第 30 卷 第 6 期
2007 年 6 月
计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS
Vol. 30 No. 6
June 2007

一种自动确定参数的 sIB 算法

叶阳东[1)] 刘 东[1)] 贾利民[2)] LI Gang[3)]

[1)](郑州大学信息工程学院计算机科学系 郑州 450052)

[2)](北京交通大学轨道交通控制与安全国家重点实验室 北京 100044)

[3)](迪肯大学信息技术学院 维多利亚 澳大利亚)

摘 要 针对 sIB 算法的压缩变量参数的确定问题,采用最小描述长度原理,构建一种自动确定参数的 AsIB 算法.算法使用一种有效的编码方案对数据分析模型和相应的数据进行描述,将最小描述长度的模型作为选择标准,从而有效发现了数据蕴含的特征模式数目.实验表明:AsIB 算法所采用的编码方案有效,在不设定模式数目的情况下,能够正确发现数据集所蕴含的模式.该算法解决了现 sIB 算法对先验知识的依赖问题,将能拓展其在多维数据的自动降维分析和模式提取等方面的应用.

关键词 IB 理论;sIB 算法;AsIB 算法;最小描述长度原理;模型选择

中图法分类号 TP18

An sIB Algorithm for Automatically Determining Parameter

YE Yang-Dong[1)] LIU Dong[1)] JIA Li-Min[2)] LI Gang[3)]

[1)](*Department of Computer Science, School of Information Engineering, Zhengzhou University, Zhengzhou* 450052)

[2)](*State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing* 100044)

[3)](*School of Information Technology, Deakin University, 221 Burwood Highway, Vic* 3125, *Australia*)

Abstract To solve the problem of determining the compression variable parameter for sIB algorithm, this paper proposes an AsIB algorithm for automatically determining parameter based on minimum description length principle. An efficient encoding scheme is designed to estimate the description length of the solution model of sIB algorithm and the original data given the model respectively, and the minimum description length model is selected as a criterion to find the number of feature patterns hidden in dataset. Experiment results show that the encoding scheme in AsIB is efficient to recover the true feature pattern in dataset without the requirement of setting category number of feature pattern. AsIB algorithm removes the dependency of empirical knowledge for sIB algorithm, which widens its applications in areas such as automatic dimension reduction and pattern extraction, etc.

Keywords IB theory; sIB algorithm; AsIB algorithm; minimum description length principle; model selection

1 引 言

1999 年,Tishby、Pereira 和 Bialek 开创性地提出一种基于信息论的数据分析方法——IB 方法[1](The Information Bottleneck Method).该方法通过将数据对象压缩到一个事先定义好的“瓶颈”变量中极大地保持其与另一数据对象的相关性,有效地发

收稿日期:2006-03-31;修改稿收到日期:2007-01-26.本课题得到国家自然科学基金(600332020)和河南省自然科学基金(0411012300)资助.**叶阳东**,男,1962 年生,博士,教授,博士生导师,研究领域为知识工程、智能系统、数据库.E-mail:ieydye@zzu.edu.cn.**刘 东**,男,1981 年生,硕士,研究方向为机器学习.**贾利民**,男,1963 年生,博士,教授,博士生导师,研究领域为计算智能、智能控制与自动化.**LI Gang**,男,1973 年生,博士,讲师,研究方向为机器学习、数据挖掘.

现了数据对象间隐含存在的模式. IB 方法已经在文本聚类[2-4]、图像聚类[5-6]、星系光谱分析[7]、基因表达式分析[8]、神经系统编码分析[9]、自然语言处理[10]、语音识别[11]等多个领域得到了应用,且取得了令人满意的效果. 2002 年,Slonim 将 IB 的原理、算法与应用在其博士论文中进行了系统的整理与归纳,并将其命名为 IB 理论[12](The Information Bottleneck Theory). 由于 IB 理论的理论基础好且应用范围广泛,近几年来,更多的研究人员被吸引到该领域的研究中,使得 IB 理论得到不断的扩充和完善,相继出现了 IBSI[13]、CCIB[14]、MIB[15-16]、GIB[17]、SDR[18]等相关的研究成果.

sIB 算法由 Slonim 于 2002 年提出[2],该算法将待分析的数据对象按照其与另一数据对象的相关性进行"硬"划分,使得划分在一起的对象充分体现出源数据对象蕴含的某个特征模式. 相对于 1999 年 Tishby 等提出的 iIB[1]、dIB[1]以及 Slonim 提出的 aIB[19]等 IB 算法,sIB 算法具有较低的时间和空间复杂度且保证可以得到问题的局部优解. 这种良好的特性有益于 sIB 算法的实际应用. 但是,sIB 算法存在两个固有的缺点:(1)随机选取的初始解导致算法容易陷入局部解;(2)压缩变量参数需要由用户指定. 如果没有关于数据的先验知识,该参数的确定极其困难. 这在一定程度上使得 sIB 算法偏离了无指导学习的本质意义并有应用的局限性. 为了克服缺点(1),文献[12,20]提出了多种方法对 sIB 算法得到的局部解进行优化,取得了较好的效果. 但针对缺点(2)进行的工作还很少. 2004 年,Still 等人根据 IB 算法对有限数据样本概率分布的依赖性,推导出概率偏差值与压缩变量参数的关系,得到了特定有限数据集的最大特征模式数目[21]. Niu 等人从 sIB 算法在文本聚类的实际应用问题出发,将文本特征选取与寻找文本真实聚类数目联系在一起,设计出一种聚类目标函数,伴随该目标函数的优化,最终发现了文本数据集的特征子集和真实聚类数目[22]. Still 等人的工作对有限数据集的依赖性很强,成功发现与压缩变量参数相对应的类别数依赖于概率估计的偏差;Niu 提出的方法仅针对文本聚类.

本文旨在为 sIB 算法引入一种内在的机制使其自动确定压缩变量参数,这一问题可以通过模型选择来有效解决. 最小描述长度(Minimum Description Length,MDL)原理能够作为选择某种学习模型的依据[23]. 其基于信息编码优化理论,用描述长度作为模型选择的优化标准,把具有最小描述长度的模型作为选定的模式. 理论上可以证明,如果有足够多的数据,用最小描述长度原理设计的模型能够收敛到理想或真正的模型上[24].

本文提出了一种能够自动确定参数的 AsIB 算法,其相对于已有工作的不同之处在于:

(1)提出了一种自动确定参数的 AsIB 算法,相对于现 sIB 算法能够正确发现模式数目;

(2)将最小描述长度原理引入 IB 理论的研究中,用模型选择方法解决了压缩变量参数确定的问题;

(3)设计了一种有效的数据和模型编码方案,用其计算出的描述长度能够较好地刻画相应数据的模式特征.

第 2 节描述了相关的背景知识;第 3 节介绍 AsIB 算法;第 4 节进行算法的实验分析;最后给出结论和下一步的工作.

2 背景知识

2.1 标记约定

大写字母 X,Y,T 表示离散随机变量;小写字母 x,y,t 表示以上随机变量的某个特定取值;$\mathcal{X}$、$\mathcal{Y}$、$\mathcal{T}$表示离散随机变量 X,Y,T 的值域集合;$p(x)$等价于 $p(X=x)$,表示随机变量 X 取值为 x 的概率;$X\sim p(x)$表示 X 服从概率分布 $p(x)$;$(X,Y)\sim p(x,y)$表示随机变量 X 和 Y 服从联合概率分布 $p(x,y)$,其边缘分布为 $p(x)=\sum_y p(x,y)$,$p(y)=\sum_x p(x,y)$;$p(y|x)$,$p(t|x)$和 $p(y|t)$表示条件概率分布;$|\mathcal{X}|$,$|\mathcal{Y}|$,$|\mathcal{T}|$分别表示 $\mathcal{X}$、$\mathcal{Y}$ 和 $\mathcal{T}$中包含元素值的个数.

2.2 相关定义

定义 1. 离散随机变量 X 和 Y 之间的互信息 $I(X;Y)$定义为

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}.$$

互信息度量了随机变量 X 与 Y 相互包含的程度,即两者公有信息量的多少. 它反映了两个变量的关联程度. 当变量 X 与 Y 相互独立即 $p(x,y)=p(x)\times p(y)$ 时,两者之间的互信息 $I(X;Y)=0$. 互信息满足对称性:$I(X;Y)=I(Y;X)$.

定义 2. $p(x)$和 $q(x)$之间的相对熵或 Kullback Leibler(KL)距离定义为

$$D_{KL}(p\parallel q)=\sum_{x\in X}p(x)\log\frac{p(x)}{q(x)}.$$

约定 $0\log(0/q)=0$，$p\log(p/0)=\infty$. KL 距离描述了两个概率分布的偏离程度. 当 $p(x)=q(x)$ 时，两者的 KL 距离为 0. 当分布差异变大时，相应 KL 距离随之变大. KL 距离不满足对称性，即 $D(p\|q)\neq D(q\|p)$，也不满足三角不等式.

定义 3.　$p(x)$ 和 $q(x)$ 之间的 Jensen Shannon (JS)距离定义为

$$JS_{\pi_1,\pi_2}(p\|q)=\pi_1 D_{KL}(p\|r)+\pi_2 D_{KL}(q\|r).$$

其中，π_1 和 π_2 是权值，满足 $0<\pi_1,\pi_2<1$，$\pi_1+\pi_2=1$. 概率分布 $r=\pi_1 p+\pi_2 q$.

JS 距离的大小反应了概率分布 $p(x)$ 和 $q(x)$ 来源于同一分布的可能性. JS 距离是对称的，但不满足三角不等式.

定义 4.　给定离散随机变量 X，Y 和 Z，当 Z 的条件分布依赖于 Y 且独立于 X 即 $p(z|x,y)=p(z|y)$ 时，称 X，Y 和 Z 形成 Markov 链，记作 $X\rightarrow Y\rightarrow Z$.

2.3　IB 理论

IB 理论起源于香农著名的率失真理论，率失真理论寻求源随机变量 $X\sim p(x)$ 压缩到压缩变量 T 的编码方案并通过失真度量函数 $d(x,t)$ 度量该过程中产生的失真，使该编码所产生的期望失真小于预先指定的值 D 的条件下压缩信息 $I(X;T)$ 达到最小. 该理论可形式化表示为

$$R(D)=\min_{\{p(t|x):E(d(x,t))\leq D\}} I(X;T) \tag{1}$$

式(1)中，$E(d(x,t))$ 是以 $p(t|x)$ 为自变量的失真度量函数 $d(x,t)$ 的数学期望.

率失真理论存在的问题是失真度量函数 $d(x,t)$ 的选取缺乏客观性，IB 理论通过定义变量 X 的相关变量 Y，推导出了一个合理的失真度量函数，有效地解决了率失真理论存在的问题. Tishby 在文献[1]中详细介绍了 IB 理论的思想：假设具有相关性的随机变量对 (X,Y) 满足 $(X,Y)\sim p(x,y)$，寻找变量 X 的压缩表示 T，使得 T 尽可能多地保留关于 Y 的相关信息. 这就好象在压缩 X 到 T 的过程中，X 所隐含的关于 Y 的相关信息同样被挤压到 T 中，压缩变量 T 中包含的元素数目即为压缩变量参数. 由于压缩变量参数 $|\mathcal{T}|$ 比之于 $|\mathcal{X}|$ 要少得多，相当于形成了一个细细的“瓶颈”，当将瓶颈一侧的大量数据压缩到该瓶颈中的时候，该数据与另一数据的相关信息同样被压缩到该瓶颈中，这正是 IB 名称的由来. 文献[1]从率失真理论推导出了 IB 理论的数学形式化表示，如式(2)，同时求出了该目标函数的形式解，如式(3)～(5).

$$R(D)=\min_{\{p(t|x):I(T;Y)\geq D'\}} I(X;T) \tag{2}$$

$$p(t|x)=\frac{p(t)}{Z(x,\beta)}\exp(-\beta D_{KL}(p(y|x)\|p(y|t))) \tag{3}$$

$$p(y|t)=\frac{1}{p(t)}\sum_{x\in X}p(y|x)p(t|x)p(x) \tag{4}$$

$$p(t)=\sum_{x\in X}p(t|x)p(x) \tag{5}$$

式(3)中，$Z(x,\beta)=\sum_{t\in T}p(t)\exp(-\beta D_{KL}(p(y|x)\|p(y|t)))$，是一个概率归一化函数. 由式(4)可知：概率 $p(y|t)$ 和 $p(t)$ 都是关于 $p(t|x)$ 的函数. 这表明等号右边的变量同样依赖于等号左边的 $p(t|x)$. 式(3)仅仅是一个形式解，问题的最终求解需要借助上述 3 个等式间形成的迭代算法.

以上 3 个概率分布描述了使用 IB 理论对源数据 X 进行划分得到的结果：概率值 $p(t|x)$ 表示将 x 指派到 t 的概率，该值越大，表示数据 x 被指派到 t 的可能性越大. 概率值 $p(y|t)$ 反应了 t 在相关变量 Y 上的条件概率分布，在 IB 理论中称其为“t 的质心分布”. $p(t)$ 描述了 t 满足的概率分布. IB 得到的数据分析结果有效地描述了变量 X 隐含的与 Y 相关的特征模式. 尽管概率值 $p(t|x)$ 可以取[0,1]区间中的任意值，但本文只关注于“硬”划分：数据分析结果中每个数据元素属于且仅属于一个划分，条件概率 $p(t|x)$ 取值只能为 0 或 1.

3　自动确定参数的 sIB 算法——AsIB

本文将 sIB 算法的压缩变量参数确定作为模型选择问题. 根据最小描述长度原理，分别对 sIB 算法得到的数据分析模型和相应的数据进行编码描述，并将最小描述长度的模型作为算法的选择标准，构建了一种能自动确定参数的 sIB 算法——AsIB.

3.1　sIB 算法

Slonim 等人在文献[2]中将 IB 目标函数变换为等价的式(6)且取平衡参数 $\beta=\infty$，得到 sIB 算法的目标函数 $F=I(T;Y)$. 针对联合概率分布 $p(x,y)$ 和用户输入的压缩变量参数 k，将数据集 X 随机地划分为 k 个子集，在此基础上迭代地将 X 中的每个元素 x 从其所在的 t 中取出，然后将 x 重新分配到满足 $t^{\text{new}}=\arg\min_{t\in T}\cos t(\{x\},t)$ 的 t^{new} 中，这里 $\cos t(\{x\},t)$ 表示将 x 指派到 t 引起的互信息 $I(T;Y)$ 值的减小，其值可以由式(7)计算. 迭代结束后，sIB 算法得到了将数据 X 划分为 k 个子集的数据分析

结果 T.

$$F_{\max} = I(T;Y) - \beta^{-1} I(X;T) \tag{6}$$

$$\cos t(\{x\},t) = (p(x)+p(t)) \times JS_{\pi_1,\pi_2}(p(y \mid x), p(y \mid t)) \tag{7}$$

式(7)中,π_1,π_2 为权值且满足 $\pi_1 = \dfrac{p(x)}{p(x)+p(t)}$, $\pi_2 = \dfrac{p(t)}{p(x)+p(t)}$.

sIB 算法的时间和空间复杂度分别为 $O(\ell|\mathcal{X}||\mathcal{T}||\mathcal{Y}|)$和 $O(|\mathcal{X}|^2)$[2](其中ℓ是一个正的常数),相对于其它 IB 算法,这种具有较低时空开销的算法更适合于实际应用.不仅如此,该算法在迭代结束时,总是保证可以得到局部稳定解.但是,sIB 算法需要用户提供压缩变量数目作为其不可缺少的输入参数,由于实际应用中,该参数往往是未知的,这在一定程度上导致了 sIB 算法应用的局限性.

3.2 最小描述长度原理

对给定数据样本,有可能存在解释这些数据的多个模型.需要一种合适的准则去度量模型的优劣.著名的论断——"奥卡姆剃刀"(Occam's razor),从符合事实的模型中选择最简单的.最小描述长度原理正是这一哲学思想的具体体现.文献[24]指出,给定数据的最优模型是模型本身的描述长度和给定模型下数据的描述长度之和最小的模型,即

$$Model^* = \arg\min_{Model}(DesLen(Model) + DesLen(Data|Model))$$

其中 $DesLen(Model)$是模型的描述长度,$DesLen(Data|Model)$是在给定模型下数据的描述长度.

本文使用最小描述长度原理选择压缩变量参数 k 的取值.k 值较小的模型比较简单,但该模型对源数据的拟合度较差.反之,k 值较大的模型相对复杂但对源数据具有较好的拟合度.模型的复杂度和模型对源数据的适应度之间存在一种平衡关系.根据最小描述长度原理,描述长度最小的模型应该是最优模型.

3.3 AsIB 算法的编码方案

使用最小描述长度原理确定 sIB 算法的最优数据分析模型时,需要设计有效的编码方案计算模型本身的描述长度和给定模型下数据的描述长度.

3.3.1 模型的编码方案

对 sIB 算法的数据分析模型进行编码包括:压缩变量参数 k 的编码和分析结果 T_k 的编码.具体编码方案如下:

①压缩变量参数 k 编码为 $\log^*(k)$, $\log^*$ 表示整数的通用编码长度[24].当事先不知道整数 k 的值域范围时,通用编码长度是最优长度,可以证明 $\log^*(x) \approx \log_2(x) + \log_2 \log_2(x) + \cdots$[25].

②分析结果 T_k 编码为 $\log S_2(n,k)$,其中 $S_2(n,k)$ 表示将 n 个数据划分为 k 个不相交非空集合的所有可能的划分个数,即第 2 类司特林数.

对于特定的输入参数 k,sIB 算法得到的数据分析模型的描述长度为

$DesLen(k) + DesLen(T_k) = \log^*(k) + \log S_2(n,k)$,其中,$DesLen(k)$ 表示编码 k 得到的描述长度;$DesLen(T_k)$表示编码 T_k 得到的描述长度;n 表示数据集 X 中元素的个数$|X|$,对数 log 以 2 为底.

3.3.2 训练数据的编码方案

条件熵 $H(X|Y)$可以作为在已知变量 Y 的条件下编码变量 X 中的一个元素所需要的位数.在已知模型(k,T_k)的条件下对源数据 X 进行编码的描述长度 $DesLen(X|k,T_k)$可以用下式来计算:

$$DesLen(X \mid k,T_k) = n \times H(X \mid T_k) = n \times (H(X) - I(X;T_k)),$$

上式中 $I(X;T_k)$为 X 与 T_k 的互信息;$H(X)$表示 X 的熵;n 表示数据集 X 中元素的个数$|\mathcal{X}|$.

由以上可知,AsIB 算法用于模型选择的总的描述长度 $DesLen$ 如下:

$$DesLen = \log^*(k) + \log S_2(n,k) + n \times (H(X) - I(X;T_k)) \tag{8}$$

3.4 AsIB 算法描述

将最小描述长度原理与 sIB 算法相结合,提出一种自动确定参数的 sIB 算法——AsIB,具体描述如下.

算法 1. 自动确定参数的 AsIB 算法.

输入:联合分布 $p(x,y)$

输出:将 X 划分为 $Kbest$ 个子集的数据分析结果 T_{Kbest}

算法步骤:

1. 根据待分析数据选择参数范围$[Kmin, Kmax]$;
2. 构造空数组 T 用以存放对应于不同参数值的 sIB 算法的解;
3. 构造空数组 $DesLen$ 用以存放对应于不同参数值的描述长度;
4. for all $K \in [Kmin, Kmax]$
5. $T(K) \leftarrow sIB(K)$;
6. $DesLen(K) \leftarrow \log^*(k) + \log S_2(n,K) + n \times (H(X) - I(X;T_k))$;
7. if $DesLen(K) < DesLen(Kbest)$ then
8. $Kbest = K$;
9. end if

10.　end for

11. return $T(Kbest)$；

算法为了节省计算资源，事先指定了参数范围 $[Kmin, Kmax]$（缺省条件下，$Kmin=1$，$Kmax=n$），针对每一个可能的 k 值调用 sIB 算法得到数据分析结果 T_k，然后用式(8)计算总的描述长度. 最终返回具有最小描述长度的模型作为最优输出，这使得模型选择的过程简单，并可得到满足要求的解.

3.5　AsIB 算法时空复杂度分析

本算法第 1～3 步时间复杂度为常数；4～10 步构成一个循环：其中，sIB 算法中将一个数据合并到某个簇中，其时间复杂度为 $O(|\mathcal{T}||\mathcal{Y}|)$，第 5 步 sIB 算法迭代结束时，其时间复杂度为 $O(\ell|\mathcal{X}||\mathcal{T}||\mathcal{Y}|)$，$\ell$ 表示算法收敛需要的迭代次数；第 6～9 步计算描述长度的时间复杂度为常数；考虑整个循环结构，可以得出 AsIB 算法总的时间复杂度为 $O(k\ell|\mathcal{X}||\mathcal{T}||\mathcal{Y}|)$，其中 k 代表算法不同压缩变量参数的数目. 同理，可以分析出 AsIB 算法的空间复杂度为 $O(k|\mathcal{T}|^2)$. 对比于原 sIB 算法，AsIB 算法的时间和空间复杂度多出一乘项 k. 由于实际中 k 取值为有限的整数，AsIB 具有较低的时空复杂度和相当的应用价值.

4　实验与性能分析

4.1　实验数据

实验数据是取自 20-Newsgroup 数据集的 9 个数据子集，其中每个数据集由 500 篇文档组成，它们是随机地从 20-Newsgroup 数据集中一些新闻话题文档中选出的. 对选出的文档使用 Bow 工具包进行预处理[2]，主要包括：去除文本标题；大写字母转换为小写字母；所有的阿拉伯数字合并为单一的数字符号；删除非希腊数字(non-alpha-numeric)符号、停用词(stop-words)和仅出现一次的单词；选择互信息最大的 2000 个单词作为文档的相关变量集. 表 1 给出了按以上步骤抽取出的每个数据集的细节.

抽出的 9 个文本数据集是“按类别有序的”. 例如，Binary 数据集中的 500 个文档，前 250 个文档来源于新闻类别 talk. politics. mideast，而后 250 个均来自于 talk. politics. misc. 其它数据集具有类似的特性. 基于该数据集的有序性问题，本文对源数据集中的文档进行随机组合得到了 9 个无序的数据集：S-Binary_1，2，3，S-Multi5_1，2，3 和 S-Multi10_1，2，3，并且进行了相应的对比实验.

表 1　IB 算法实验的数据集

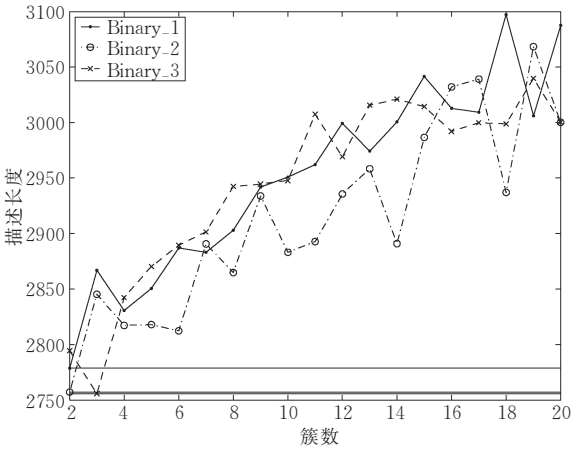
数据集名称	真实类别数	文档所源自的新闻话题类别	每个类别的文档数目	总规模
Binary_1，2，3	2	talk. politics. mideast， talk. politics. misc	250	500
Multi5_1，2，3	5	comp. graphics，rec. motorcycles，rec. sport. baseball， sci. space，talk. politics. mideast	100	500
Multi10_1，2，3	10	alt. atheism，comp. sys. mac. hardware，misc. forsale， rec. autos，rec. sport. hockey，sci. crypt，sci. electronics， sci. med，sci. space，talk. politics. guns	50	500

4.2　实验方法与结果

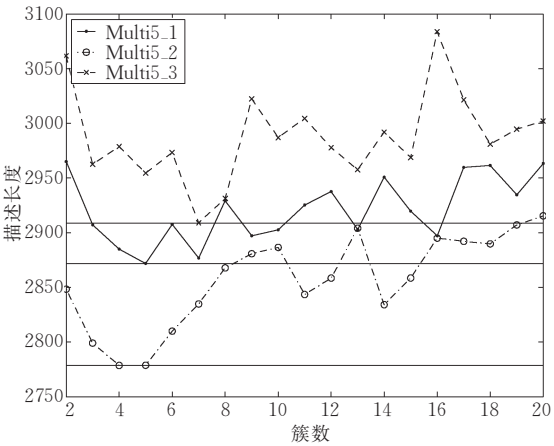
研究 IB 算法的实验数据集是“相关”数据集：分析对象和其相关变量内在隐含着某种相关性. 在具有这种特性的数据集上，传统的模式分析方法很难有效地发现特征模式，而 IB 算法则适合于处理这类问题. 由于实验数据是源于 20-Newsgroup 的文档-单词相关数据集，本文的算法分析结果可以认为是：已知单词相关变量的条件下对文档聚类的结果. 本文将针对相应实验数据集测试 AsIB 算法的有效性.

4.2.1　有序数据集的实验

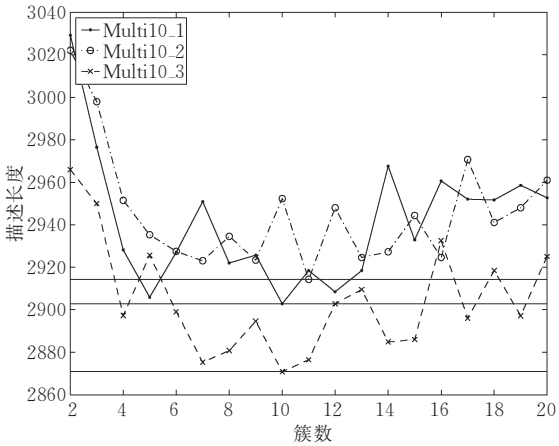
设置 $Kmin=2$ 且 $Kmax=20$，调用 AsIB 算法自动发现每个数据集的文档类别数并对文档进行聚类. 图 1 显示了 AsIB 算法在 9 个有序的数据集上对应于不同簇数目的描述长度. 图 1(a)表明：描述长度随着簇数目的增加而逐渐增大；数据集 Binary_1 和 Binary_2 在 $k=2$ 时描述长度最小；在数据集 Binary_3 上算法分析得到的类别数为 3，接近于真实类别数 2. 图 1(b)和图 1(c)给出了类似的结果，由此可知：AsIB 算法在数据集 Multi5_1、Multi10_1 和 Multi10_3 上正确无误地发现文档类别数，在数据集 Multi5_2、Multi5_3 和 Multi10_2 上发现的类别数与文档真实类别数差异很小. 表 2 给出了 AsIB 算法得到的结果：能够正确无误地发现 9 个数据集中 5 个数据集的类别数，3 个数据集上发现的类别数与真实类别数差异仅为 1 个左右. 由此可见，AsIB 算法在自动发现具有相关性的文档数据集隐含的类别数时是有效的.



(a) Binary



(b) Multi5



(c) Multi10

图 1 有序数据集上描述长度与簇数的关系

表 2 AsIB 算法在有序数据集上的实验结果

数据集	发现的类别数	真实类别数	差异值
Binary_1	2	2	0
Binary_2	2	2	0
Binary_3	3	2	1
Multi5_1	5	5	0
Multi5_2	4	5	1
Multi5_3	7	5	2
Multi10_1	10	10	0
Multi10_2	11	10	1
Multi10_3	10	10	0

4.2.2 随机数据集的实验

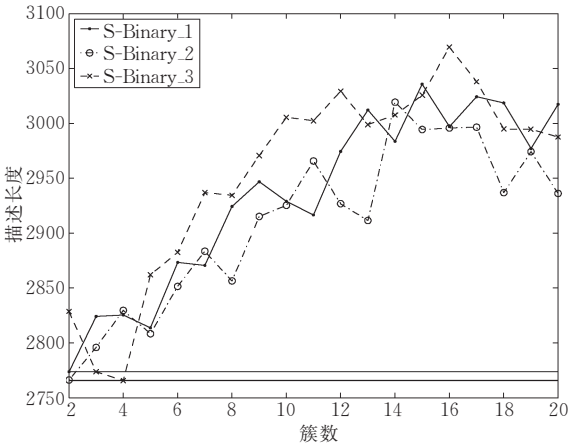
实验具体设置完全类似于 4.2.1 节,图 2 显示了 AsIB 算法在随机的 9 个数据集上对应于不同簇数的描述长度. 图 2(a)表明:描述长度随着簇数的增加而逐渐增大;数据集 S-Binary_1 和 S-Binary_2 在 $k=2$ 时描述长度最小;在数据集 S-Binary_3 上算法得到 $k=4$,与 S-Binary_3 中真实类别数差异为 2. 图 2(b)和图 2(c)表明:AsIB 算法在数据集 S-Multi5_1、S-Multi5_2、S-Multi10_1 和 S-Multi10_3 上正确无误地发现了文档类别数;在数据集 S-Multi5_3 和 S-Multi10_2 上发现的类别数与文档类别数差异很小. 表 3 给出的 AsIB 算法得到的结果与 4.2.1 节中的实验结果接近,从而进一步验证了 AsIB 算法的有效性.

表 3 AsIB 算法在随机数据集上的实验结果

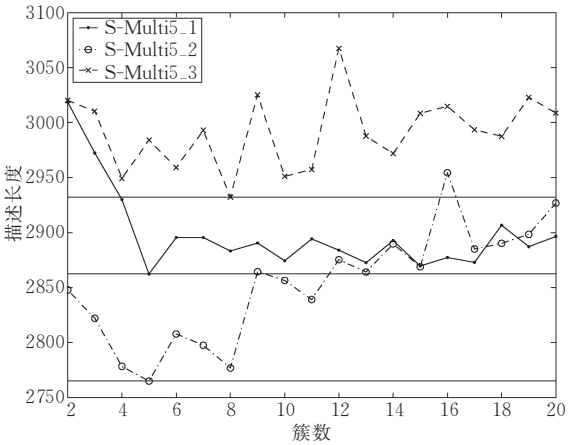
数据集	发现的类别数	真实类别数	差异值
S-Binary_1	2	2	0
S-Binary_2	2	2	0
S-Binary_3	4	2	2
S-Multi5_1	5	5	0
S-Multi5_2	5	5	0
S-Multi5_3	8	5	3
S-Multi10_1	10	10	0
S-Multi10_2	11	10	1
S-Multi10_3	10	10	0

4.2.3 增加 sIB 算法调用次数的实验

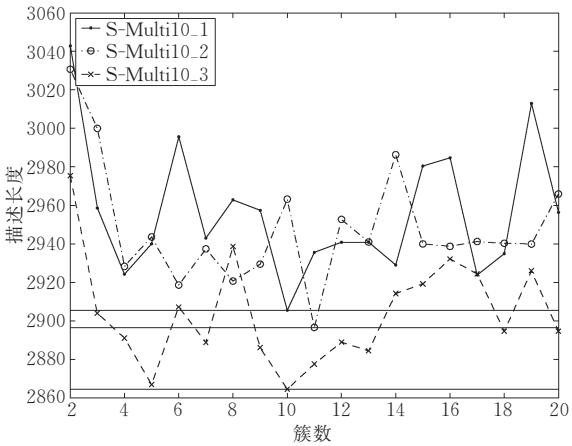
为了避免 sIB 算法的局部解对 AsIB 算法正确发现特征模式数目的影响,实验针对 AsIB 算法没有正确发现特征模式数的有序和随机数据集,将 sIB 算法的运行次数增大一倍,进一步测试 AsIB 算法的性能. 实验的结果如图 3 和表 4 所示. 其中:在 Multi5_2、Multi10_2、S-Multi10_2 上正确地发现了



(a) S-Binary

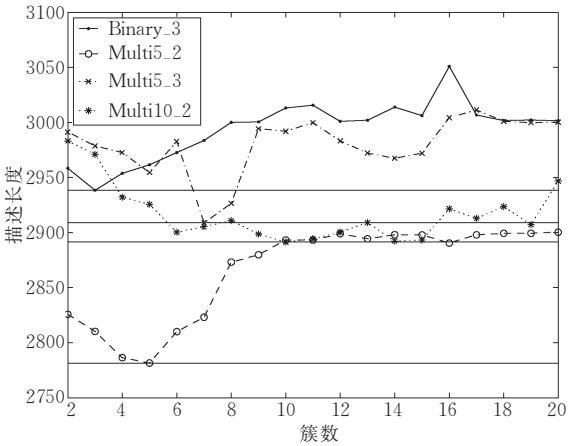


(b) S-Multi5

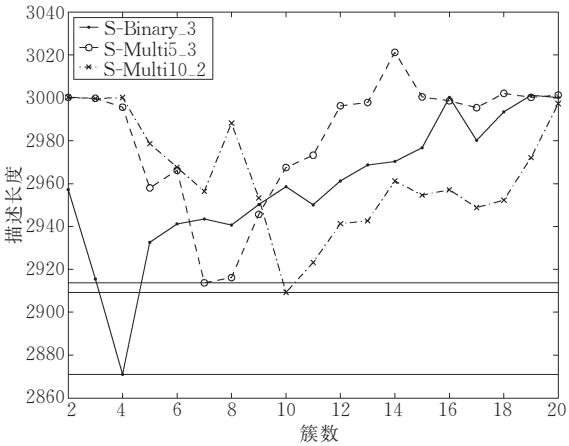


(c) S-Multi10

图 2　随机数据集上描述长度与簇数的关系



(a) 有序数据集



(b) 随机数据集

图 3　增大 sIB 算法运行次数时描述长度与簇数的关系

表 4　增大调用次数时针对发生错误的数据集的实验结果

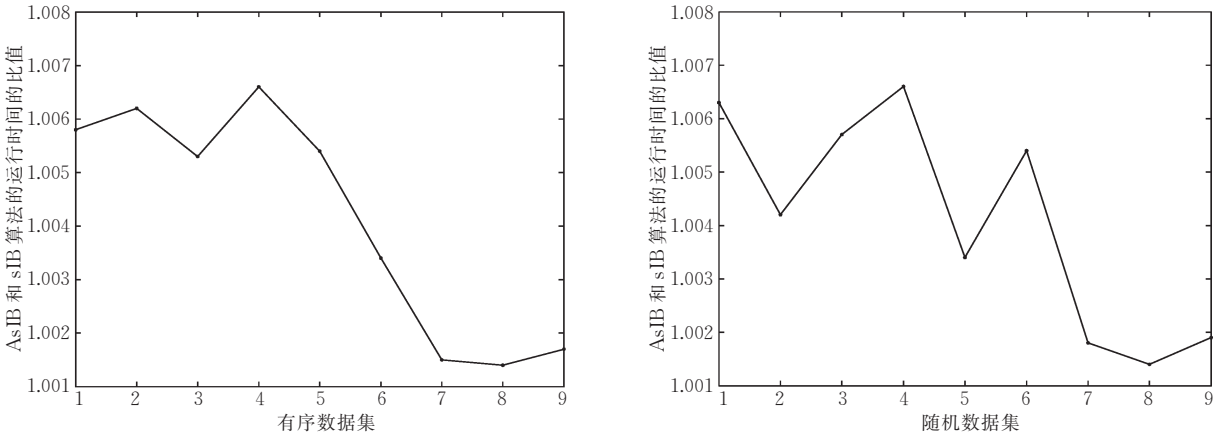
数据集	发现的类别数	真实类别数	差异值
Binary_3	3	2	1
S-Binary_3	4	2	2
Multi5_2	5	5	0
Multi5_3	7	5	2
S-Multi5_3	7	5	2
Multi10_2	10	10	0
S-Multi10_2	10	10	0

真实类别数；从实验结果中发现，没有能正确发现真实模式类别数目的数据集是相同的数据集：Multi5_3、S-Multi5_3、Binary_3、S-Binary_3.

4.2.4　AsIB 算法与 sIB 算法时间对比

在 sIB 算法调用次数相同的情况下，针对以上 18 个数据集进行 AsIB 算法和 sIB 算法运行时间的对比分析. 图 4 中的横轴表示数据集，纵轴表示

AsIB 算法与 sIB 算法运行时间的比值. 两组实验结果表明：AsIB 算法与 sIB 算法运行时间比值最大为 1.0066，说明在基本 sIB 算法调用次数相同的情况下，两者的运行时间几乎相等.



(a) 有序数据集 (b) 随机数据集

图 4 AsIB 算法与 sIB 算法运行时间比值

4.3 实验结果分析

从以上实验结果可知：

(1) 在 18 个数据集上，AsIB 算法正确无误地在 11 个数据集上发现了相应的文档类别数，在剩余的 7 个数据集上发现的类别数与真实文档类别数差异很小，充分表明算法的有效性.

(2) AsIB 算法在有序和随机数据集上得到的结果较为接近，该特性表明 AsIB 算法具有良好的鲁棒性.

(3) 增加 sIB 算法的调用次数可以有效避免该算法陷入局部解，有利于 AsIB 算法正确发现特征模式数目. 当 sIB 算法的调用次数增大一倍时，AsIB 算法在 14 个数据集上正确无误地确定出相应的特征模式数目. AsIB 算法发现相应数据集类别数目的平均误差为 0.389 个.

(4) 当 AsIB 算法不能完全正确地发现类别数目时，发生错误的数据集是相同的数据集. 这表明 AsIB 算法错误判定类别数目的原因可能是由数据集自身隐含的“噪声”所导致. 该问题有待于进一步的研究.

(5) 在基本 sIB 算法调用次数相同的情况下，AsIB 算法与 sIB 算法的运行时间几乎相等. 但这是以增加基本 sIB 算法调用次数为前提的，如何在减少基本 sIB 算法调用次数的情况下增强 AsIB 算法的时间优势是我们下一步的工作之一.

5 结论及下一步的工作

sIB 算法是现存的效率较高、可以从复杂数据中提取相关特征模式的 IB 算法之一. 但与压缩变量参数相对应的特征模式数目需要作为先验知识提供给该算法作为输入参数，这在一定程度上影响了其在实际中的应用. 本文将 sIB 算法的参数确定作为模型选择问题，根据最小描述长度原理，分别对 sIB 算法得到的数据分析模型和相应的数据进行编码描述，并将最小描述长度的模型作为算法的选择标准. 在此基础上，构建了一种自动确定参数的算法——AsIB. 在研究 IB 算法的公共数据集上的实验结果表明：AsIB 算法所采用的编码方案有效，在不设定模式数目的情况下，能够正确发现数据集所蕴含的模式. 本文的主要贡献可以概括如下：

(1) 将参数确定机制与 sIB 算法进行有机结合，构建了一种自动的 AsIB 算法，解决了 sIB 算法有效性的基本问题.

(2) 设计了一种有效的数据和模型编码方案，用其计算出的描述长度能够较好地刻画相应数据的模式特征.

(3) 成功解决了 sIB 算法输入参数的确定问题. 在不需要任何先验知识的条件下，能够以相当高的精度自动地正确确定数据集中包含的特征模式数.

AsIB 算法应用于非相关数据集的研究是我们下一阶段的主要工作.

致 谢 在 IB 理论与各种算法的研究过程中，美国 Princeton 大学的 Slonim 博士为我们提供了帮助，在此对 Slonim 先生表示感谢. 感谢朱真峰和张洁在算法实验实施方面给予的帮助！

参 考 文 献

[1] Tishby N, Pereira F, Bialek W. The information bottleneck method//Proceedings of the 37th Allerton Conference on Communication, Control and Computing. Illinois, USA, 1999: 368-377

[2] Slonim N, Friedman N, Tishby N. Unsupervised document classification using sequential information maximization//Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland, 2002: 129-136

[3] Slonim N, Tishby N. Document clustering using word clusters via the information bottleneck method//Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Athens, Greece, 2000: 208-215

[4] Slonim N, Tishby N. The power of word clusters for text classification//Proceedings of the 23rd European Collquium on Information Retrieval Research. Darmstadt, Germany, 2001: 1-12

[5] Goldberger J, Gordon S, Greenspan H. Unsupervised image-set clustering using an information theoretic framework. IEEE Transactions on Image Processing, 2006, 15(2): 449-458

[6] Winston H H, Shih C F. Visual cue cluster construction via information bottleneck principle and kernel density estimation//Proceedings of the 4th International Conference Image and Video Retrieva. Singapore, 2005: 82-91

[7] Slonim N, Somerville R, Tishby N, Lahav O. Objective classification of galaxies spectra using the information bottleneck method. Monthly Notices of the Royal Astronomical Society, 2001, 323(2): 270-284

[8] Tishby N, Slonim N. Data clustering by Markovian relaxation and the information bottleneck method//Proceedings of the 13th Annual Conference on Neural Information Processing Systems. Colorado, USA, 2001: 640-646

[9] Schneidman E, Bialek W, Berry M J. An information theoretic approach to the functional classification of neurons//Proceedings of the 15th Annual Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada, 2002: 197-204

[10] Gorodetsky M. Methods for discovering semantic relations between words based on co-occurrence patterns in corpora [Masters dissertation]. School of Computer Science and Engineering, Hebrew University, Jerusalem, 2002

[11] Hecht R, Tishby N. Extraction of relevant speech features using the information bottleneck method//Proceedings of the InterSpeech. Lisbon, Portugal, 2005

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

[13] Chechik G, Tishby N. Extracting relevant structures with side information//Proceedings of the 16th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2002: 857-864

[14] Gondek D, Hofmann T. Non-redundant data clustering//Proceedings of the 4th IEEE International Conference on Data Mining. Brighton, UK, 2004: 75-82

[15] Friedman N, Mosenzon O, Slonim N, Tishby N. Multivariate information bottleneck//Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence. Seattle, Washington, USA, 2001: 152-161

[16] Elidan G, Friedman N. Learning hidden variable networks: The information bottleneck approach. Journal of Machine Learning Research, 2005, (6): 81-127

[17] Chechik G, Globerson A, Tishby N, Weiss Y. Information bottleneck for gaussian variables. Journal of Machine Learning Research, 2005, (6): 165-188

[18] Cloberson A, Tishby N. Sufficient dimensionality reduction. Journal of Machine Learning Research, 2003, (3): 1307-1331

[19] Slonim N, Tishby N. Agglomerative information bottleneck//Proceedings of the 13th Annual Conference on Neural Information Processing Systems. Colorado, USA, 1999: 617-623

[20] Peltonen J, Sinkkonen J, Kaski S. Sequential information bottleneck for finite data//Proceedings of the 21st International Conference on Machine Learning. Madison, USA, 2004: 647-654

[21] Still S, Bialek W. How many clusters? An information-theoretic perspective. Neural Computation, 2004, 16(12): 2483-2506

[22] Niu Z Y, Ji D H, Tan C L. Document clustering based on custer validation//Proceedings of the 11st International Conference on Information and Knowledge Management. Virginia, USA, 2004: 501-506

[23] Rissanen J. Modeling by shortest data description. Automatica, 1978, (14): 465-471

[24] Barron A, Rissanen J, Yu B. The minimum description length principle in coding and modeling. IEEE Transactions on Information Theory, 1998, 44(6): 2737-2760

[25] Rissanen J. Universal prior for integers and estimation by minimum description length. Annals of Statistics, 1983, 11(2): 416-431

YE Yang-Dong, born in 1962, professor, Ph. D. supervisor. His current research interests include knowledge engineering, intellectual system, database system.

LIU Dong, born in 1981, M. S.. His current research interests include machine learning and data mining.

JIA Li-Min, born in 1963, Ph. D., professor, Ph. D. supervisor. His current research interests include computational intelligence, intelligent control and automation.

LI Gang, born in 1973, Ph. D., lecturer. His current research interests include machine learning and data mining.

Background

This work is supported by the National Natural Science Foundation of China under grant No. 600332020 with title "Research on Intelligent Synthetical transportation Information System of High Railway and its Key Technologies", and the Natural Science Foundation of Henan province under grant No. 0411012300 titled "Research on Application of Theories of Modeling of Hybrid System based on Distributed Agent". This paper aims to solve the problem of the analysis of multi-source and multi-dimension data and focuses on the methods and technologies for analysis of huge data using IB theory. The research has achieved certain results in areas such as improving the accuracy of IB algorithms and introducing IB into image retrieval. By compressing one variable to the "Bottleneck" variable and maximally preserving its relevance to another variable, IB theory effectively settles the problems that many other classical feature extracting methods can't solve. Along with the embedded applications of IB theory in many areas, the shortcomings of IB algorithms in the case of enlarging and changing of searching spaces have emerged. One of these problems can be summarized as: Determining the parameter of compression variable for IB algorithms. In order to determine the compression variable parameter, this paper proposes an sIB algorithm for automatically determining parameter based on minimum description length principle. An efficient encoding scheme is designed to estimate the description length of the solution model of sIB algorithm and the original data given the model respectively, and the minimum description length model is selected as a criterion to find the number of feature patterns hidden in dataset. The proposed AsIB algorithm removes the dependency of empirical knowledge for sIB algorithm, and the research results could widen its applications in areas such as automatic dimension reduction and pattern extraction, etc.