ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 1,Jan. 2016,pp. 130-138
DOI:10.16337/j.1004-9037.2016.01.013

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

# 一种基于赋权联合概率模型的聚类算法

姬 波 叶阳东 卢红星

(郑州大学信息工程学院,郑州,450001)

**摘 要**:序列化信息瓶颈(Sequential information bottleneck, sIB)算法是一种广泛使用的聚类算法。该算法采用联合概率模型表示数据,对样本和属性的相关性有较好的表达能力。但是sIB算法采用的联合概率模型假设数据各个属性对聚类的贡献度相同,从而削弱了聚类效果。本文提出了赋权联合概率模型概念,采用互信息度量属性重要度,并构建赋权联合概率模型来优化数据表示,从而达到突出代表性属性、抑制冗余属性的目的。UCI数据集上的实验表明,基于赋权联合概率模型的WJPM_sIB算法优于sIB算法,在F1评价下,WJPM_sIB算法聚类结果比sIB算法提高了5.90%。

**关键词**:聚类;属性权重;联合概率模型;序列化信息瓶颈算法;互信息

**中图分类号**:TP181 **文献标识码**:A

## Clustering Algorithm Based on Weighting Joint Probability Model

Ji Bo, Ye Yangdong, Lu Hongxing

(School of Information Engineering, Zhengzhou University, Zhengzhou, 450052, China)

**Abstract**: Sequential information bottleneck (sIB) algorithm is one of the widely used clustering algorithms. The sIB algorithm applies the joint probability model to describe data, which has good ability to express the relationship between data samples and data attributes. However, the sIB algorithm suggests that all data attributes are equally important, which influences the clustering effect. To address the issue, the paper proposes the weighting joint probability model. The proposed model applies the mutual information measurement to the important level of data attributes so that to highlight representative attributes and depress redundancy attributes. Experiments on UCI datasets show that the proposed the weighting joint probability model (WJPM) sIB algorithm based on WJPM improves the F1 measure by 5.90% than the sIB algorithm.

**Key words**: clustering; attribute weight; joint probability model; sequential information bottleneck algorithm; mutual information

## 引 言

信息瓶颈(Information bottleneck, IB)方法[1-2]是起源于信息论的数据分析方法。基于IB方法的一系列聚类算法包含贪婪IB(aIB)算法[3]、序列化IB(Sequential IB, sIB)算法[3]、迭代的序列化IB

**基金项目**:国家自然科学基金(61170223)资助项目;国家自然科学基金、河南人才培养联合基金(U1204610)资助项目;河南省科技攻关计划(132102210404)资助项目。
**收稿日期**:2014-03-25;**修订日期**:2014-05-12

(Iterative sIB)算法[4]和图像语义标注算法[5]等。在这些算法中,sIB 算法由于具有执行效率高、可获得较优局部解等优点,被广泛应用于各种聚类问题[6-7]。sIB 算法的思路是采用联合概率模型表示数据,并在聚类目标簇数目确定的情况下,通过计算簇融合代价来迭代更新数据集合的质心分布。但是,联合概率模型假设数据各个属性对聚类的贡献度相同,从而导致 sIB 算法中未区分属性重要度,因而削弱了 sIB 算法的聚类效果。与 sIB 算法的联合概率模型不同,KNN, Rocchio, K-modes 和 K-means 等聚类算法均采用向量空间模型[8-9]来进行数据表示。研究表明,在向量空间模型上采用属性赋权来突出代表性属性、抑制冗余属性,可以有效地达到提高聚类效果目的[10-14]。因此,本文在联合概率模型基础上提出了赋权联合概率模型,即在构建联合概率模型时,采用互信息[15-16]度量属性以得到最优数据表示模型,在此基础上,构建了 WJPM_sIB 算法。

## 1　IB 方法和 sIB 算法

### 1.1　IB 方法

IB 方法起源于香农信息论。令信源为 $X$,输出编码为 $Y$,则转移概率分布为 $p(y_j|x_i)$,失真函数为 $d(x_i,y_j)$。图 1 中的信源编码器的目标是:在给出一个失真的限制 $D$ 的条件下,选择最优编码使信息传输率 $R$ 尽量小。

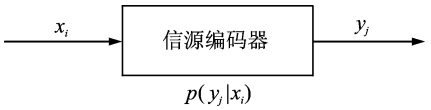



图 1　信源编码器

Fig. 1　Source encoder

率失真函数 $R(D)$为

$$R(D)=\min_{P_D} I(X;Y) \tag{1}$$

式中:$P_D=\{p(b_j\mid a_i):\overline{D}\leqslant D\}$;$\overline{D}=\sum_{i=1}^{n}\sum_{j=1}^{m}p(a_i,b_j)d(a_i,b_j)$。

$R(D)$可以有效衡量编码时的压缩程度,因此在通信中得以广泛应用。但是,其主要问题是难以选择出合适的失真函数。因此,IB 方法引入了原变量 $X$ 的相关变量 $Y$,并采用互信息度量代替了失真函数。IB 方法可以表示为

$$\hat{R}(\hat{D})=\min_{\{p(t|x):I(T;Y)\geqslant\hat{D}\}} I(T;X) \tag{2}$$

式中:$\hat{D}$ 为阈值;$I(T;X)$为 $X$ 和压缩变量 $T$ 的互信息;$I(T;Y)$为 $Y$ 和压缩变量 $T$ 的互信息;$p(t|x)$为转移概率。

IB 方法的目标是寻找原变量 $X$ 的压缩表示 $T$,使互信息 $I(T;X)$最小化的同时使互信息 $I(T;Y)$最大化,即在尽可能地压缩数据的同时尽可能地保存原有结构。

### 1.2　sIB 算法

sIB 算法是基于 IB 方法的聚类算法。算法主要步骤如下:

(1) 数据划分为 $k$ 个簇(随机划分或指定划分);

(2) 从已知簇 $t$ 中不重复依次取出其中所有单个样本 $x$,计算 $x$ 融合到其他簇中的融合代价 $\mathrm{cost}(\{x\},t)$,并将 $x$ 分配到融合代价最小的簇 $t'$。

$$t'=\arg\min_{t\in T}\mathrm{cost}(\{x\},t) \tag{3}$$

式中:$\mathrm{cost}(\{x\},t)=(p(x)+p(t))\times JS_{\pi_1,\pi_2}(p(y\mid x),p(y\mid t))$ 表示互信息值的减小量;$\pi_1$, $\pi_2$ 为权值。

(3) 返回第(2)步进行下一轮迭代。直到某轮迭代后,所有簇均未发生变化时,算法结束。

## 2　一种基于赋权联合概率模型的聚类算法

### 2.1　向量空间模型和联合概率模型

向量空间模型是以属性作为特征,将样本映射为高维特征空间的向量 $\boldsymbol{d}_x=(d_{x(1)},d_{x(2)},\cdots,d_{x(n)})$ 的数据表示模型。向量空间模型 $\boldsymbol{D}$ 为 $m\times n$ 矩阵,$m$ 为样本总数,$n$ 为特征数目。

$$\boldsymbol{D}=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix} \tag{4}$$

IB 方法中采用的联合概率模型是通过对向量空间模型 $\boldsymbol{D}$ 进行数据归一化(数据总和为 1)来得到联合概率矩阵 $\boldsymbol{D}'$

$$\boldsymbol{D}'=\begin{bmatrix} d'_{11} & d'_{12} & \cdots & d'_{1n} \\ d'_{21} & d'_{22} & \cdots & d'_{2n} \\ \cdots & \cdots & & \cdots \\ d'_{m1} & d'_{m2} & \cdots & d'_{mn} \end{bmatrix} \tag{5}$$

归一化为

$$d'_{ij}=\frac{d_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}d_{ij}} \tag{6}$$

IB 方法的联合概率模型与向量空间模型数据来源相同,描述目标相仿。但 $\boldsymbol{D}'$ 更侧重于描述样本与属性的相关性。

## 2.2 赋权联合概率模型

联合概率模型侧重于表达样本与属性的相关性,但其假设数据各个属性对聚类的贡献度相同,并导致 sIB 算法中未区分属性重要度,因而削弱了 sIB 算法的聚类效果。本文提出了赋权联合概率模型的构建方法(见图 2)。首先在向量空间模型上称量属性权重,得到赋权向量空间模型;然后归一化得到赋权联合概率模型。赋权联合概率模型结合了向量空间模型和联合概率模型的优点,既能充分地表达出样本与属性的相关性,也能充分地表达属性重要程度。

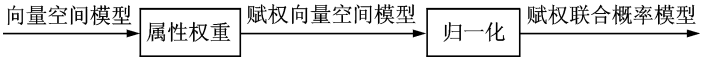



图 2 赋权联合概率模型

Fig. 2 Weighting joint probability model

## 2.3 基于赋权联合概率模型的聚类过程

基于赋权联合概率模型的聚类过程如图 3 所示。


图 3 基于赋权联合概率模型的聚类过程

Fig. 3 Clustering process based on weighting joint probability model

包括如下步骤。

(1) 样本映射。将样本映射为向量空间模型 $\boldsymbol{D}$。

(2) 平均互信息权重度量。

(a) 将 $\boldsymbol{D}$ 中列向量 $\boldsymbol{\alpha}_i$ 和 $\boldsymbol{\alpha}_j$ 组成一个新矩阵 $\boldsymbol{D}_i$;

(b) 采用式(6)归一化 $\boldsymbol{D}_i$ 得到联合概率 $p(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j)$和边缘概率 $p(\boldsymbol{\alpha}_i)$,$p(\boldsymbol{\alpha}_j)$;

(c) 计算互信息

$$I(\boldsymbol{\alpha}_i;\boldsymbol{\alpha}_j)=\sum_{i=1}^{m}\sum_{j=1}^{2}p(\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_j)\log\frac{p(\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_j)}{p(\boldsymbol{\alpha}_i)\,p(\boldsymbol{\alpha}_j)} \tag{7}$$

(d) 计算平均互信息

$$E(\boldsymbol{\alpha}_i)=\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}I(\boldsymbol{\alpha}_i;\boldsymbol{\alpha}_j) \tag{8}$$

(e) 计算属性权重

$$W(\boldsymbol{\alpha}_i)=\frac{E(\boldsymbol{\alpha}_i)}{\sum_{i=1}^{n}E(\boldsymbol{\alpha}_i)} \tag{9}$$

(3) 权重规约及赋权。上述计算会为每个属性赋予一个独特权重,这导致了权重的过度离散化。因此,采用数据规约方法将数据范围近似的若干权重值归整为一个值。权重规约为

$$\begin{gathered}w_i'=\left\lceil\frac{w_i}{\sum_{i=1}^{n}(w_i)/\theta}\right\rceil=\lceil w_i\times\theta\rceil\\ w''_i=\frac{w_i'}{\sum_{i=1}^{n}w_i'}\\ \sum_{i=1}^{n}(w_i)=1,\theta=\text{attr\_num}\times\text{scale}\end{gathered} \tag{10}$$

式中:$\theta$ 为规约系数,$\theta\geqslant 1$,$\theta$ 决定了离散程度,$\theta$ 越大,权重离散度越大,$\theta=1$ 时相当于未规约;attr_num 为属性数;scale 为调节系数,

最后,采用 $\boldsymbol{W}''$ 对 $\boldsymbol{D}$ 进行赋权,赋权公式如下

$$\boldsymbol{\alpha}_i'=w''_i\cdot\boldsymbol{\alpha}_i \tag{11}$$

式中:$\boldsymbol{\alpha}_i$ 为一个向量;$w''_i$ 为一个标量。

(4) 归一化。将赋权向量空间模型转换为赋权联合概率模型。

(5) 聚类。

### 2.4 WJPM_sIB 算法

基于赋权联合概率模型的 WJPM_sIB 聚类算法如下:

输入:向量空间矩阵 $\boldsymbol{D}$;目标簇个数 $k$;权重调节系数集合 SCALE;平衡因子 $Beta$。

输出:$\boldsymbol{D}$ 到 $k$ 簇的多个划分集 $\boldsymbol{TT}$。

步骤 1　计算权重向量 $\boldsymbol{W}$。

步骤 2　权重规约

For every scale∈SCALE:

(1) 计算权重规约系数 $\theta$,$\theta$=attr_num×scale,attr_num 为属性数;

(2) 规约,得到权重向量 $\boldsymbol{W}''$;

(3) 赋权,得到赋权矩阵 $\boldsymbol{D}_i$;

(4) 归一化 $\boldsymbol{D}_i$,得到赋权联合概率矩阵 $\boldsymbol{D}_i'$。

End for

步骤 3　聚类

For every $\boldsymbol{D}_i'\in\boldsymbol{D}$

(1)　随机初始化 $\boldsymbol{D}_i'$ 到 $k$ 簇的划分;

(2)　While not Done

(3)　　Done=TRUE;

(4) For every $x \in X$:

(5) 从当前簇 $t(x)$ 中移除 $x$,形成单独的簇 $\{x\}$;

(6) $t' = \arg\min_{t \in T} \cos t(\{x\}, t)$

(7) 把 $x$ 合并到 $t'$ 中;

(8) End for

(9) End while

(10) 得到一个划分 $\boldsymbol{T}$,加入集合 $\boldsymbol{TT}$

End for

输出 $\boldsymbol{TT}$

算法时间复杂度为 $O(c \cdot l|X||T||Y|)$。其中 $|X|$ 为样本数;$|Y|$ 为属性数;$c$ 为集合 SCALE 中元素数;$l$ 为聚类循环次数。

## 3 实验评价与结果

### 3.1 实验数据集

实验数据取自 20-Newsgroup 数据集、Reuters-21578 数据集和 4Univesities 数据集,如表 1 所示。采用 rainbow 软件的数据集预处理参数如下:

(1) 20-Newsgroup,"-istext-avoid-uuencode -skip-header -O 2";

(2) Reuters-21578, "-O 2";

(3) 4Univesities,"-skip-html -no-stoplist -O 2"。

**表 1 实验数据集**

**Tab. 1 Experimental dataset**

| 名称 | 来源 | 类别 | 每组文档数 | 总文档数 | 属性 |
|---|---|---|---|---|---|
| NS2_1,2,3 | | 2 | 250 | 500 | 2 000 |
| NS5_1,2,3 | 20Newsgroup | 5 | 100 | 500 | 2 000 |
| NS10_1,2,3 | | 10 | 50 | 500 | 2 000 |
| RS2_1,2,3 | | 2 | 250 | 500 | 2 000 |
| RS5_1,2,3 | Reuters | 5 | 100 | 500 | 2 000 |
| RS8_1,2,3 | | 8 | 50 | 400 | 2 000 |
| US7_1,2,3 | 4Universities | 7 | 60 | 420 | 2 000 |
| RM2_1,2,3 | Reuters | 2 | 250 | 4 000 | 2 000 |
| NM20_1,2,3 | 20Newsgroup | 20 | 250 | 5 000 | 2 000 |

### 3.2 实验评价方法

本文采用 $F1$ 评价来评估算法的优劣。$F1$ 评价是基于准确率 $PR$ 和召回率 $RE$ 的一个综合评价指标。$F1$ 越高,反映出聚类结果越有优势。

$$PR = \frac{\sum_{c=1}^{k} \frac{a_c}{a_c + b_c}}{k}, RE = \frac{\sum_{c=1}^{k} \frac{a_c}{a_c + c_c}}{k}, F1 = \frac{2 \times PR \times RE}{PR + RE} \tag{12}$$

式中:$a_c$ 为正确归到簇 $c$ 的样本数目;$b_c$ 为不应该归到簇 $c$ 但却归到簇 $c$ 的样本数目;$c_c$ 为错误的拒绝了原本属于簇 $c$ 的样本数目;$k$ 为簇数;$m$ 是数据样本总数。

### 3.3 实验结果

WJPM_sIB 算法的基本参数与 sIB 算法一致,平衡因子 Beta $=\infty$; 循环次数 Restarts $=$ 15 。权重

规约调节系数 SCALE = {0.1, 0.2, 0.5, 1, 2, 3, 5, 10, 20, 50, 100, 200, 500, 1 000}。

3.3.1　赋权效果的实验结果与分析

在 *F*1-Measure 评价下的实验结果见表 2 和图 4。从中可以看出:WJPM_sIB 算法在全部 27 个数据集的 26 个占优,在 1 个数据集上与 sIB 算法相同;最高提高 16.87%,平均提高 5.90%。

**表 2　赋权实验结果与分析**

**Tab. 2　Weighting experiment result & analysis**

| 数据集 | sIB | WJPM_sIB | 增长比例/% |
|---|---|---|---|
| NS2_1 | 93.970 0 | 94.770 0 | 0.85 |
| NS2_2 | 92.795 0 | 93.390 0 | 0.64 |
| NS2_3 | 92.605 0 | 93.635 0 | 1.11 |
| NS5_1 | 91.955 0 | 93.765 0 | 1.97 |
| NS5_2 | 94.620 0 | 95.840 0 | 1.29 |
| NS5_3 | 96.015 0 | 96.415 0 | 0.42 |
| RS2_1 | 91.620 7 | 91.736 8 | 0.13 |
| RS2_2 | 94.059 3 | 95.504 9 | 1.54 |
| RS2_3 | 94.679 2 | 95.744 8 | 1.13 |
| RS5_1 | 79.299 6 | 84.988 2 | 7.17 |
| RS5_2 | 78.439 7 | 80.953 5 | 3.20 |
| RS5_3 | 80.444 7 | 89.027 9 | 10.67 |
| RS8_1 | 64.124 8 | 73.177 9 | 14.12 |
| RS8_2 | 65.939 9 | 74.369 7 | 12.78 |
| RS8_3 | 67.897 6 | 67.897 6 | 0.00 |
| US7_1 | 40.923 3 | 47.434 0 | 15.91 |
| US7_2 | 43.519 2 | 49.473 5 | 13.68 |
| US7_3 | 45.278 3 | 50.178 1 | 10.82 |
| NS10_1 | 64.055 1 | 74.859 5 | 16.87 |
| NS10_2 | 65.881 5 | 73.596 7 | 11.71 |
| NS10_3 | 69.302 7 | 71.810 0 | 3.62 |
| NM20_1 | 54.917 5 | 61.866 3 | 12.65 |
| NM20_2 | 56.171 5 | 61.409 2 | 9.32 |
| NM20_3 | 57.641 7 | 61.913 4 | 7.41 |
| RM2_1 | 92.867 6 | 92.907 7 | 0.04 |
| RM2_2 | 92.692 1 | 92.752 3 | 0.06 |
| RM2_3 | 92.431 9 | 92.567 1 | 0.15 |
| 平均 | | | 5.90 |

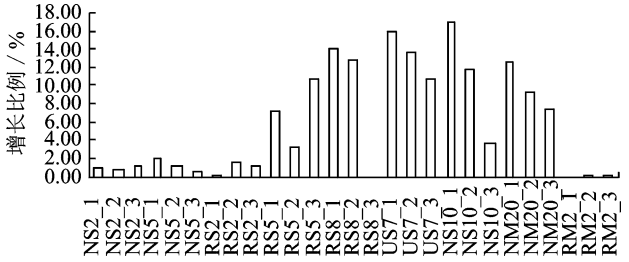



图 4　WJPM_SIB 增长比例直方图

Fig. 4　Improvement scale histogram for WJPM_SIB

3.3.2 权重规约系数 $\theta$ 分析

为了考察权重规约程度对赋权的影响，设定[0.1,1 000]范围内的 14 个调节系数，SCALE = {0.1, 0.2,0.5,1,2,3,5,10,20,50, 100,200,500,1 000}。

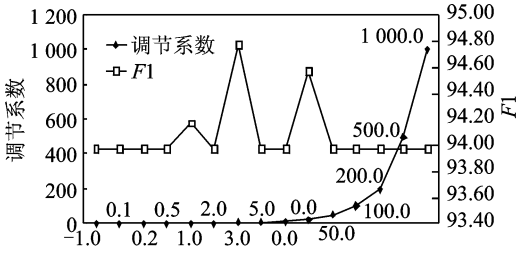
(1) 调节系数变化对比实验

图 5(a)中给出了数据集 NS2_1 上 $F1$ 评价和调节系数相关变化曲线，图 5(b)中给出了算法用时和调节系数相关变化曲线。其中，调节系数－1.0 代表未赋权。从中可以看出：

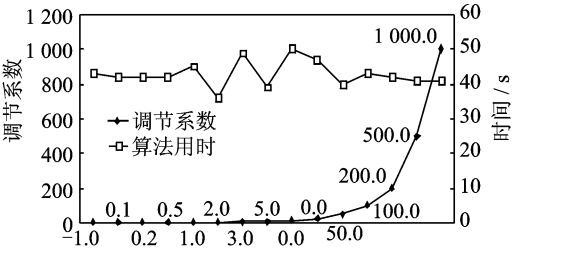
(a) 调节系数对聚类结果影响显著。调节系数为 3.0 时，$F1$ 最大。调节系数为 0.1,0.2,0.5,5.0, 10.0 时，$F1$ 与未赋权时相同。

(b) 调节系数对算法用时无显著影响。调节系数 10.0 时，运行时间最长，调节系数为 2.0 时，运行时间最短。这说明规约时调节系数的取值对运行时间影响不大，权重规约的引入并未降低算法效率。

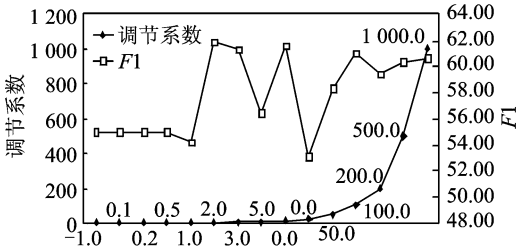
图 5(c,d)中给出了数据集 NM20_1 的结果，从中可以看出和数据集 NS2_1 类似的结果。



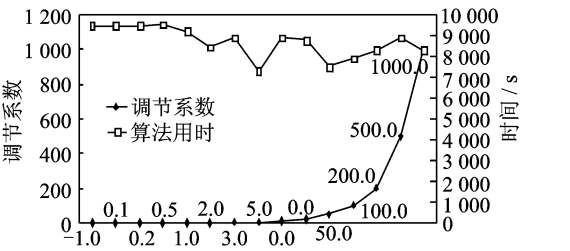

(a) $F1$评价和调节系数变化曲线(NS2_1)
(a) $F1$measure and adjust coefficient curve (NS2_1)




(b) 算法用时和调节系数变化曲线(NS2_1)
(b) Algorithm time consuming and adjust coefficient curve (NS2_1)




(c) $F1$评价和调节系数变化曲线(NM20_1)
(c) $F1$measure and adjust coefficient curve (NM20_1)




(d) 算法用时和调节系数变化曲线(NM20_1)
(d) Algorithm time consuming and adjust coefficient curve (NM20_1)

图 5 调节系数变化对比实验

Fig. 5 Adjustment coefficient comparison

(2) 经验调节系数

上述实验中反映了权重规约系数 $\theta$ 的取值会显著影响聚类质量。同时，实验中通过应用一系列调节系数后取最优的方式获得了最佳聚类质量。表 3 中给出了 27 个数据集的最佳调节系数、最佳规约系数、规约前权重向量的相异权重数和最佳规约后权重向量的相异权重数。从中可以看出：

(a) 所有数据集的规约前权重向量的相异权重数在[1 269,1 993]之间，接近数据集实际属性数，这说明规约前互信息权重度量几乎为每个属性设定了一个独特权重；

(b) 数据集 RS8_3 的最佳规约后权重向量的相异权重数为 1，这代表 RS8_3 最佳规约后全部属性权重相同，其最佳规约等同于未加权的情况；RS5_1 和 RM20_2 分别为 85 和 435。除此之外的 24 个数据集的最佳规约后权重向量的相异权重数在[2,28]的范围，这说明最佳规约后达到了将近似权重值合

并的效果,而且合并是有效的。

(c) 去除两种较为异常的数据集 RS5_1 和 RM20_2 后,规约前权重向量的相异权重数的平均值为 1 783,数量级为 $10^3$;最佳规约后的平均值为 7,数量级为 $10^1$。这说明最佳规约将相异权重数降低了 2 个数量级,降低幅度非常显著。

(d) 除了数据集 RS5_1 和 RM20_2 的最佳调节系数较大外,其他数据集的最佳调节系数在[0.5,20]的较小范围内。这说明在实际使用中,可以将调节系数经验设定在[0.5,20]的区间。

**表 3　调节系数经验值**

**Tab. 3　Adjustment coefficient empirical value**

| 数据集 | 属性数 | 最佳调节系数 | 最佳规约系数 | 规约前相异权重数 | 规约后相异权重数 |
|---|---|---|---|---|---|
| NS2_1 | 1 952 | 3 | 5 856 | 1 707 | 4 |
| NS2_2 | 1 947 | 5 | 9 735 | 1 705 | 6 |
| NS2_3 | 1 894 | 1 | 1 894 | 1 679 | 2 |
| NS5_1 | 1 948 | 1 | 1 948 | 1 866 | 2 |
| NS5_2 | 1 931 | 20 | 38 620 | 1 839 | 20 |
| NS5_3 | 1 936 | 1 | 1 936 | 1 859 | 2 |
| RS2_1 | 1 478 | 1 | 1 478 | 1 269 | 2 |
| RS2_2 | 1 533 | 1 | 1 533 | 1 328 | 2 |
| RS2_3 | 1 541 | 1 | 1 541 | 1 342 | 2 |
| RS5_1 | 1 922 | 100 | 192 200 | 1 864 | 85 |
| RS5_2 | 1 882 | 1 | 1 882 | 1 819 | 2 |
| RS5_3 | 1 890 | 2 | 3 780 | 1 815 | 3 |
| RS8_1 | 1 868 | 2 | 3 736 | 1 775 | 3 |
| RS8_2 | 1 875 | 5 | 9 375 | 1 811 | 6 |
| RS8_3 | 1 861 | 0.5 | 930.5 | 1 762 | 1 |
| US7_1 | 1 952 | 20 | 39 040 | 1 928 | 28 |
| US7_2 | 1 929 | 20 | 38 580 | 1 888 | 25 |
| US7_3 | 1 940 | 20 | 38 800 | 1 914 | 25 |
| NS10_1 | 1 904 | 20 | 38 080 | 1 792 | 20 |
| NS10_2 | 1 875 | 3 | 5 625 | 1 754 | 4 |
| NS10_3 | 1 885 | 10 | 18 850 | 1 771 | 10 |
| NM20_1 | 2 000 | 2 | 4 000 | 1 993 | 3 |
| NM20_2 | 2 000 | 1 000 | 2 000 000 | 1 992 | 435 |
| NM20_3 | 2 000 | 1 | 2 000 | 1 993 | 2 |
| RM2_1 | 2 000 | 1 | 2 000 | 1 991 | 2 |
| RM2_2 | 2 000 | 1 | 2 000 | 1 990 | 2 |
| RM2_3 | 2 000 | 1 | 2 000 | 1 989 | 2 |

## 4　结束语

序列化 IB(sIB)算法采用的联合概率模型能较好地表达样本和属性的相关性,但该模型假设数据各个属性对聚类的贡献均匀,从而影响了 sIB 算法的聚类效果。本文提出采用互信息度量属性重要度来构建赋权联合概率模型,并指出通过权重规约可以进一步达到得到最优数据表示的目的。同时,在 WJPM_sIB 算法上通过实验给出了权重规约系数 $\theta$ 的经验取值。下一步可能的工作包括互信息赋权方法之外的其他赋权方法的适用性和权重规约系数思想的理论分析等问题。

**参考文献:**

[1] Fabrizio R, Nicola D M. Applying the information bottleneck to statistical relational learning [J]. Machine Learning, 2012, 86(1): 89-114.

[2] Gedeon T, Parker A E, Dimitrov A G. The mathematical structure of information bottleneck methods [J]. Entropy, 2012,

14, 456-479.

[3] Slonim N. The information bottleneck: Theory and application [D]. Jerusalem, Israel: The Hebrew University of Jerusalem, 2002.

[4] Yuan H Q, Ye Y D. Iterative sIB algorithm [J]. Pattern Recognition Letters, 2011, 32(4): 606-614.

[5] 夏利民,谭立球,钟洪. 基于信息瓶颈算法的图像语义标注[J]. 模式识别与人工智能, 2008, 21(6):812-818.
Xia Limin, Tan Liqiu, Zhong Hong. Semantic annotation of image based on information bottleneck method[J]. Pattern Recongnition and Artifical Intelligence, 2008, 21(6): 812-818.

[6] Lou Z, Ye Y D, Zhu Z. Information bottleneck with local consistency [J]. Lecture Notes in Computer Science, 2012, 7458: 285-296.

[7] Slonim N, Friedman N, Tishby N. Unsupervised document classification using sequential information maximization [C]// Proc of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2002: 129-136.

[8] 陆玉昌,鲁明羽,李凡,等. 向量空间法中单词权重函数的分析和构造[J]. 计算机研究与发展, 2002, 39(10): 1205-1210.
Lu Yuchang, Lu Mingyu, Li Fan et al. Analysis and construction of word weighing function in VSM[J]. Journal of Computer Research and Development, 2002, 39(10): 1205-1210.

[9] Soucy P, Mineau G W. Beyond TFIDF weighting for text categorization in the vector space model [C]// Proc of Int'l Joint Conf Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2005: 1130-1135.

[10] 王骏,王士同,邓赵红. 特征加权距离与软子空间学习相结合的文本聚类新方法[J]. 计算机学报, 2012, 35(8): 1655-1665.
Wang Jun, Wang Shitong, Deng Zhaohong. A novel test clustering algorithm based on feature weighting distance and soft subspace learning[J]. Chinease Journal of Computers, 2012, 35(8): 1655-1665.

[11] 黄剑雄,丁建立. 基于集成赋权模糊积分的信息系统风险评价[J]. 数据采集与处理, 2011, 26(4): 485-489.
Huang Jianxiong, Ding Jianli. Evaluation on risks of information system based on integrated weight fuzzy integral method [J]. Journal of Data Acquisition and Processing, 2011, 26(4): 485-489.

[12] He Z Y, Xu X F, Deng S C. Attribute value weighting in k-modes clustering [J]. Expert Systems with Applications, 2011, 38(12): 15365-15369.

[13] Jing L, Ng M K, Huang Z X. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(8): 1026-1041.

[14] Sparck Jones K. IDF term weighting and IR research lessons [J]. Journal of Documentation, 2004, 60(6): 521-523.

[15] Shannon C E. A mathematical theory of communication [J]. The Bell System Technical Journal, 1948, 27: 379-423.

[16] Yates R B, Neto B R. Modern information retrieval [M]. Boston, MA, USA: Addison-Wesley Longman Publishing Co Inc, 1999.

**作者简介:**

**姬波**(1973-),男,博士,副教授,研究方向:人工智能、模式识别和信息论,E-mail:iebji@zzu.edu.cn。

**叶阳东**(1962-),男,博士,教授,博士生导师,研究方向:知识工程、机器学习和数据库, E-mail: ieydye@zzu.edu.cn。

**卢红星**(1965-),通信作者,男,副教授,研究方向:模式识别, E-mail: iehxlu@zzu.edu.cn。