一种改进的 K - means 算法最佳聚类数确定方法

边 鹏[1,2] 赵 妍[3] 苏玉召[1,2]
[1](中国科学院国家科学图书馆 北京 100190)
[2](中国科学院研究生院 北京 100049)
[3](郑州航空工业管理学院计算机科学与应用系 郑州 450015)

【摘要】对 BWP 方法进行研究,从嵌入式 NSTL 个性化推荐的文本聚类需求入手,分析 BWP 方法的不足,提出一种改进的 K - means 算法最佳聚类数确定方法。对单一样本类的类内距离计算方法进行优化,扩展 BWP 方法适用的聚类数范围,使原有局部最优的聚类数优化为全局最优。实验结果可以验证该方法具有良好性能。
【关键词】K - means 聚类 聚类数 文本聚类 推荐系统
【分类号】TP18 G350

An Improved Method for Determining Optimal Number of Clusters in K - means Clustering Algorithm

Bian Peng[1,2] Zhao Yan[3] Su Yuzhao[1,2]
[1] (National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2] (Graduate University of Chinese Academy of Sciences, Beijing 100049, China)
[3] (Computer Science and Application Department, Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou 450015, China)

【Abstract】Based on the text clustering requirement from the embedded NSTL Recommending System, this paper researches on the BWP algorithm, and analyzes the shortage of the BWP. Then an improved algorithm is proposed to optimize the calculation of the distance within the single sample cluster. The improved algorithm enlarges the range of clusters number based on the BWP. Moreover, it changes the partial optimum into the whole optimum. At last, the test result shows it is effective and efficient.
【Keywords】K - means cluster Cluster number Text clustering Recommending system

1 引 言

随着个性化信息服务的蓬勃发展,机器学习技术得到了长足进步,K - means 聚类算法作为一种典型机器学习技术,得到了广泛应用。K - means 算法可以方便快捷地判定数据集合中样本的类别,属于“无监督的学习”,但是该算法必须事先指定一个聚类数 k,影响了其自动化水平,难以满足人们日益增长的个性化信息服务需求。

为此,不断有人提出自动确定最佳聚类数 k 的方法,这些方法的一般过程是设计一个函数指标,然后使用聚类

收稿日期:2011 - 07 - 12
收修改稿日期:2011 - 08 - 02

算法划分样本数据集合,聚类数在一定范围内变化,计算得到不同的函数指标值,最后找出最符合条件的指标值对应的聚类数作为最佳聚类数。

其中,较为典型的方法有 Calinski - Harabasz(CH)指标[1]、Davies - Bouldin(DB)指标[2]、Krzanowski - Lai(KL)指标[3]、Weighted Inter - intra(Wint)指标[4]、In - Group Proportion(IGP)指标[5],周世兵等[6]借鉴 Rousseeuw[7]的方法,提出了类间类内划分(Between - Within Proportion,BWP)指标,在寻找最佳聚类数方面进行了研究,并取得了一定成效。但该方法在单一样本类的处理方面存在不足,本文针对 BWP 方法的局限性,提出了改进的方法。实验表明,该方法具有更好的有效性。

2 K - means 聚类算法

K - means 算法由 MacQueen[8]提出,该算法是一种基于划分的聚类算法,它通过不断的迭代来实现聚类,当算法收敛到结束条件时就终止迭代过程,得出聚类结果[9]。传统的 K - means 算法[10]涉及到以下公式:

设聚类的样本集为 :$X = \{x_1, x_2, \cdots, x_n\}$,n 为样本总数,得到 C 个聚类中心为$\{z_1, z_2, \cdots, z_c\}$。令 a_j(j = 1,2,…,C)表示聚类的 C 个类别,n_j 表示第 j 类的样本数,则 :

$$z_j = \frac{1}{n_j} \sum_{x \in a_j} x$$

定义目标函数:

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n_j} d_{ij}(x_j, x_i)$$

目标函数 J 为每个样本数据点到相应聚类中心的距离平方和,即聚类的最小均方误差。

传统的 K - means 算法步骤[10]如下:

(1)随机指定 C 个样本点 $z_1(1), z_2(1), \cdots, z_c(1)$ 为初始聚类中心;

(2)按照距离最近的原则,对样本集合聚类,确定每个样本的类属关系;

(3)使用 $z_j = \frac{1}{n_j} \sum_{x \in a_j} x$, 计算新的聚类中心 $z_1(k), z_2(k), \cdots, z_c(k)$,k 表示迭代次数;

(4)重复执行步骤(2) - (4),直到聚类中心稳定为止。

K - means 算法容易实现对大规模数据集的聚类[9],且其简单直观并易于理解,因而得到广泛应用,但是其缺点也比较明显,除了抗干扰性较差、仅能处理数字属性等问题外,还需要预知类数目[11],虽然,X - means 算法[12]能进行最佳类数目估计,但是目前关于 k 值的准确界定依然还是一个难题。周世兵等[6]借鉴了 Rousseeuw[7]的指标说明方法,提出了类间类内划分指标,在寻找最佳聚类数方面进行了尝试,并取得了一定成效。

3 BWP 指标含义、不足及改进思路

3.1 BWP 指标含义

BWP 指标[6]的定义如下:

定义 1:令 K = {X, R} 为聚类空间, 其中 $X = \{x_1, x_2, \cdots, x_n\}$, 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的最小类间距离 b(j,i) 为该样本到其他每个类中样本平均距离的最小值, 即:

$$b(j,i) = \min_{1 \le k \le c, k \ne j} \left(\frac{1}{n_k} \sum_{p=1}^{n_k} ||x_p^{(k)} - x_i^{(j)}||^2 \right) \quad (1)$$

其中: k 和 j 表示类标,$x_i^{(j)}$ 表示第 j 类的第 i 个样本,$x_p^{(k)}$ 表示第 k 类的第 p 个样本,n_k 表示第 k 类中的样本个数,$\| \cdot \|^2$ 表示平方欧氏距离。

定义 2:令 K = {X, R} 为聚类空间, 其中 $X = \{x_1, x_2, \cdots, x_n\}$, 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的类内距离 w(j,i) 为该样本到第 j 类中其他所有样本的平均距离, 即:

$$w(j,i) = \frac{1}{n_j - 1} \sum_{q=1, q \ne i}^{n_j} ||x_q^{(j)} - x_j^{(i)}||^2 \quad (2)$$

其中:$x_q^{(j)}$ 表示第 j 类的第 q 个样本,并且 q≠i, n_j 表示第 j 类中的样本个数。

定义 3:令 K = {X, R} 为聚类空间, 其中 $X = \{x_1, x_2, \cdots, x_n\}$, 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的聚类距离 baw(j,i) 为该样本的最小类间距离和类内距离之和, 即:

$$baw(j,i) = b(j,i) + w(j,i) = \min_{1 \le k \le c, k \ne j} \left(\frac{1}{n_k} \sum_{p=1}^{n_k} ||x_p^{(k)} - x_i^{(j)}||^2 \right) + \frac{1}{n_j - 1} \sum_{q=1, q \ne i}^{n_j} ||x_q^{(j)} - x_j^{(i)}||^2 \quad (3)$$

定义 4:令 K = {X, R} 为聚类空间, 其中 $X = \{x_1, x_2, \cdots, x_n\}$, 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的聚类离差距离 bsw(j,i) 为该样本的最小类间距离和类内距离之差, 即:

$$bsw(j,i) = b(j,i) - w(j,i) = \min_{1 \le k \le c, k \ne j} \left(\frac{1}{n_k} \sum_{p=1}^{n_k} ||x_p^{(k)} - x_i^{(j)}||^2 \right) - \frac{1}{n_j - 1} \sum_{q=1, q \ne i}^{n_j} ||x_q^{(j)} - x_j^{(i)}||^2 \quad (4)$$

定义5:令 K = {X, R} 为聚类空间,其中 X = $\{x_1, x_2, \cdots, x_n\}$,假设 n 个样本对象被聚类为 c 类,定义第 j 类的第 i 个样本的类间类内划分指标 BWP(j,i) 为该样本的聚类离差距离和聚类距离的比值,即:

$$BWP(j,i) = \frac{bsw(j,i)}{baw(j,i)} = \frac{b(j,i) - w(j,i)}{b(j,i) + w(j,i)} \quad (5)$$

$avg_{BWP}(k)$ 表示数据集聚成 k 类时的平均 BWP 指标值,k_{opt}表示最佳聚类数。

$$avg_{BWP}(k) = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} BWP(j,i) \quad (6)$$

$$k_{opt} = \frac{argmax}{2 \leqslant k < n}\{avg_{BWP}(K)\} \quad (7)$$

最后,通过式(6)和式(7)找出使 $avg_{BWP}(k)$ 值(简称 BWP 值)最大的 k,即为最佳聚类数 k_{opt}。

3.2 BWP 指标的不足

BWP 指标对于聚类数较少,特别是远远小于样本总数的情况下,通常在聚类数 $c \leqslant Int(\sqrt{n})$(n 是样本总数)时有良好的表现,因此,BWP 方法将聚类数的搜索范围设为[2, $Int(\sqrt{n})$]。

笔者在设计嵌入式 NSTL 的个性化推荐系统时,实现同类检索词推荐需要对经过去重的样本数据聚类,由于用户输入的检索词(样本)在某个领域(类)较为集中,导致某些领域的检索词数较少,出现真实聚类数 $c > Int(\sqrt{n})$ 的情况。如果放大聚类数的搜索范围为[2, n),则随着聚类数接近样本总数,$avg_{BWP}(k)$ 的值呈现向数值 1 递增的趋势,使得聚类数 c = n - 1。

表1 sY4C 样本聚合

样本编号	属性1	属性2
1	99.209254	98.827389
2	98.669109	100.0419
3	199.71121	100.203441
4	100.314827	199.324139
5	99.763016	199.29908
6	200.43229	200.676832
7	199.865907	200.006268

表1是一个样本总数为7、真实聚类数为4的样本集合,命名为 sY4c 数据集合,使用 BWP 的聚类方法,受限于聚类数搜索范围[2, $Int(\sqrt{7})$],聚类数为2。若将聚类数搜索范围扩大为[2, 7),则当聚类数为6时,得到最大的 BWP 值,即分为6类。$avg_{BWP}(k)$ 的值随聚类数的变化如图1所示。

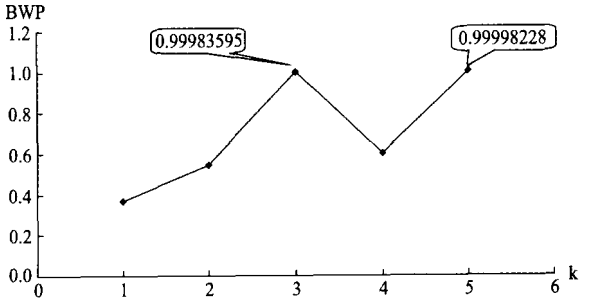

图1 sY4c 的 $avg_{BWP}(k)$ 值随聚类数的变化

BWP 值呈现向数值 1 递增的趋势是因为当类内样本数为1时,w(j,i) =0,无论 b(j,i) 等于多少,BWP(j,i) ≡1,而 BWP(j,i) 可能的最大值就是1。这样,随着越来越多的单一样本类出现,BWP = 1 的情况也越多,进而不断推高 $avg_{BWP}(k)$。这种问题在处理原来 BWP 有效的数据集时更加突出。如图2所示,数据集合 Y3c 是3类30个样本,聚类数搜索范围若在[2, Int($\sqrt{30}$)],聚类数为3;聚类数搜索范围若在[2, 30),聚类数为29。

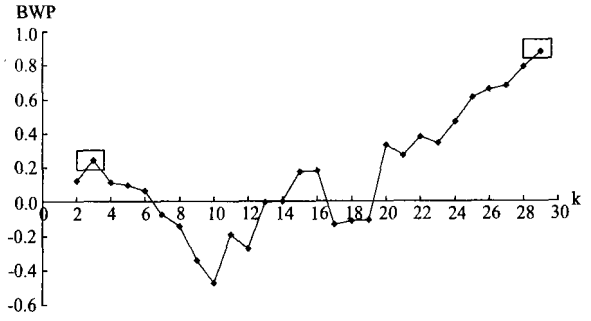

图2 聚类数范围扩展为[2, n)后,
Y3c 的 $avg_{BWP}(k)$ 值随聚类数的变化

笔者认为:当 w(j,i) 值很小趋近于0时,表示样本的类内距离很小,样本聚为一类的合理性升高。但是,当 w(j,i) =0 时,说明类内没有相似样本,仅有单一样本,无论其类间距离 b(j,i) 是多少都将 BWP 值赋为最大值1,这是 BWP 指标的一个不足。

3.3 BWP 指标的改进思路

要弥补 BWP 指标的不足,就需要找到样本单一成类的条件。举个简单的例子来说明单一样本独立成一类的过程。如图3所示,方块和菱形代表样本,单看左侧一列菱形距离方块从上而下依次变远,可以直观判断,菱形越来越可能与方块不属于一类。为了能够将这种直观判断量化,笔者在左侧原图上为方块所属区域加

了一个圆圈阴影(现实中阴影可能不都是圆形的),同时,将这个同等大小的阴影也赋给菱形,于是就形成了图3的右侧一列。如果两个阴影相重叠,则不能说明菱形与方块不属于同类;如果阴影不重叠且阴影间空隙越大,则说明菱形与方块不属于同类。量化阴影后进行分析,就可以判断单一样本独立成一类的合理性。

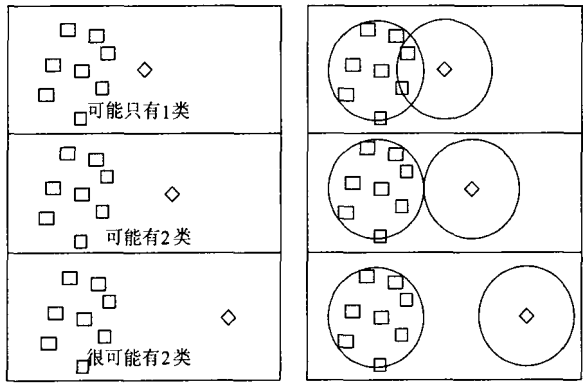

图3 单一样本独立成一类的过程

4 对BWP聚类有效性指标的改进

量化阴影的方法有很多,为了能够与原有BWP指标衔接,笔者选取BWP的类内距离w(j,i)表示方块部分的阴影,选取BWP指标的最小类间距离b(j,i)和类内距离w(j,i)为参考,对单一样本类提出4种阴影表示指标 $w_{single1}$、$w_{single2}$、$w_{single3}$、$w_{single4}$ 替代原来的类内距离w(j,i)。

4.1 最小类间距离b(j,i)/2

从图3中可以看出,菱形阴影的大小可能与离它最近的类有关系,当菱形阴影与方块阴影相分离的时候,正好是两个阴影相切的时候,这时,再放一个菱形的话,距离原来菱形的距离应不大于最小类间距离的一半,假设类内距离为最小类间距离的一半。

定义6:令 $K = \{X, R\}$ 为聚类空间,其中,$X = \{x_1, x_2, \cdots, x_n\}$,假设n个样本对象被聚类为c类,则:

$$w_{single1}(j,i) = b(j,i) \div 2 \text{ 且 } n_j = 1$$

进一步表示如下:

$$w_{single1}(j,i) = \min_{1\leqslant k\leqslant c, k\neq j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}||x_p^{(k)} - x_i^{(j)}||^2\right) \div 2 \text{ 且 } n_j = 1 \tag{8}$$

其中,第j类仅有一个样本,因此,i=0,以下定义同理,不再赘述。而且,k和j表示类标,$x_i^{(j)}$ 表示第j类的第i个样本,$x_p^{(k)}$ 表示第k类的第p个样本,n_k 表示第k类中的样本个数,$\|\cdot\|^2$ 表示平方欧氏距离。

4.2 最近类的最大类内距离

$w_{single1}$ 对阴影的估算只是机械地考虑了最小类间距离,未注意到离它最近类的类内距离是变化的,并且可能影响菱形的阴影大小,因此,假设菱形的类内距离是最近类的最大类内距离,若最近类仍是单一样本类,则用式(10)估计菱形的类内距离。即:

定义7:令 $K = \{X, R\}$ 为聚类空间,其中,$X = \{x_1, x_2, \cdots, x_n\}$,假设n个样本对象被聚类为c类,则:

$$w_{single2}(j,i) = \max_{1\leqslant m\leqslant c, 1\leqslant h\leqslant n_m} w(k,h),$$

$$\text{且}\frac{1}{n_m}\sum_{p=1}^{n_m}||x_p^{(k)} - x_i^{(j)}||^2 = \min_{1\leqslant k\leqslant c, k\neq j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}||x_p^{(k)} - x_i^{(j)}||^2\right), n_j = 1$$

进一步表示如下:

$$w_{single2}(j,i) = \max_{1\leqslant m\leqslant c, 1\leqslant h\leqslant n_m} \frac{1}{n_k - 1}\sum_{q=1, q\neq h}^{n_k}||x_q^{(k)} - x_h^{(k)}||^2$$

$$\text{且}\frac{1}{n_m}\sum_{p=1}^{n_m}||x_p^{(k)} - x_i^{(j)}||^2 = \min_{1\leqslant k\leqslant c, k\neq j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}||x_p^{(k)} - x_i^{(j)}||^2\right), n_j = 1 \tag{9}$$

其中,k和j表示类标,$x_i^{(j)}$ 表示第j类的第i个样本,$x_p^{(k)}$ 表示第k类的第p个样本,n表示第k类中的样本个数,$\|\cdot\|^2$ 表示平方欧氏距离,m类是距离j类最小的类。第j类仅有一个样本,因此,i=0。

4.3 所有样本类内距离的最大值

$w_{single2}$ 只考虑了离菱形最近的类,并未从全局角度考虑聚类的情况,因此,可以假设菱形的类内距离是所有类中的最大类内距离。

定义8:令 $K = \{X, R\}$ 为聚类空间,其中,$X = \{x_1, x_2, \cdots, x_n\}$,假设n个样本对象被聚类为c类,则:

$$w_{single3}(j,i) = \max_{2\leqslant k\leqslant c, 1\leqslant h\leqslant n_k} w(k,h) \text{ 且 } n_j = 1$$

进一步表示如下:

$$w_{single3}(j,i) = \max_{2\leqslant k\leqslant c, 1\leqslant h\leqslant n_k} \frac{1}{n_k - 1}\sum_{q=1, q\neq h}^{n_k}||x_q^{(k)} - x_h^{(k)}||^2 \text{ 且 } n_j = 1 \tag{10}$$

其中,$x_q^{(k)}$ 表示第k类中的第q个样本,n_k 表示第k类样本的个数。

4.4 最小的类内距离最大值

$w_{single3}$ 考虑到了全局的聚类情况,但与阴影最大的类相比,全局中阴影最小的类对于单一样本类的参考意义更大,因此,可以假设菱形的类内距离是类内距离

最大值集合中的最小值。

定义9:令 $K = \{X, R\}$ 为聚类空间,其中,$X = \{x_1, x_2, \cdots, x_n\}$,假设 n 个样本对象被聚类为 c 类,则:

$$w_{single4}(j,i) = \min_{1\leqslant k\leqslant c}\left(\max_{1\leqslant h\leqslant n_k} w(h,k)\right) \text{且 } n_j = 1$$

进一步表示如下:

$$w_{single4}(j,i) = \min_{1\leqslant k\leqslant c}\left(\max_{1\leqslant h\leqslant n_k} \frac{1}{n_k - 1}\sum_{q=1,q\neq h}^{n_k} ||x_q^{(k)} - x_h^{(k)}||^2\right) \text{且 } n_j = 1 \quad (11)$$

其中,$x_q^{(k)}$ 表示第 k 类中的第 q 个样本,n_k 表示第 k 类样本的个数。

5 改进的 BWP 算法

(1)选择聚类数的搜索范围 [2,n)。

(2)从 2 循环至 n-1。

①调用 K-means 算法;

②利用式(5) 计算单个样本的 BWP 指标值,若类内仅有一个样本,则分别调用式(8)-(11)计算单个样本的 BWP 指标值;

③利用式(6)计算平均 BWP 指标值。

(3)利用式(7) 计算最佳聚类数。

(4)输出最佳聚类数、有效性指标值和聚类结果,对应式(8)-(11),分别有 BWP1-BWP4。

6 实验与分析

本文通过对改进前的 BWP 方法和改进后的 BWP 方法进行数据实验,分析改进的效果。其中,改进前的方法根据聚类数范围的不同,分别简称 BWP 和 BWP0;改进后的方法分别有 4 个新指标:对应 $w_{single1}$、$w_{single2}$、$w_{single3}$、$w_{single4}$ 分别简称 BWP1、BWP2、BWP3、BWP4,通过实验,分析哪个指标效果最佳。为了克服 K-means 算法初始聚类中心对聚类结果产生影响,实验对所有指标均采用 100 种随机数据作为 K-means 聚类的种子,并选择 Weka 自带的 getSquaredError( )函数用来判断最佳的初始聚类中心。本文在方法实验中使用了 Java、MySQL Server5.1 和 Weka3.6 等工具,硬件平台环境是 Intel(R)Core(TM)2 Duo CPU 2.4GHz 和 2GB 内存,操作系统是 Windows XP。

6.1 验证改进效果

sY4c 是二维小样本数据,真实类数为 4,样本没有重复,在聚类数为 4 的情况下通过 K-means 可以正确聚类。sY3c、sY4c1、sY5c、sY6c、sY7c、sY8c 也是同理构造,实验数据如表 2 所示:

表 2 二维小样本数据实验

数据集	样本数	真实聚类数	BWP 方法在 [2, Int($\sqrt{n}$)] 范围上的聚类数	BWP0 在 [2,n)范围上的聚类数	BWP1 在 [2,n)范围上的聚类数	BWP2 在 [2,n)范围上的聚类数	BWP3 在 [2,n)范围上的聚类数	BWP4 在 [2,n)范围上的聚类数
sY3c(含单一实例类)	5	3	2	4	3	3	4	3
sY4c1(含单一实例类)	5	4	2	4	2	2	4	4
对 sY4c(含单一实例类)	7	4	2	6	4	4	4	4
对 sY5c(含单一实例类)	9	5	3	5	5	5	5	5
对 sY6c(含单一实例类)	10	6	3	8	6	6	6	6
对 sY7c(含单一实例类)	14	7	3	13	7	7	7	7
对 sY8c(含单一实例类)	25	8	2	24	7	8	8	8

从表 2 的实验可以看出,改进后的方法对二维小样本效果明显,特别是 BWP4 指标正确率最高,图 4 是 BWP4 对之前数据集合 sY4c 的聚类结果示例。

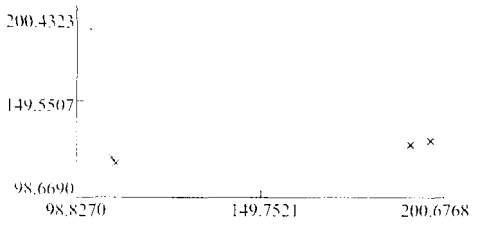

图 4 用 BWP4 对 sY4c 聚类的结果

Y50c 是二维数据,真实类数为 50,每个类含 1 至 3 个样本,样本没有重复,在聚类数为 50 的情况下通过 K-means 方法无法准确聚类,仅能近似聚类。Y100c 也是同理构造,实验数据及结果如表 3 所示。

从表 3 不难发现,改进后的指标对二维较大样本的聚类数更加接近真实聚类数,聚类效果优于 BWP 方法。图 5 是 Y50c 数据集的 BWP 值趋势图,刻画了 BWP 值随聚类数的变化趋势。从图 5 中看到,打开聚类数范围限制后,在前段(聚类数较少时),几种方法的重合度吻合得很好,但是从中段低值开始,原来的 BWP 结果曲线随着聚类数目的增加,呈现明显递增趋势,BWP 值在样本数最大值附近达到最大;而改进后的

表3 二维较大样本数据实验

数据集	样本数	真实聚类数	BWP在[2, Int(√n)]范围上的聚类数	BWP0在[2,n)范围上的聚类数	BWP1在[2,n)范围上的聚类数	BWP2在[2,n)范围上的聚类数	BWP3在[2,n)范围上的聚类数	BWP4在[2,n)范围上的聚类数
对Y50c(含单一实例类)的聚类数	97	50	10	96	40	40	40	40
对sY100c(含单一实例类)	196	100	4	195	69	69	69	69

4个指标值在达到最大值后,都能较好地控制递增的趋势,特别是BWP3、BWP4两个指标,在最大值出现后,没有出现随着聚类数的增加而递增的明显趋势,BWP3在聚类数接近最大样本量时表现得比BWP4方法还要稳定。

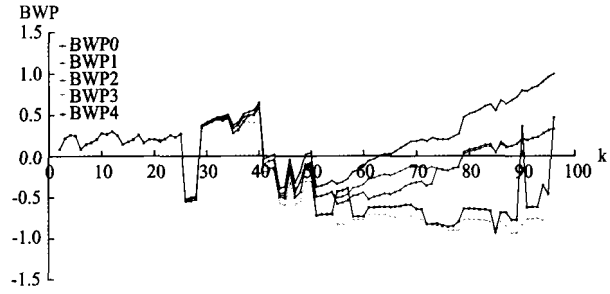

图5 Y50c数据集的BWP值趋势图

图6是sY100c数据集的BWP值趋势图,刻画了BWP值随聚类数的变化趋势,结论与图5相同。

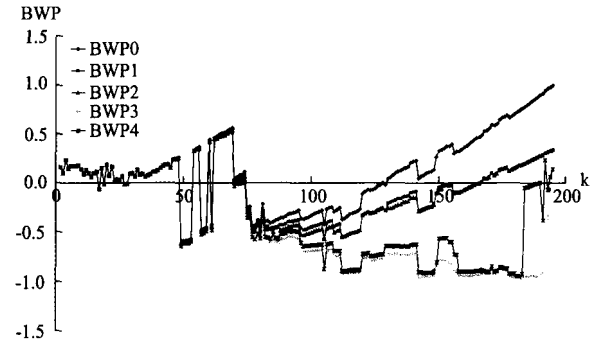

图6 Y100c数据集的BWP值趋势图

6.2 验证有效性未降低

文献[6]人工构造了三类数据集合,本文也按照其特点构造了类似的三个集合,用来验证当真实聚类数在范围[2, Int(√n)]之内时,本文所提的改进方法仍能达到原有BWP方法的效果,聚类结果如表4所示:

表4 有效性未降低的数据实验

数据集	样本数	真实聚类数	BWP在[2, Int(√n)]范围上的聚类数	BWP0在[2,n)范围上的聚类数	BWP1在[2,n)范围上的聚类数	BWP2在[2,n)范围上的聚类数	BWP3在[2,n)范围上的聚类数	BWP4在[2,n)范围上的聚类数
SM1	387	2	2	386	2	2	2	2
SM2	2 400	4	4	2 399	4	4	4	4
Y3c	131	3	3	130	3	3	3	3

从表4可见,当真实聚类数在范围[2, Int(√n)]之内时,如果采用改进的方法进行聚类,即使将聚类数的搜索范围扩大为[2,n),最终的聚类数也会落在[2, Int(√n)]之内。

6.3 嵌入式NSTL的检索词聚类

本文的实验使用了NSTL嵌入式资源服务的推荐系统原型中的检索词数据,事先邀请中国科学院国家科学图书馆的研究生参与使用该原型系统进行文献检索,历时两周。笔者运用K-means聚类分析检索词的相关性,将与用户输入的检索词同类的其他检索词推荐给用户。根据这些数据进行了实验研究,但用户的检索词可能会相差较大,导致类别较多,出现大于Int(√n)(n为样本总数)的情况,甚至很多类别中仅包含一个样本。这种情况适用本文确定最佳聚类数的方法。

数据集合含有3 308个非登录用户在嵌入式NSTL系统中使用的检索词,经过分类、去除停用词、去重和Weka的字符串-词向量函数StringToWordVector(样本对于不存在的属性维度赋值为空),形成123个样本187维的数据集合,样本分布较为稀疏,只有少部分样本比较相似。从表5和图7的实验结果可以看到,使用BWP方法,最佳聚类有2类,每类中的检索词相差甚远,没有起到聚类的作用;使用BWP0方法,最佳聚类有120类,只得出2个有效类,各含两个样本,其他119个样本都分到一类中,聚类的效果较差;当使用BWP4时,最佳聚类数为101类,找到12个非单一样本类,其中,11个类比较适合检索词推荐,聚类效果显著提高。

表5 数据集合 text 的文本聚类实验

数据集	样本数	维度	BWP 在 [2, Int(√n)]范围」的聚类数	BWP0 在 [2,n)范围上的聚类数	BWP1 在 [2,n)范围上的聚类数	BWP2 在 [2,n)范围上的聚类数	BWP3 在 [2,n)范围上的聚类数	BWP4 在 [2,n)范围上的聚类数
text	123	187	2	122	122	122	2	101

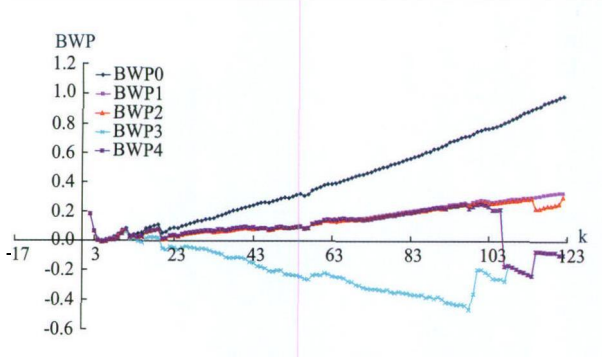

图7 text 数据集的 BWP 值趋势图

上述实验验证了新方法的有效性，改进后的4种 BWP 方法在未降低原有 BWP 方法有效性的前提下，进一步提高了其聚类的有效性，特别是 BWP4 方法效果表现最佳，并在文本聚类的应用中得到明显的改进效果。

7 结 语

BWP 方法产生的最优聚类数是局部最优，而不是全局最优，本文对 BWP 方法进行了改进，扩展了聚类数的范围，使其产生的最优聚类数在全局最优成为可能，提高了 BWP 方法对含单一样本类数据集合的适应能力，同时，该方法对原来 BWP 方法有效的数据集合也依然有效，将其应用在嵌入式 NSTL 个性化信息服务中取得了良好效果。本文的不足之处是：聚类的效果仍然受限于 K - means 算法本身，它的准确度依赖于 K - means 在聚类数正确时、能够正确分类的性能，并未对 K - means 算法进行优化。

参考文献：

[1] Calinski R, Harabasz J. A Dendrite Method for Cluster Analysis [J]. *Communications in Statistics*, 1974,3(1):1 - 27.

[2] Davies D L, Bouldin D W. A Cluster Separation Measure[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979,1(2):224 - 227.

[3] Dudoit S, Fridlyand J. A Prediction - based Resampling Method for Estimating the Number of Clusters in a Dataset[J]. *Genome Biology*, 2002,3(7):1 - 21.

[4] Dimitriadou E, Dolnicar S, Weingessel A. An Examination of Indexes for Determining the Number of Cluster in Binary Datasets [J]. *Psychometrika*, 2002,67(1):137 - 160.

[5] Kapp A V, Tibshirani R. Are Clusters Found in One Dataset Present in Another Dataset? [J]. *Biostatistics*, 2007,8(1):9 - 31.

[6] 周世兵，徐振源，唐旭清，K - means 算法最佳聚类数确定方法[J]. *计算机应用*，2010,30(8):1995 - 1998.

[7] Rousseeuw P J. A Graphical Aid to the Interpretation and Validation of Cluster Analysis[J]. *Journal of Computational and Applied Mathematics*, 1987,20(1):53 - 65.

[8] MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations[C]. In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1967:281 - 297.

[9] 谢娟英，蒋帅，王春霞，等，一种改进的全局 K - 均值聚类算法[J]. *陕西师范大学学报：自然科学版*，2010,38(2):18 - 22.

[10] 李飞，薛彬，黄亚楼，等，初始中心优化的 K - means 聚类算法[J]. *计算机科学*，2002,29(7):94 - 96.

[11] 姜园，张朝阳，仇佩亮，等. 用于数据挖掘的聚类算法[J]. *电子与信息学报*，2005,27(4):655 - 662.

[12] Pelleg D, Moore A. X - means: Extending K - means with Efficient Estimation of the Number of Clusters[C]. In: *Proceedings of the 17th ICML*. 2000: 727 - 734.

（作者 E - mail:bianpeng@mail.las.ac.cn）