基于数据选择模型的 IB 算法

娄铮铮,杨 晨,叶阳东
(郑州大学信息工程学院,河南郑州 450052)

摘 要: 针对数据对象自身模式特征明确程度的不同给 IB(Information Bottleneck)方法数据分析带来的问题,定义一个"基于明确因素"的数据选择模型,使得 IB 方法可从数据集中选取模式特征较为明确的数据对象并对其进行模式分析,提出 DSIB (Data Selection Information Bottleneck)算法.DSIB 算法采用数据压缩过程中所产生的信息损失作为数据对象模式特征是否明确的判定条件,使用"边选择边学习"的顺序"抽取-合并"策略来优化 DSIB 目标函数.实验结果表明:随着数据选择标准的不断提高,DSIB 算法在提高数据分析精度的同时所牺牲的召回率较小;与未做选择的数据分析算法相比,DSIB 算法可更好地识别出数据中所固有的内在模式.

关键词: IB 方法; 数据选择; 簇; 模式特征
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2014)09-1839-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.09.027

An IB Algorithm Based on Data Selection Model

LOU Zheng-zheng, YANG Chen, YE Yang-dong
(*School of Information Engineering, Zhengzhou University, Zhengzhou, Henan* 450052, *China*)

Abstract: In the original IB (Information Bottleneck) algorithms, all the data points are employed to learn the cluster patterns. However, in many real-world applications, some data show clear coherent behavior and can be summarized well, while some data present weak tendencies to be assigned to any particular pattern. For such situations, this paper proposes a DSIB (Data Selection Information Bottleneck) algorithm which has the ability to select data points with clear coherent behavior and find their corresponding cluster patterns. To realize this goal, the DSIB algorithm takes the information loss as the data selection criterion, which is generated when we try to compress the data point into one of the clusters. The DSIB algorithm adopts sequential "draw-and-merge" procedure to select the data and learn the cluster patterns. This learning process can take full account of each datum's natural pattern. Experimental results show that with the improvement of the data selection criterion, the DSIB algorithm can improve the clustering precision while the expense of the recall is small. In our evaluation, the DSIB algorithm is found to be consistently superior to all the other clustering methods we examine.

Key words: information bottleneck (IB) method; data selection; cluster; patterns

1 引言

IB 方法(Information Bottleneck method)[1]是 Tishby 等人于 1999 年提出的一种基于信息论的数据分析方法.该方法在做数据分析时,通过将数据对象压缩到一个"瓶颈"变量中,同时最大化地保存数据中的信息量,可有效地发现数据对象间所蕴含的内在模式.IB 方法已在众多领域中取得成功的应用,其中包括:文档数据分析[2~4]、图像模式识别[5~7]、语音识别[8]、社区发现[9]、监督量化码本学习[10]、范畴数据分析[11]等.

目前,针对 IB 方法所设计的优化算法有 aIB 算法[2]、sIB 算法[3]、CD-sIB 算法[11]、GsIB 算法[12]、isIB 算法[13]、DaIB 算法[14]、DsIB[14]算法等.这些算法在对数据进行分析时,每一个数据对象都被强制地指派到一个相应的簇中.然而在实际应用中,有些数据对象与数据中所蕴含的某一模式的拟合程度较强,自身模式特征较为突出;而有些数据对象与数据中所有内在模式的拟合度均不强,自身模式特征较弱.在应用中,若不考虑数据对象自身模式特征的明确程度,强制地将所有数据对象指派到某些簇中时,由于模式特征较弱的数据对象与数据内在结构之间的拟合度不强,它们的加入将在一定程度上使簇模式偏离数据所固有的内在模式,从而影响 IB 方法数据分析的性能.

收稿日期:2013-05-13;修回日期:2013-12-21;责任编辑:覃怀银
基金项目:国家自然科学基金(No.61170223)

本文通过定义一个“基于明确因素”的数据选择模型 DSM(Data Selection Model),将数据对象模式特征是否明确这一因素融入到 IB 方法中,使得 IB 方法有选择的对数据进行模式分析,提出 DSIB 算法(Data Selection Information Bottleneck).与传统的无监督数据分析算法相比,如 k-means、DBSDAN[15]、Normalized Cuts[16]、sIB[3]、Affinity Propagation[17],DSIB 算法允许对模式特征较弱的数据对象不做任何簇指派,从而可获取精度更高的簇.这种特点在购物篮数据分析、web 挖掘、生物信息分析等领域均具有重要的应用价值[18,19].

文献[3]为消弱模式特征较弱的数据对象给数据分析带来的负面影响,将数据对象与其所在簇质心之间的距离作为模式特征是否明确的判定标准,进而实现对数据的选择分析.该方法是一个“先学习再选择”的过程,通过 IB 算法先学习一个模式划分,再对数据进行选择.其结果易受原 IB 算法所得簇划分质量的影响.MCRD 算法[20]采用一个“先选择再学习”的过程来实现对数据的选择分析,首先采用基于 IB 方法的单类学习算法 OCRD[20]从数据中选取一部分数据对象,然后再采用 IB 算法[3]对所选取的数据对象进行模式划分.其结果易受 OCRD 算法的数据选择及 IB 算法簇结构的学习两过程的影响.上述两种方法的簇模式学习过程与数据选择过程相对分离,簇模式的学习并未充分考虑每一个数据对象模式特征的明确程度.

DSIB 算法采用数据压缩过程中所产生的信息损失作为数据对象模式特征是否明确的判定条件,通过顺序“抽取-合并”的优化策略来确保簇结构的学习过程与数据的选择过程同时实施,是一个“边选择边学习”的数据分析算法.在 20-Newsgroup 和 Reuters21578 文档数据集上的实验结果表明:具有“边选择边学习”特点的 DSIB 算法的数据选择分析的性能优于文献[3]及文献[20]中所提出的数据选择分析方法的性能;另外,通过对数据的选择分析,DSIB 算法的数据分析性能优于未做数据选择的 IB 算法[3]、k-means 算法及 Normalized Cuts 算法[16].

本文主要贡献可总结如下:

(1) 针对数据自身特征明确程度的不同,定义一个数据选择模型,使得 IB 方法有选择的对数据进行模式分析;

(2) 采用数据压缩过程中所产生的信息损失作为数据对象模式特征是否明确的判定条件,提出基于数据选择模型的 DSIB 目标函数;

(3) 提出一个顺序的 DSIB 算法来优化 DSIB 目标函数,该算法为一个“边选择边学习”的优化过程.

2 IB 方法

IB 方法在做数据分析时,采用变量 X 表示待分析的数据对象 $X=\{x_1,x_2,\cdots,x_n\}$,Y 表示描述数据对象的特征变量,其取值域为 $Y=\{y_1,y_2,\cdots,y_m\}$.假设 n_{ij} 为特征 y_j 在数据对象 x_i 中出现的次数(例如,某一单词在文档中出现的次数),则 X 和 Y 之间的联合分布 $p(x_i,y_j)=\frac{n_{ij}}{\sum_{i'}\sum_{j'}n_{i'j'}}$.令 $T=\{t_1,t_2,\cdots,t_l\}$ 为压缩“瓶颈”,由变量 T 表示.

给定变量 X 与 Y 之间的联合分布,IB 方法在做数据分析时,将数据模式的提取视为一个数据压缩的过程,即将变量 X 压缩到“瓶颈”变量 T 中,与此同时,使 T 最大化的保存特征 Y 中所蕴含的信息量,其中 X 到 T 的编码方案 $p(t|x)$ 揭示了待分析数据对象间所蕴含的内在模式.IB 方法可形式化的描述为:

$$R(D)\equiv\min_{|p(T|X):I(T;Y)\geqslant D|}I(X;T) \tag{1}$$

其中 $I(X;T)=\sum_x\sum_t p(x,t)\log\frac{p(x,t)}{p(x)p(t)}$ 为互信息[21].从式(1)中可以看出,IB 方法是在满足信息保存限制的条件下,即“瓶颈”变量 T 中所保存的信息量 $I(T;Y)$ 应满足 $I(T;Y)\geqslant D$,在所有可能的编码方案中选择使压缩信息 $I(T;X)$ 最小化的一个编码方案 $p(T|X)$.

为求解最优压缩编码方案 $p(T|X)$,文献[1]采用拉格朗日乘数法,将式(1)改写为如下的 IB 目标函数:

$$F_{\min}(p(T|X))=I(X;T)-\beta I(T;Y) \tag{2}$$

其中,$\beta\in[0,\infty)$ 是拉格朗日因子,用于平衡信息源的压缩和相关信息的保存.本文在做数据分析时,仅考虑“硬”聚类,即 $p(t|x)$ 取值仅为 0 或 1,β 取值为 ∞,将 IB 方法的重点放在相关信息的保存上.此时 IB 方法目标函数可重写为[4]:

$$F_{\max}(p(T|X))=I(T;Y) \tag{3}$$

3 基于数据选择模型的 IB 算法

3.1 数据选择模型

针对数据对象自身模式特征明确程度的不同给数据分析带来的问题,本文提出一个数据选择模型 DSM 来指导 IB 方法对数据进行选择分析.其定义如下:

定义 1 数据选择模型 DSM 是一个 5 元组 (X,X',T,s,f),其中

(1) $X=\{x_1,x_2,\cdots,x_n\}$ 为待分析的数据对象集合;

(2) X' 是 X 的数据子集;

(3) $T=\{t_1,t_2\cdots t_l\}$ 代表数据分析过程中所得到的簇;

(4) s 为一选择函数,$X'=s(X;T)$;

(5) f 为一学习函数,$T=f(X')$.

图 1 为数据选择模型的示意图.数据选择模型中的

s 函数为数据选择函数,该函数通过对比数据对象与簇结构之间的拟合程度来判定数据对象的簇结构特征是否明确,从而作出对数据的选择.学习函数 f 则根据相应的目标将选择出的数据对象指派到一个合适的簇中,从而学习出数据中所蕴含的内在结构.数据的选择和簇结构的学习是一个循环的过程,直到所学习到的簇结构达到一个稳定状态为止.

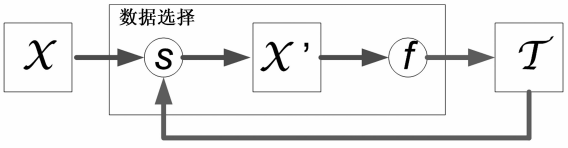

图1　数据选择模型DSM示意图

3.2　基于数据选择模型的 IB 算法

本节在数据选择模型的基础上,提出一个具有选择性分析能力的 IB 算法:DSIB 算法.该算法在对数据进行模式分析时,仅从数据中选取簇结构特征较为明确的数据对象并从中学习簇的模式结构,以便改善 IB 方法数据分析的性能.

3.2.1　DSIB 目标函数

从信息论角度来讲,模式特征较为明确的数据对象自身的归属性较强,当 IB 方法将其压缩到其所应归属的某一个簇中时,所产生的信息损失相对较小;而模式特征较弱的数据对象自身的不确定较强,在将其压缩到任何一个簇中时,会产生较大的信息损失.基于此,我们将数据对象分配到某一簇中所产生的信息损失作为数据选择的判定条件,给出如下的基于数据选择模型的 DSIB 目标函数:

$$\max_{p(T|X)} I(T;Y) \tag{4a}$$

$$\text{s.t.}\Delta I(x,t)<\lambda,\text{if } p(t|x)=1 \tag{4b}$$

其中 $p(t|x)=1$ 表示数据对象 x 被压缩到簇 t 中.在满足式(4b)的约束条件情况下,DSIB 在寻求 X 到 T 的最优压缩表示 $p(T|X)$时,在式(4a)中力图使 T 最大化的保存特征变量 Y 中所蕴含的信息量.式(4a)对应于数据选择模型中的 f 函数.在式(4b)中,若 x 被压缩到簇 t 中,则该压缩过程所产生的信息损失 $\Delta I(x,t)$必须小于阈值 λ.该式对簇中的每一个数据对象做了约束,从而实现数据的选择.若将数据对象 x 压缩到所有的簇 t 中时,所产生的信息损失 $\Delta I(x,t)$均不小于阈值 λ,则该数据对象便被视为簇结构特征不明确的数据对象,不对其进行压缩.约束条件(4b)即为数据选择的判定标准,对应于数据选择模型中的数据选择函数 s,其所选取的数据子集 $X'=\{x|\exists\, t\in T:p(t|x)=1\}$.

3.2.2　信息损失

DSIB 算法采用数据压缩过程中所产生的信息损失作为数据对象模式特征是否明确的判定条件,从而实现对数据的选择分析.假设数据对象 x 在没有被压缩到某一簇中时数据中的信息量为 I_x^{bef},数据对象 x 被压缩到某一簇 t 中形成新簇 t'时数据中的信息量为 I_x^{aft},则该压缩过程所产生的信息损失 $\Delta I(x,t)$可计算为:

$$\Delta I(x,t)=I_x^{\text{bef}}-I_x^{\text{aft}} \tag{5}$$

下面我们给出欲将某一数据对象 x 指派到某一簇 t 中形成一个新簇 t'时所产生的信息损失的计算方法.首先,$p(t')$和 $p(y|t')$可计算如下:

$$p(t')=p(x)+p(t) \tag{6}$$

$$p(y|t')=\frac{p(x)}{p(t')}p(y|x)+\frac{p(t)}{p(t')}p(y|t) \tag{7}$$

当 x 合并到簇 t 并形成簇 t'时,$\Delta I(x,t)$为:

$$\begin{aligned}\Delta I(x,t)&=I_x^{\text{bef}}-I_x^{\text{aft}}=I(T^{\text{bef}};Y)-I(T^{\text{aft}};Y)\\&=p(t)\sum_y p(y|t)\log\frac{p(y|t)}{p(y)}\\&\quad+p(x)\sum_y p(y|x)\log\frac{p(y|x)}{p(y)}\\&\quad-p(t')\sum_y p(y|t')\log\frac{p(y|t')}{p(y)}\end{aligned}$$

将式(6)和(7)代入上式得:

$$\begin{aligned}\Delta I(x,t)=&\ p(t)\sum_y p(y|t)\log\frac{p(y|t)}{p(y)}\\&+p(x)\sum_y p(y|x)\log\frac{p(y|x)}{p(y)}\\&-\sum_y p(x)p(y|x)\log\frac{p(y|t')}{p(y)}\\&-\sum_y p(t)p(y|t)\log\frac{p(y|t')}{p(y)}\\=&\ p(x)\sum_y p(y|x)\log\frac{p(y|x)}{p(y|t')}\\=&+p(t)\sum_y p(y|t)\log\frac{p(y|t)}{p(y|t')}\end{aligned}$$

由文献[21]中的 JS 距离的定义可得:

$$\Delta I(x,t)=p(t')\cdot\text{JS}_{\pi_1,\pi_2}(p(Y|x),p(Y|t)) \tag{8}$$

其中,$\pi_1=\dfrac{p(x)}{p(t')}$,$\pi_2=\dfrac{p(t)}{p(t')}$.

3.2.3　DSIB 算法

为实现数据的选择和簇结构的学习两过程的同时实施,本文采用顺序“抽取-合并”的过程来优化 DSIB 目标函数(4).其基本思想为:顺序的分析每一个数据对象 x,若数据对象 x 已被指派到某一簇 t 中时,即 $p(t|x)=1$,则将该数据对象从当前簇 t 中抽取出来;然后再根据公式(8)计算欲将该数据对象压缩到每一个簇中所产生的信息损失 $\Delta I(x,t)$;若存在某些簇,满足 $\Delta I(x,t)<\lambda$,则该数据对象被视为模式特征较为明确的数据对象,并将该数据对象“合并”到簇 $\hat{t}$ 中,其中 $\hat{t}=\arg\min_{t\in T}\Delta I(x,t)$;否则,该数据对象被视为模式特征不明确的数据对象,不对其进行压缩.DSIB 算法的具体实施过程如算

法 1 所示.从该算法中我们可以看出,算法的第 6 ~ 8 步计算了数据对象到每一个簇特征的明确程度,第 9 ~ 12 步骤为数据的选择和簇结构的学习过程,只有当 $\min(\Delta I(x,t)) < \gamma$ 时才将数据对象合并到相应的簇中.因此,在 DSIB 算法中数据的选择和簇结构的学习是同时进行的.

算法 1 DSIB 算法

输入:数据集 $X=\{x_1,x_2,\cdots,x_n\}$;簇的个数 $l=|T|$;参数 λ
输出:X 到 T 的压缩表示 $p(T|X)$
初始化:$p(t|x)$←X 到 T 的随机初始划分;
主循环:
1. Repeat
2. For every $x\in X$
3. If $\exists t, p(t|x)=1$
4. 把 x 从当前簇 t 中抽取出来,令 $p(t|x)=0$;
5. End
6. For every $t\in T$
7. 根据式(8)计算 $\Delta I(x,t)$;
8. End
9. If $\min(\Delta I(x,t))<\gamma$
10. 将 x 合并到簇 $\hat{t}$ 中,其中 $\hat{t}=\arg\min_{t\in T}\Delta I(x,t)$,令 $p(\hat{t}|x)=1$;
11. 根据式(6)和(7)更新簇 $\hat{t}$ 的相关信息;
12. End
13. End
14. Until $p(t|x)$不再发生变化为止

在 DSIB 算法中,参数 λ 是数据对象模式特征是否明确的判定阈值.该参数取值越小,DSIB 算法对数据对象模式特征的明确程度要求越高,所选择的数据对象也越少.

3.3 时间复杂度分析

在 DSIB 算法的主循环中,步骤 3 ~ 5 为数据抽取的过程,其时间复杂度为 $O(1)$;步骤 6 ~ 8 为信息损失的计算,其时间复杂度为 $O(|Y|)$;步骤 9 ~ 12 为数据选择及合并的过程,时间复杂度为 $O(|Y|)$.因此,整个 DSIB 算法的时间复杂度为 $O(k|X||Y|)$,其中 k 为算法停止时所迭代的次数.在实验部分 4.4.3 节中我们将看到,通过数十次的迭代循环,DSIB 算法中的 $p(t|x)$便不再发生任何改变,算法收敛到目标函数的一个局部优化解.可见,DSIB 算法的时间复杂度与数据集的规模线性相关.

4 实验与性能分析

4.1 实验数据集

本文采用 20-Newsgroup 文档数据集的 9 个子数据集、Reuters-21578 文档数据集的 4 个子数据集,共计 13 个数据集来评估 DSIB 算法数据分析的性能.这些数据集的相关信息如表 1 所示.20-Newsgroup 文档数据集的 9 个子数据集的选取方法如文献[3]所示.针对 Reuters-21578 数据集,去除含有多个类标签的文档,从中选取规模最大的 10 个类别,共得到 7285 篇文档.由于最大两个类别分别含有 3713 篇文档与 2055 篇文档,而其它 8 个类中,规模最大的一个类仅包含 321 篇文档,因此这 10 个类极其不平衡.在此,我们将该 10 个类拆分为 4 个子数据集,其中 Reuters2 是包含规模最大的两个类别的数据子集,Reuters8 为剩下的 8 个类别所组成的一个数据集,然后再将这 8 个类别拆分为 2 个子数据集,每个包含四个类别,形成 Reuters4_1、Reuters4_2 两个数据集.

针对 20-Newsgroup 数据集,从所有单词中选择对共现矩阵互信息贡献度最大的 2000 个单词作为文档的特征[3],而针对 Reuters 数据集,选取 1000 个互信息贡献度最大的单词作为描述文档的特征单词,形成最终的共现矩阵,作为本文的实验数据.

表 1 数据集描述

名称	类别	特征词数	每类文档数	总规模
$\text{Binary}_{1,2,3}$	2	2000	250×2	500
$\text{Multi5}_{1,2,3}$	5	2000	100×5	500
$\text{Multi10}_{1,2,3}$	10	2000	50×10	500
Reuters2	2	1000	3713,2055	5768
Reuters4_1	4	1000	321,245,142,110	818
Reuters4_2	4	1000	298,197,114,90	699
Reuters8	8	1000	321,245,142,110,298,197,114,90	1517

4.2 实验评估方法

假设 $C=\{c_1,c_2,\cdots,c_k\}$ 为数据集中真实的类标签,如果数据对象 x 属于类 c,则 $q(c|x)=1$,否则 $q(c|x)=0$.为评估各种算法的性能,首先为每一个学习得到的簇 t 分配一个对应的真实类标签 c,用函数 $h(t)=\arg\max_c|\{x|p(t|x)=1\wedge q(c|x)=1\}|$来表示,然后采用精确率(Precision,$P$)、召回率(Recall,$R$)和 $F1$ 度量($F1$-measure,$F1$)作为评价算法性能的指标[4],其计算方法如下:

$$P=\frac{1}{|T|}\sum_{t\in T}\frac{|\{x|p(t|x)=1\wedge q(h(t)|x)=1\}|}{|\{x|p(t|x)=1\}|} \tag{9}$$

$$R=\frac{1}{|T|}\sum_{t\in T}\frac{|\{x|p(t|x)=1\wedge q(h(t)|x)=1\}|}{|\{x|q(h(t)|x)=1\}|} \tag{10}$$

$$F1=\frac{2PR}{P+R} \tag{11}$$

精度度量了每一个学习到的簇中数据对象的纯度,而召回率衡量了数据分析方法的查全率,$F1$ 度量为精度和召回率的调和平均值.

4.3 实验设计

本文欲给出 DSIB 算法与 k-means 算法,Normalized Cuts(NCuts)算法[16],sIB 算法[3],文献[3]中改进的 IB 算法[3]及 MCRD 算法[20]的对比试验.在五个对比算法中,前三种算法在做数据分析时均没考虑单个数据对象自身模式特征的明确程度,而强制地将每个数据对象都指派到某一簇中;后两种算法可实现对数据的选择分析.

4.4 实验结果

4.4.1 数据选择实验

图 2 给出了随着参数 λ 的减小,DSIB 算法在部分数据集上数据分析的精度和召回率变化的曲线图.DSIB 算法在本文其他数据集上的运行结果呈现出类似的效果图,这里不再一一给出.在该图中,随着 λ 的减小,DSIB

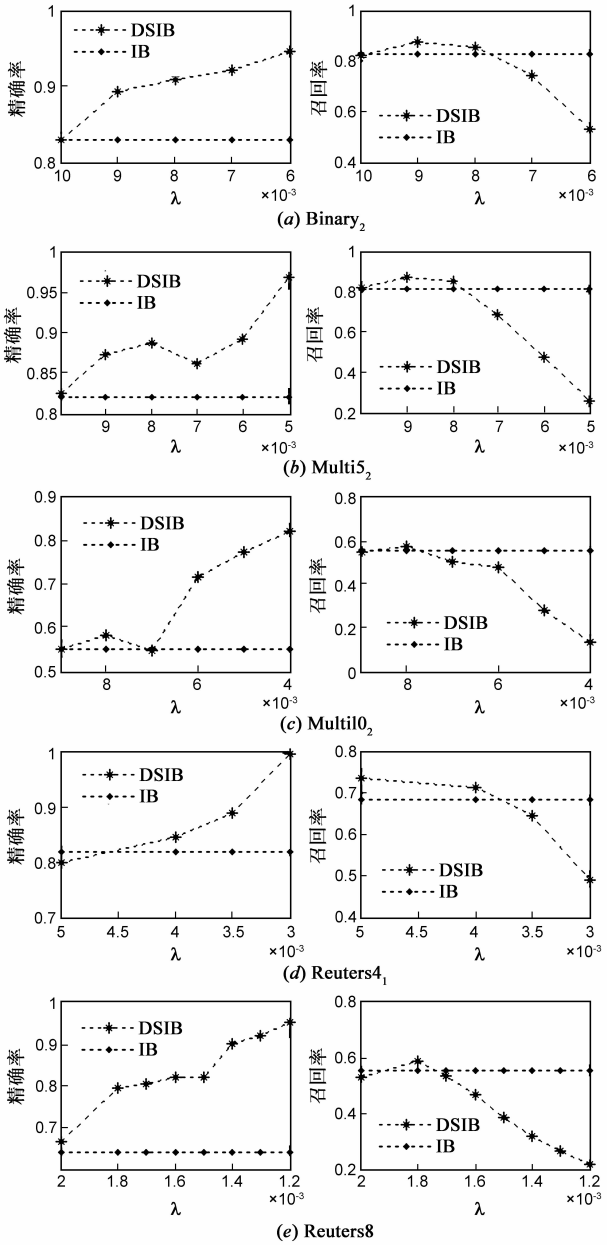

图2　精度和召回率随参数λ减小的曲线图

算法所得的精度整体上呈现出越来越高的趋势,而召回率呈现出先小幅度提高,后随之下降的趋势.

随着 λ 的减小,DSIB 算法对所选取数据对象模式特征的明确程度要求越高,所选择数据对象之间所蕴含的模式越容易被发现,相应的簇越纯,精度越高.图 2 中 DSIB 算法所得的精度曲线图便验证了这一期望.另外,随着 λ 的减小,DSIB 算法所选择的数据对象个数会越小,所识别的数据对象也越少,相应的召回率也期望越小.图 2 中的召回率曲线图验证了该期望.

需要注意的是,当 λ 取值在某一区间时,DSIB 算法仅判定少数数据对象的模式特征不够明确,此时的召回率相对于未做数据选择的 IB 算法会有所的提高.分析其原因,可能是因为被判定为簇结构特征不明确的数据对象对簇结构学习过程的负面影响较大,削弱了的簇与簇之间的区别程度.这些数据对象的存在,使得 IB 方法不能很好的发现数据中所蕴含的内在结构.而 DSIB 算法在做数据分析时,充分考虑了每一个数据对象对簇结构的影响程度,排除这些模式特征不明确的数据对象,从而可更好的发现数据中所蕴含的内在模式.表 2 是 IB 算法[3]与 DSIB 算法对 Reuters4_1 数据集做分析时所得的混淆矩阵.其中 T 表示算法学习到的簇,C 表示数据中真实的类标签.从该表中我们可以得出,IB 算法对全部的 818 个数据对象做分析时,$c2$ 类别中的数据对象被分到 $t2$ 和 $t3$ 两个簇中,而 $c3$ 类别和 $c4$ 类别中的数据对象则被合并到同一个簇 $t4$ 中.此时,原 IB 算法并没有很好地识别出数据中所固有的 4 个真实类别.而 DSIB 仅对 818 个数据对象中的 756 个数据对象做模式分析,并较好地识别出数据中所固有的 4 个真实类别.因此,DSIB 算法通过对数据对象有选择的进行分析,可消弱模式特征不明确的数据对象所带来的负面影响,进而可识别出更多的真实类别.此时,在一定的 λ 取值区间内,相应的精度和召回率都会有所提高.

表 2　IB 算法与 DSIB 算法在 Reuters4_1 数据集上的混淆矩阵

	IB (818)				DSIB (756)				
T \ C	$c1$	$c2$	$c3$	$c4$	T \ C	$c1$	$c2$	$c3$	$c4$
$t1$	295	0	16	1	$t1$	287	3	8	1
$t2$	8	159	3	2	$t2$	1	229	0	0
$t3$	0	86	3	0	$t3$	23	1	92	0
$t4$	18	0	120	107	$t4$	2	2	4	103

图 3 给出了 DSIB 算法与改进的 IB 算法(记为 Imp-IB)[3]及 MCRD 算法[20]在做数据选择分析时的精度-召回率对比曲线图.DSIB 算法的数据选择和簇结构的学习同时实施,而改进的 IB 算法及 MCRD 算法数据的选择与簇结构的学习相对分离.从该图中我们可以得出,

具有“边选择边学习”能力的 DSIB 算法在提高数据分析的精度时,所牺牲的召回率相对于改进的 IB 算法与 MCRD 算法较小.

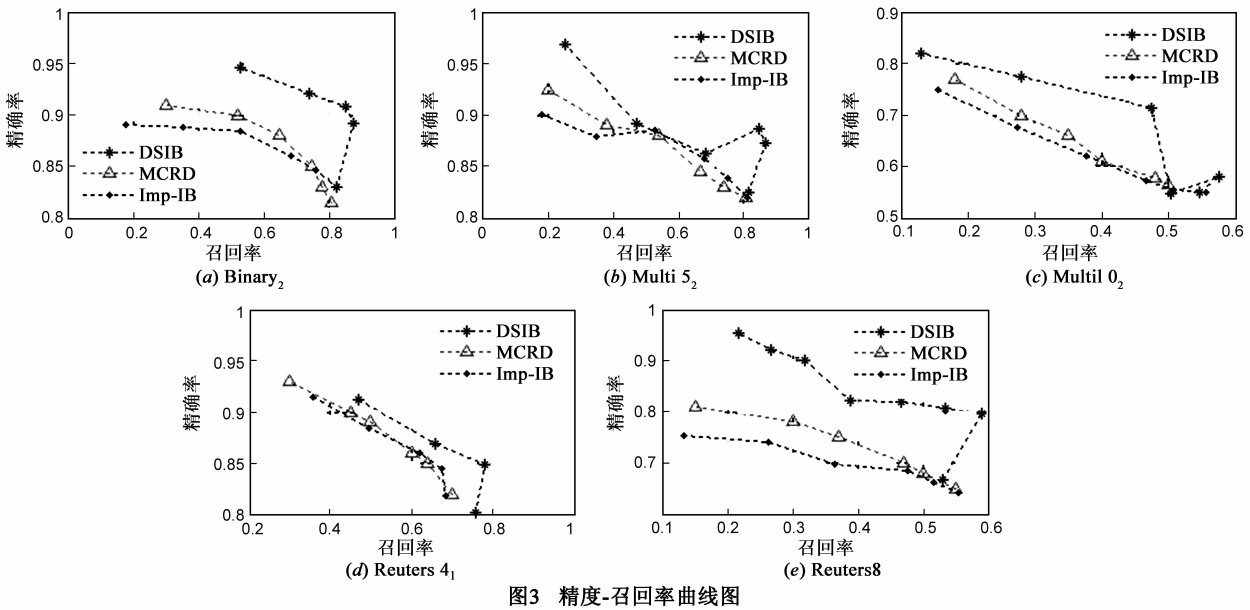

图3 精度-召回率曲线图

4.4.2 DSIB 算法与未做数据选择分析算法的对比实验

上节实验中我们注意到,当参数 λ 取值在某一区间时,DSIB 算法数据分析的精度和召回率相对于 IB 算法[3]都会有提高.在此,我们给出 DSIB 算法与未做数据选择的原 IB 算法[3]的对比实验,如表 3 所示.该组实验给出的是在 10 次随机初始划分上运行结果的平均值与标准差.从该表中可以得出,通过有选择的对数据进行模式分析,消弱模式特征较弱的数据对象给数据分析带来的负面影响,DSIB 算法数据分析的性能在精度、召回率、$F1$ 度量三方面均明显高于未做数据选择的 IB 算法,平均性能分别提高 7.3%、7.2%、7.3%.

表 3 DSIB 算法与 IB 算法的对比实验

Datasets	DSIB			IB			Improvement		
	P%	R%	$F1$%	P%	R%	$F1$%	P%	R%	$F1$%
Binary$_1$	86.1±4.2	83.5±4.8	84.8±4.5	83.7±8.7	83.4±8.9	83.5±8.8	2.4	0.2	1.3
Binary$_2$	81.2±7.3	78.0±8.3	79.5±7.8	75.1±10.6	74.7±10.5	74.9±10.5	6.1	3.3	4.6
Binary$_3$	88.4±8.8	83.0±9.9	85.6±9.4	71.9±16.9	71.0±16.9	71.5±16.9	16.5	12.0	14.1
Multi5$_1$	86.9±6.6	86.2±6.8	86.6±6.7	78.6±9.4	78.4±9.5	78.5±9.4	8.3	7.8	8.1
Multi5$_2$	89.6±1.9	88.4±1.9	89.0±1.9	78.7±9.1	79.0±8.6	78.8±8.8	10.9	9.4	10.2
Multi5$_3$	88.2±8.5	87.1±9.2	87.6±8.8	84.8±9.9	84.4±10.2	84.6±10.0	3.4	2.7	3.0
Multi10$_1$	61.7±4.9	62.1±4.8	61.9±4.8	55.9±6.0	57.3±6.2	56.6±6.1	5.8	4.8	5.3
Multi10$_2$	62.0±3.4	62.0±3.5	62.0±3.4	56.1±4.3	57.5±4.5	56.8±4.4	5.9	4.5	5.2
Multi10$_3$	63.8±3.1	64.4±3.3	64.1±3.2	57.6±4.7	58.8±4.9	58.2±4.8	6.2	5.6	5.9
Reuters2	84.9±0.0	87.9±0.0	86.4±0.0	81.9±2.5	84.1±3.4	83.0±2.9	3.0	3.8	3.4
Reuters4$_1$	85.8±7.1	85.4±10	85.5±8.5	74.4±5.8	63.3±8.8	68.1±6.6	11.4	22.1	17.4
Reuters4$_2$	92.4±3.6	90.9±1.6	91.7±2.6	79.9±9.0	78.8±13.4	79.3±11.2	12.5	12.1	12.4
Reuters8	69.0±3.1	66.6±4.3	67.8±3.6	66.1±4.8	60.8±10.4	63.1±7.8	2.9	5.8	4.7
Avg.	80.0	78.9	79.4	72.7	71.7	72.1	7.3	7.2	7.3

图 4 为 DSIB 算法与 IB 算法、*k*-means 算法、Normalized Cuts(NCuts)算法的对比实验.该图展示的是对比算法在 10 次随机初始划分上运行结果的平均值.横坐标中的 13 个数字分别顺序对应表 3 中所罗列的 13 个数据集.从图 4 中可以看出:(1)DSIB 算法与 IB 算法在 20-Newsgroup 的 9 个子数据集上数据分析的性能明显优于 *k*-means 算法和 Normalized Cuts 算法,而 DSIB 算法对这 9 个数据集内在模式的识别能力又明显优于 IB 算法;(2)IB 算法在 Reuters-21578 的 4 个子数据集上模式识别的能力不如 *k*-means 算法与 Normalized Cuts 算法,而 DSIB 算法对这四个数据集分析的能力整体上优于 *k*-means 算法与 Normalized Cuts 算法;(3)DSIB 算法在所有数据集上的平均精度、平均召回率、平均 *F*1 度量三方面均明显优于其它对比算法.

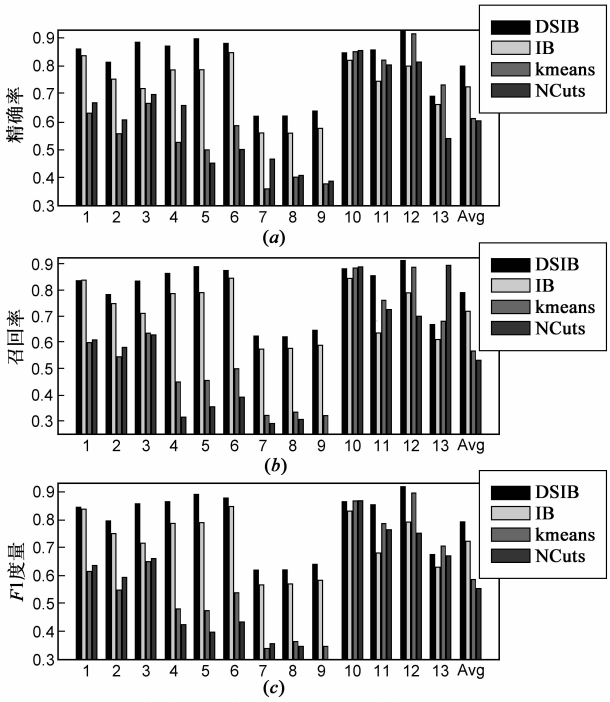

图4　DSIB算法与IB算法、*k*-means算法、Normalized Cuts算法的对比实验

4.4.3　收敛性实验

图 5 给出了在 3 个不同 λ 参数取值上,DSIB 算法的主循环部分重复迭代运行的次数.图中显示的迭代数值为 10 次随机初始化运行迭代次数的平均值.这里仅给出 DSIB 算法在 $Multi5_2$、$Multi10_2$ 与 $Reuters4_1$ 三个数据集上的收敛迭代次数,在其它数据集上具有类似的迭代收敛效果,这里不再一一给出.从图 5 中我们可以看出,DSIB 算法通过数十次的重复迭代便可得到一个稳定的压缩模式.

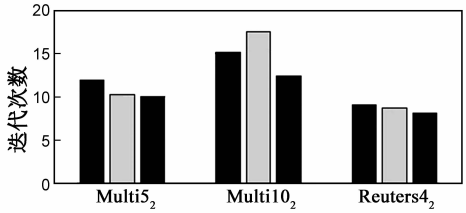

图5　DSIB算法主循环中的重复迭代次数

5　结论

由于模式特征不明确的数据对象与簇结构的拟合度不强,若强制地将这些数据对象指派到某一簇中,可能会拉偏数据中所固有的簇结构,从而影响 IB 方法数据分析的性能.针对该问题,本文通过定义一个“基于明确因素”的数据选择模型,并在此基础上提出 DSIB 算法,使得 IB 方法有选择的对数据进行模式分析.DSIB 算法根据数据对象到“瓶颈”变量的压缩过程中所产生的信息损失来判定数据对象自身模式特征是否明确,从而实现对数据的选择;采用“边选择边学习”的顺序“抽取-合并”策略来优化 DSIB 目标函数,使得簇结构的学习过程随时考虑每一个数据对象对簇结构的影响程度.实验结果表明:通过对数据进行选择性分析,DSIB 算法有效地挖掘出高精度的簇;DSIB 算法在提高数据分析的精度时,所牺牲的召回率较小,可更准确的找到更多的数据对象;DSIB 算法的数据内在模式的识别能力优于未做选择分析的原 IB 算法、*k*-means 算法及 Normalized Cuts 算法;DSIB 算法可快速收敛到一个稳定的压缩模式,具有较高的学习效率.

参考文献

[1] Tishby N, Pereira F, Bialek W. The information bottleneck method[A]. Proceedings of 37th Allerton Conference on Communication, Control and Computing[C]. Monticello, IL: IEEE Press, 1999. 368 – 377.

[2] Slonim N, Tishby N. Agglomerative information bottleneck [A]. Proceedings of Advances in Neural Information Processing Systems[C]. Denver, CO: MIT Press, 1999. 617 – 623.

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

[4] Slonim N. The information bottleneck: theory and application [D]. Jerusalem: The Hebrew University of Jerusalem, 2002.

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

[6] Bardera A, Rigau J, Baoda I, et al. Image segmentation using information bottleneck method[J]. IEEE Transactions on Image Processing, 2009, 18(7): 1601 – 1612.

[7] Lou Z, Ye Y, Yan X. The multi-feature information bottleneck with application to unsupervised image categorization[A]. Proceedings of 23rd International Joint Conference on Artificial Intelligence[C]. Beijing, China: AAAI Press, 2013. 1508 – 1515.

[8] Hecht R, Noor E, Tishby N. Speaker recognition via gaussian information bottleneck[A]. Proceedings of InterSpeech[C]. Brighton, UK: ISCA Press, 2009. 1567 – 1570.

[9] 沈华伟,程学琪,陈海强,刘悦. 基于信息瓶颈的社区发现[J]. 计算机学报,2008,31(4):677 – 686.
SHEN Hua-wei, CHENG Xue-qi, CHEN Hai-qiang, LIU Yue. Information bottleneck based community detection in network[J]. Chinese Journal of Computers, 2008, 31(4): 677 – 686. (in Chinese)

[10] Lazebnik S, Raginsky M. Supervised learning of quantizer codebooks by information loss minimization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(7): 1294 – 1309.

[11] 叶阳东,何锡点,贾利民. 面相范畴类型数据的 sIB 算法[J]. 电子学报,2009,37(10):2165 – 2172.
YE Yang-dong, HE Xi-dian, JIA Li-min. CD-sIB: a kind of sIB algorithm orient to categorical data[J]. Acta Electronica Sinica, 2009, 37(10): 2165 – 2172. (in Chinese)

[12] 袁华强,叶阳东,刘东. 遗传顺序 IB 算法[J]. 电子学报,2009,37(8):1804 – 1809.
YUAN Hua-qiang, YE Yang-dong, LIU Dong. Genetic sequential IB algorithm[J]. Acta Electronica Sinica, 2009, 37(8): 1804 – 1809. (in Chinese)

[13] 朱真峰,叶阳东,Gang Li. 基于变异的迭代 sIB 算法[J]. 计算机研究与发展,2007,44(11):1832 – 1838.
ZHU Zhen-feng, YE Yang-dong, LI Gang. Iterative sIB algorithm based on mutation[J]. Journal of Computer Research and Development, 2007, 44(11): 1832 – 1838. (in Chinese)

[14] Ye Y, Ren Y, Li G. Using local density information to improve IB algorithms[J]. Pattern Recognition Letters, 2011, 32(2): 310 – 320.

[15] Ester M, Kriegel H-P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[A]. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining[C]. Portland, Oregon: AAAI Press, 1996. 226 – 231.

[16] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888 – 905.

[17] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972 – 976.

[18] Gupta G, Ghosh J. Bregman bubble clustering: a robust, scalable framework for locating multiple, dense regions in data[A]. Proceedings of 6th International Conference on Data Mining[C]. Piscataway, NJ: IEEE Press, 2006. 232 – 243.

[19] Xiong Y, Zhu Y, Yu P S, et al. Towards cohesive Anomaly mining[A]. Proceedings of 27th AAAI Conference on Artificial Intelligence [C]. Bellevue, Washington: AAAI Press, 2013. 984 – 990.

[20] Crammer K, Talukdar P P, Pereira F. A rate-distortion one-class model and its application to clustering[A]. Proceedings of 25th International Conference on Machine Learning[C]. New York: ACM Press, 2008. 184 – 191.

[21] Cover T M, Thomas J A. Elements of information theory[M]. New York: John Wiley and Sons, 1991.

作者简介

娄铮铮 男,1984 年 4 月出生于河南原阳. 郑州大学在读博士研究生. 主要研究方向为机器学习、模式识别、计算机视觉.
E-mail: iezzlou@gmail.com

杨 晨 男,1987 年 7 月出生于河南平顶山. 硕士研究生. 主要研究方向为机器学习.
E-mail: yangchenwo@126.com

叶阳东(通信作者) 男,1962 年 10 月出生于河南潢川. 工学博士、郑州大学信息工程学院教授、博士生导师. 主要研究方向为智能系统、机器学习、数据库.
E-mail: yeyd@zzu.edu.cn