m} \lambda_i)}$ 确定 n 个主成分，得到 $\boldsymbol{D}_{p\times n}$；

4）计算 $\boldsymbol{D}_{p\times n}$ 中每个数据对象的 k′dist 值，并按从小到大的顺序排列，绘出 k′dist 曲线图；

5）分析 k′dist 曲线图，观察有几个平缓的曲线，确定聚类数值 k 并求出每段平缓曲线上所包含数据对象的均值，存放在集合 Q 中；

6）利用距离矩阵表示 $\boldsymbol{D}_{p\times n}$ 中两两数据对象之间的欧式距离 $d(i,j)$；

7）计算 $\boldsymbol{D}_{p\times n}$ 中每个对象的 ε 邻域内所包含的对象个数，如果满足最小个数 $pts_{\min}$，则将这个数据对象加入到高密度数据对象集合 H 中；

8）从数据对象集合 Q(利用 k′dist 曲线所求出的每段平缓曲线所包含对象的均值)中选取 Q_1 作为第一个聚类中心 k_1，将 k_1 放入初始聚类中心集合 M 中；

9）计算高密度数据对象集合 H 中所有对象与选取的第一个初始聚类中心 k_1 的距离 d，从高密度数据对象集合 H 中，找出距离初始聚类中心集合 M 中的所有数据对象最远的数据对象 k_2，将 k_2 从高密度集合 H 中删除，并将 k_2 加入到集合 M 中；

10）从高密度数据对象集合 H 中找出距离初始聚类中心集合 M 中的所有数据对象距离最远的数据对象 k_3，从高密度集合中删除 k_3，将 k_3 加入到集合 M 中；

11）从高密度集合 H 中找距离集合 M 距离最远的对象，直到找到 k 个初始中心为止；

12）从集合 M 中的这 k 个中心出发，将其他对象分到离它距离最近的簇，用 K-means 算法进行聚类，计算本次的聚类质量评价函数值 S_i；

13）从集合 Q 中选取第二个均值 Q_2 作为第一个聚类中心 k_1，重复(6)～(12)，直到选取 Q_k 为第一个初始聚类中心结束。

14）比较每次的聚类质量评价函数值 S_i，从中选取聚类质量评价函数值 S_i 最大的一组作为结果。

3　实验结果及分析

3.1　PCA-KDKM 算法与各改进 K-means 算法的对比

实验平台：英特尔 Xeon(至强)E5450@3.00 GHz 四核、4 G 内存、500 G 硬盘，Windows 7 旗舰版系统。

实验描述：实验采用专为机器学习和数据挖掘算法设计的公共数据库 UCI 来进行实验[21]。UCI 中的每个数据对象都有分类编号，因此可以将 PCA-KDKM 算法得出的聚类结果与 UCI 数据库中数据对象的分类标志进行一致性统计，来评价 PCA-KDKM 算法的聚类质量。采用 UCI 数据库中的 Iris、Glass Identification、Wine 和 Hayes-Roth 4 组数据。为了对比 K-means 算法与 PCA-KDKM 算法的准确度，对 4 组数据不作任何修改。

基于先前实践经验，PCA-KDKM 算法进行实验时对 Iris 和 Glass Identification 这两组数据指定领域 ε

为 1,某点 p 的ε范围内含数据对象的最小个数 $pts_{\min}$ 为 50;对于 Wine 指定领域ε为 10 000,某点 p 的ε范围内含数据对象的最小个数 $pts_{\min}$ 为 50;对于 Hayes-Roth 指定领域,ε为 600,某点 p 的ε范围内含数据对象的最小个数 $pts_{\min}$ 为 40。PCA-KDKM 算法得到的初始中心是稳定的,所以对 PCA-KDKM 算法进行 k 次实验即可。K-means 算法进行了 50 次实验,与 PCA-KDKM 算法的实验对比结果如表 1、2 所示。

表 1 PCA-KDKM 算法与 K-means 算法聚类准确率比较

Tab. 1 Comparison of clustering accuracy between PCA-KDKM algorithm and K-means algorithm

算法	数据集	准确率/%		
		最高	最低	平均
K-means	Iris	86.52	78.31	84.30
	Glass	71.68	45.65	62.13
	Wine	69.23	61.54	65.40
	Hayes-Roth	83.73	70.58	77.84
PCA-KDKM	Iris	95.35	90.68	93.55
	Glass	86.40	79.53	86.86
	Wine	78.46	67.36	75.83
	Hayes-Roth	88.73	84.63	86.38

表 2 PCA-KDKM 算法与 K-means 算法聚类质量评价函数值 *S* 比较

Tab. 2 Comparison of clustering quality evaluation function value S between PCA-KDKM algorithm and K-means algorithm

算法	数据集	Smax	Smin	Savg
K-means	Iris	0.345 2	0.246 5	0.303 2
	Glass	0.491 2	0.382 1	0.416 5
	Wine	0.481 2	0.412 6	0.434 6
	Hayes-Roth	0.618 2	0.502 3	0.562 3
PCA-KDKM	Iris	0.818 7	0.7187	0.755 7
	Glass	0.846 5	0.798 5	0.814 5
	Wine	0.764 1	0.613 6	0.734 1
	Hayes-Roth	0.818 2	0.712 3	0.782 4

从表 1、表 2 中可以看出,对于 Iris、Glass、Hayes-Roth 3 个数据集合,PCA-KDKM 算法在聚类准确率和聚类质量评价函数值方面比 K-means 算法的聚类准确率明显提高,聚类准确度为 90%以上;对于 Wine 数据对象集合来说,PCA-KDKM 算法比 K-means 算法的准确率提高不明显,但从聚类质量评价函数值 S 来看,PCA-KDKM 算法比 K-means 算法的聚类效果好。

在 UCI 数据集中选取 4 个数据集 yeast、abalone、magic 和 skin 进行实验。将 PCA-TDKM 算法分别与李敏等[12]的 CFSFDP-KM 算法、庄瑞格等[13]的 QMC-KM 聚类算法和熊开玲等[14]的 KNE-KM 聚类优化算法进行聚类准确率及聚类误差平方和的比较。文献[12-14]的 3 个算法都是针对 K-means 算法的初始 k 个聚类中心进行了优化,都在一定程度上克服了 K-means 聚类算法随机选择 k 个初始聚类中心所造成聚类不稳定的问题。通过实验得到聚类精度的比较结果、误差平方和的比较结果分别如表 3、表 4。其中 HIG 表示最高聚类准确率,LOW 表示最低聚类准确率,AVG 表示平均聚类准确率。

由表 3 可知,基于 PCA 的 PCA-KDKM 算法的聚类效果稳定;对于 yeast、abalone 和 skin 数据集,PCA-KDKM 算法的聚类准确率都达到了 84.35%以上,高于其他三种算法;在 magic 数据集上,PCA-KDKM 算

法聚类准确率为 79.63%,优于其他三种算法。

表 3　4 种算法聚类准确率对比

Tab. 3　Clustering accuracy comparison of 4 algorithms

数据	聚类准确率/%											
	PCA-KDKM			KNE-KM			QMC-KM			CFSFDP-KM		
	HIG	LOW	AVG	HIG	LOW	AVG	HIG	LOW	AVG	HIG	LOW	AVG
yeast	86.56	86.56	86.56	79.93	71.32	73.54	72.14	70.36	71.45	68.54	61.78	64.25
abalone	84.35	84.35	84.35	80.45	76.46	78.93	75.67	73.62	74.34	70.94	67.95	69.34
magic	79.63	79.63	79.63	70.59	64.32	68.45	76.35	69.43	73.94	77.98	70.46	74.58
skin	86.79	86.79	86.79	81.23	73.64	78.42	80.64	78.46	79.34	74.12	70.67	73.45

表 4　4 种算法误差平方和对比

Tab. 4　Squared error sums comparison of 4 algorithms

数据	误差平方和			
	PCA-KDKM	KNE-KM	QMC-KM	CFSFDP-KM
yeast	0.112	0.153	0.156	0.345
abalone	0.192	0.456	0.564	0.753
magic	54.361	69.133	70.694	78.321
skin	70.135	89.321	99.135	86.124

由表 4 可知,PCA-KDKM 算法在 4 个数据集上的误差平方和小于其他 3 种算法在 4 个数据集上的误差平方和,说明 PCA-KDKM 算法的聚类效果比其他三种算法聚类效果好。

3.2　PCA-KMKD 算法在网络微博舆情检测中的应用

采用向量空间模型 VSM(vector space model)将文本内容的处理简化为向量空间中的运算,文档中词语的权重采用 TF-IDF(term frequency-inverse document frequency)方法来计算。利用相同的数据集对 PCA-KDKM 算法及 K-means 算法进行对比试验。

首先抓取一些目前比较热门、关注度高的词汇。本实验在新浪微博上分别为“战争”“上合峰会”“房地产”以及“中美贸易战”等四个关键词的每个词汇均抓取 10 000 条数据,然后进行人工过滤筛选来保证数据的准确性及有效性。从每个词汇分类中选取 400 条以上的微博数据,每条字符长度在 20 以上的微博共筛选出 1 600 条,作为本次微博舆情聚类实验的训练数据集。采用中国科学院研究开发的 ICTCLAS 分词系统,对从新浪微博中提取的数据进行数据镜像分词、词性标注处理,并借助停用词表进行过滤,把停用词从分词结果中过滤掉,然后使用 TF-IDF 方法构造微博数据的向量空间模型(VSM 特征项矩阵),实现对网络微博文本数据的聚类。

3.3　PCA-KDKM 算法在舆情应用中的实验结果分析

采用信息检索领域广泛使用的度量标准 F 度量值,作为微博文本数据聚类结果的评价标准。该方法将查准率(W)和查全率(Z)两个因素结合在一起,使得度量更加客观公正,W 和 Z 分别由以下公式计算得出:

$$W = TP/(TP + FP),$$

$$Z = TP/(TP + FN)。$$

其中,TP 即聚类的正确文档数,为正类被划分为正类的文档数,FP 为负类被划分为正类的文档数,FN 为正类被划分为负类的文档数,$TP + FP$ 表示实际分类的文档数,$TP + FN$ 表示应有的文档数。

为了更客观地对算法的聚类性能进行评价,本研究对 K-means、PCA-KDKM、KNE-KM、QMC-KM、CFSFDP-KM 5 种算法进行实验。由于 PCA-KDKM 算法聚类中心稳定,所以只进行 k(聚类中心的个数 k)次实验,其他 4 种算法进行 30 次实验。表 5 是 K-means、PCA-KDKM、KNE-KM、QMC-KM、CFSFDP-KM 5 种算法 F 平均值实验结果对比。

表 5 5 种算法 F 平均值对比

Tab. 5 Comparison of the average of F 5 algorithm %

微博词汇类	战争	上合峰会	房地产	中美贸易战
K-means	56.8	69.5	74.6	69.7
PCA-KDKM	79.5	82.3	78.5	84.3
KNE-KM	65.4	70.6	67.9	71.4
QMC-KM	66.4	71.3	73.6	68.1
CFSFDP-KM	53.4	80.6	77.8	69.5

由表 5 可知,PCA-KDKM 算法 F 值相对稳定,不像其他 4 种算法那样波动大。此外,PCA-KDKM 算法聚类准确率相比 K-means、KNE-KM、QMC-KM、CFSFDP-KM 4 种算法聚类准确率有明显提高。运用 PCA-KDKM 算法对微博文本中提取的 5 个词汇,共计 40 000 条微博数据进行抓取,然后对数据进行聚类,所得到的聚类结果与时下讨论的热点问题事实数据一致,从每个簇中微博舆情数据的特征项,可以快速获得时下微博舆论关注的热点。以“中美贸易战”为例,最近讨论比较多的是“中兴”“特普朗”“中国芯”等,其中“中兴”热度最高,符合最近微博舆情的基本情况。

下面进行运行时间的比较。PCA-KDKM 算法在初始计算初始聚类中心时,计算量较大,有一定的时间消耗,但从试验运行情况来看,在微博关键词聚类过程中迭代次数减少了,因此总体进行微博聚类的运行时间降低。为验证 PCA-KDKM 算法的时间效率,分别选取具有代表性的 4 次聚类时间,对 K-means、KNE-KM、QMC-KM、CFSFDP-KM 算法以及 PCA-KDKM 算法进行比较。由图 2 可得 PCA-KDKM 算法在运行时间上比 K-means 算法有明显优势,也比 KNE-KM、QMC-KM、CFSFDP-KM 三种算法的运行时间短。

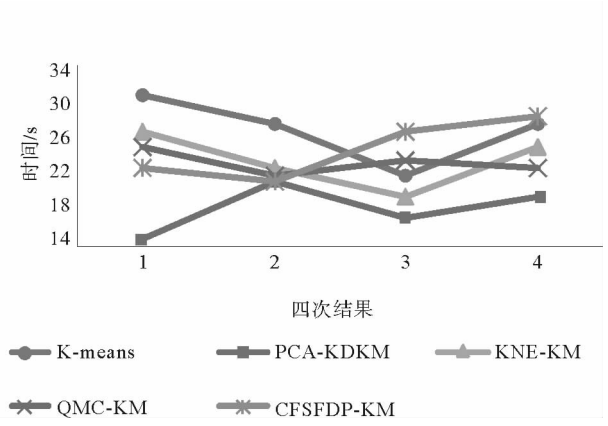

图 2 运行时间对比

Fig. 2 Comparison of running time

综上,PCA-KDKM 算法不仅显著提高了聚类的准确率,而且还提高了运行效率,能够更快更准的发现舆论热点话题。

4 结论

本文提出了 PCA-KDKM 算法:首先利用 PCA 算法进行降维,选取主属性,利用 k′dist 图自动获取 k 值并采用基于最大最小距离算法依次选取 k 个初始聚类中心,消除了当初始聚类中心选取到噪音数据以及孤立点时导致的聚类结果不稳定且容易陷入局部最优解的影响。将本文提出的 PCA-KDKM 算法应用到网络微博高关注度词汇分析中,实验结果表明:PCA-KDKM 算法在聚类的准确率和运行效率方面都比 K-means 及其他算法高,在用于微博高关注度词汇数据进行聚类分析时,可以快速而准确发现当前微博舆论的热点话题,有利于政府机构快速决策,降低社会负面影响。

参考文献:

[1]VISVANATHAN M,ADAGARLA B S,GERALD H L,et al. Cluster validation:An integrative method for cluster analysis [C]// IEEE International Conference on Bioinformatics and Biomedicine Workshop. IEEE,2009:238-242.

[2]HAN J W,KAMBER M,PEI J,等. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2012.

[3]MACQUEEN J B. On convergence of K-means and partitions with minimum average variance[J]. Annals of Mathematical Statistics,1965,36.

[4]周晓彦,安星星,刘文杰,等. 一种基于最小距离的量子 K-means 算法[J]. 小型微型计算机系统,2017,38(5):1059-1062.
ZHOU Xiaoyan,AN Xingxing,LIU Wenjie,et al. Quantum K-means algorithm based on the minimum distance[J]. Journal of Chinese Computer Systems,2017,38(5):1059-1062.

[5]郑丹,王潜平. K-means 初始聚类中心的选择算法[J]. 计算机应用,2012,32(8):2186-2188.
ZHENG Dan,WANG Qianping. Selection algorithm for K-means initial clustering center[J]. Journal of Computer Applications,2012,32(8):2186-2188.

[6]冯波,郝文宁,陈刚,等. K-means 算法初始聚类中心选择的优化[J]. 计算机工程与应用,2013,49(14):182-185.
FENG Bo,HAO Wenning,CHEN Gang,et al. Optimization to K-means initial cluster centers[J]. Computer Engineering and Applications,2013,49(14):182-185.

[7]石文峰,商琳. 一种基于决策粗糙集的模糊 C 均值聚类数的确定方法[J]. 计算机科学,2017,44(9):45-48.
SHI Wenfeng,SHANG Lin. Determining clustering number of FCM algorithm based on DTRS[J]. Computer Science,2017,44(9):45-48.

[8]周炜奔,石跃祥. 基于密度的 K-means 聚类中心选取的优化算法[J]. 计算机应用研究,2012,29(5):1726-1728.
ZHOU Weiben,SHI Yuexiang. Optimization algorithm of K-means clustering center of selection based on density[J]. Application Research of Computers,2012,29(5):1726-1728.

[9]王宏杰,师彦文. 结合初始中心优化和特征加权的 K-Means 聚类算法[J]. 计算机科学,2017,44(增 2):457-459.
WANG Hongjie,SHI Yanwen. K-means clustering algorithm based on initial center optimization and feature weighted[J]. Computer Science,2017,44(S2):457-459.

[10]陈小雪,尉永清,任敏,等. 基于萤火虫优化的加权 K-means 算法[J]. 计算机应用研究,2018,35(2):466-470.
CHEN Xiaoxue,WEI Yongqing,REN Min,et al. Weighted K-means clustering algorithm based on firefly algorithm[J]. Application Research of Computers,2018,35(2):466-470.

[11]王赛芳,戴芳,王万斌,等. 基于初始聚类中心优化的 K-均值算法[J]. 计算机工程与科学,2010,32(10):105-107.
WANG Saifang,DAI Fang,WANG Wanbin,et al. A K-means algorithm based on the optimal initial clustering center[J]. Computer Engineering & Science,2010,32(10):105-107.

[12]李敏,张桂珠. 密度峰值优化初始中心的 K-means 算法[J]. 计算机应用与软件,2017,34(3):212-217.
LI Min,ZHANG Guizhu. K-means Algorithm of optimized initial center by density peaks[J]. Computer Applications & Software,2017,34(3):212-217.

[13]庄瑞格,倪泽邦,刘学艺. 基于拟蒙特卡洛的 K 均值聚类中心初始化方法[J]. 济南大学学报(自然科学版),2017,31(1):35-41.
ZHUANG Ruige,NI Zebang,LIU Xueyi. A novel method for refining the initial points for K-means clustering based on quasi-Monte Carlo method[J]. Journal of University of Jinan (Science and Technology),2017,31(1):35-41.

[14]熊开玲,彭俊杰,杨晓飞,等. 基于核密度估计的 K-means 聚类优化[J]. 计算机技术与发展,2017,27(2):1-5.
XIONG Kailing,PENG Junjie,YANG Xiaofei,et al. K-means clustering optimization based on kernel density estimation [J]. Computer Technology and Development,2017,27(2):1-5.

[15]张雪凤,张桂珍,刘鹏. 基于聚类准则函数的改进 K-means 算法[J]. 计算机工程与应用,2011,47(11):123-127.
ZHANG Xuefeng,ZHANG Guizhen,LIU Peng. Improved K-means algorithm based on clustering criterion function[J]. Computer Engineering and Applications,2011,47(11):123-127.

[16]赵将. 基于改进 K-means 聚类的推荐方法研究[D]. 武汉:华中科技大学,2016.

[17]DUAN L,XU L D,GUO F,et al. A local-density based spatial clustering algorithm with noise[J]. Information Systems,2006,32(7):978-986.

[18]杨正宽. 基于距离的离群挖掘算法研究[D]. 重庆:重庆大学,2011.

[19]ESTER M,KRIEGEL H P,SANDER J,et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI,1996:226-231.

[20]张靖,段富. 优化初始聚类中心的改进 K-means 算法[J]. 计算机工程与设计,2013,34(5):1691-1694.
ZHANG Jing,DUAN Fu. Improved K-means algorithm with meliorated initial centers[J]. Computer Engineering & Design,2013,34(5):1691-1680.

[21]傅德胜,周辰. 基于密度的改进 K 均值算法及实现[J]. 计算机应用,2011,31(2):432-434.
FU Desheng,ZHOU Chen. Improved K-means algorithm and its implementation based on density[J]. Journal of Computer Applications,2011,31(2):432-434.

(责任编辑:傅 游)

(上接第 54 页)

[12]高明美,孙浩,刘喜华. 带干扰和投资的双二项风险模型的破产概率[J]. 统计与决策,2015(22):22-25.
GAO Mingmei,SUN Hao,LIU Xihua. Ruin probability in risk model with double binomial process of diffusion and investment[J]. Statistics & Decision,2015(22):22-25.

[13]盖维丹. 带常利率和相依结构更新风险模型的破产概率[J]. 经济数学,2016,33(2):29-33.
GAI Weidan. Ruin probability in a risk model with constant interest force and dependence structure[J]. Journal of Quantitative Economics,2016,33(2):29-33.

[14]何晓霞,姚春,胡亦钧. 利率为马氏链的离散时间风险模型的破产概率[J]. 应用概率统计,2012,28(3):270-276.
HE Xiaoxia,YAO Chun,HU Yijun. Ruin probabilities for the discrete risk models with Markov chain interest[J]. Chinese Journal of Applied Probability and Statistics,2012,28(3):270-276.

[15]刘家有. 带马氏链利率的离散风险模型的破产概率[J]. 合肥学院学报,2006,16(2):15-18.
LIU Jiayou. Ruin probabilities in a discrete time risk model with a Markov chain interest[J]. Journal of Heifei University (Natural Science),2006,16(2):15-18.

[16]李娜之,刘志平. 带马氏利率的离散时间风险模型的破产概率[J]. 数学理论与应用,2009,29(4):6-9.
LI Nazhi,LIU Qingping. Ruin probabilities for discrete time risk models with Markov interest rate[J]. Mathematical Theory and Applications,2009,29(4):6-9.

[17]牛祥秋. Markov 链利率下再保险模型的破产概率上界[J]. 经济数学,2016,33(3):45-50.
NIU Xiangqiu. Upper bound for the ruin probability under risk model of reinsurance with Markov chain interest rate[J]. Journal of Quantitative Economics,2016,33(3):45-50.

(责任编辑:傅 游)