一种面向非平衡数据的多簇IB算法

江 鹏 叶阳东 娄铮铮

(郑州大学信息工程学院 郑州 450052)

摘 要 信息瓶颈(Information Bottleneck,IB)方法在处理非平衡数据集时,倾向于将大簇中的数据对象划分到数据规模较小的小簇中,造成了聚类效果不理想的问题。针对该问题,提出了一种面向非平衡数据的多簇信息瓶颈算法(McIB)。McIB算法采用向下抽样方法来降低非平衡数据集的倾斜度,使用先划分再学习后合并的策略来优化IB算法处理非平衡数据的合并抽取过程。整个算法包含3步:首先根据分离标准来确定抽样比例参数;然后对数据进行初步的聚类,生成可信赖的多个簇;最后再利用簇之间的相似性对簇进行合并,组织多个簇代表每个实际的簇来得到最终的聚类结果。实验结果表明:所提算法能够有效地解决IB方法在非平衡数据集上的"均匀效应"问题;与其他聚类算法相比,McIB算法的性能更优。

关键词 聚类,IB算法,非平衡数据,多簇,簇合并

中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.7.044

Multi-clusters IB Algorithm for Imbalanced Data Set

JIANG Peng YE Yang-dong LOU Zheng-zheng

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

Abstract When dealing with imbalanced data sets, the original IB method tends to produce clusters of relatively uniform size, resulting in the problem of unsatisfactory clustering effect. To solve this problem, this paper proposesd a multi-clusters information bottleneck (McIB) algorithm. McIB algorithm tries to reduce the skewness of the data distributions by under-sampling method to divide the imbalanced data sets into multiple relatively uniform size clusters. Entire algorithm consists of three steps. First, a dividing measurement standard is proposed to determine the sampling ratio parameter. Second, McIB algorithm preliminary analyses the data to generate reliable multi-clusters. At last, McIB algorithm merges clusters into one bigger size cluster according to the similarity between clusters and organizes multiple clusters representing the actual cluster to obtain the final clustering results. Experimental results show that the McIB algorithm can effectively mine the pattern resided in imbalanced data sets. Compared with other common clustering algorithms, the performance of the McIB algorithm is better.

Keywords Clustering, Information bottleneck method, Imbalanced data, Multi-clusters, Cluster merging

1 引言

随着互联网技术的迅猛发展,大数据时代已经到来,大规模信息呈现爆炸式增长,现实世界中有着许多非平衡数据集[1-3],对非平衡数据中蕴藏的数据模式的挖掘显得相当重要。顾名思义,非平衡数据集是指同一个数据集中某些类的数据对象个数远远大于其他类的数据对象个数[4-6],它广泛存在于现实生活中。例如在网络入侵检测数据中,入侵数据相对于正常的登录日志来说只占很少的一部分;在医疗数据中,假设有1000例体检数据,体检结果正常的数量只有900多个,而体检结果有异常的可能只有几十个;在文档数据中,归属于每一类的文档对象个数是不定的,且时常在数量上相去甚远[2,6]。所以对非平衡数据集的数据进行分析具有很高的研究价值。

作为机器学习和模式识别中最重要的问题之一,非平衡数据的聚类问题近年来引起了国内外专家的广泛关注,取得了一系列成果,并在相关领域得到广泛应用[1,6,7]。当前国内外的针对非平衡数据的研究主要集中在有监督的学习中,一般的处理方法都集中在组合分类器算法上[8-10];近年,针对非平衡数据的特性,国内外各个研究小组提出了许多基于聚类方法的采样分类算法[10-12]。非平衡数据的非监督学习方法不多:Xiong等人[13]从聚类度量指标的角度对数据类分布对于K-means[9]算法的影响进行了研究,其中还理论上证明了K-means在处理基于欧氏空间数据时产生"均匀效应"[13,14]的原因;另外Jianqian等人[15]提出了一种基于图论的谱聚类方法;Kumar等人[6]基于K-means算法提出了一种针对非平衡数据集的聚类分析算法。但是目前针对非平衡数据的聚类算法大多是针对欧氏空间的非平衡数据,针对基于联合概率分布的非平衡共现数据还缺乏研究。

传统的机器学习算法一般都基于这样一种假定:实验数据是类平衡分布的,即各类数据的数量差别不大,不是一类数据的数量远远大于另一类的非平衡数据集,这样就使得传统

到稿日期:2015-04-01 返修日期:2015-07-19 本文受国家自然科学基金项目:多变量IB方法及算法的研究(61170223),国家自然科学基金联合基金项目:可扩展迁移学习中跨媒体复杂问题自动映射研究(U1204610)资助。

江 鹏(1990—),男,硕士,主要研究方向为机器学习,E-mail:iejiangpeng@gmail.com;叶阳东(1962—),男,博士,博士生导师,主要研究方向为智能系统、机器学习、数据库;娄铮铮(1984—),男,博士,讲师,主要研究方向为机器学习、模式识别、计算机视觉。

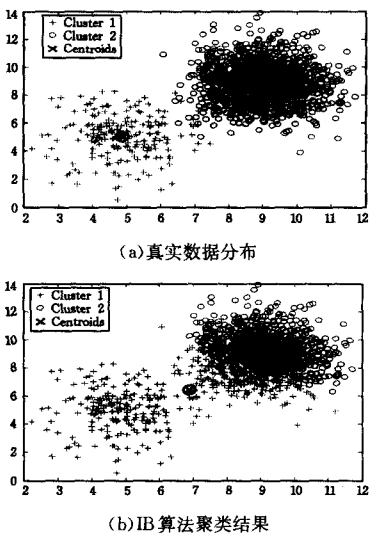
算法在平衡数据集上的性能要远远优于在非平衡数据集上的性能[16]。为了解决类不平衡问题，现有的研究方法主要集中在两个层面：1)修改数据的类分布情况(数据层面)，包括向下抽样算法、向上抽样算法；2)提出新的算法来提高少数类的识别效果(算法层面)[11,12]。向下取样方法在处理非平衡数据问题时是一种常见且有效的数据预处理方法。向下抽样方法在处理非平衡数据上具有极高的效率，它通过减少归属于大簇的数据对象个数使得数据分布相对平衡。向下抽样方法仅仅使用归属于大簇中的一部分具有代表性的子集来进行数据分析，抽样比例参数的不确定性导致其忽略了大簇中大量的有效信息，从而影响算法的识别性能。

IB方法(Information Bottleneck Method)[17]是Tishby等人于1999年提出的一种基于率失真理论[18]的数据分析方法。IB方法处理数据的过程可以看成一个数据压缩的过程，它在将数据对象压缩到一个"瓶颈"变量的同时，还要最大化地保存源信息中所蕴含的信息量，从而发现数据中所蕴含的内在模式。IB方法能够从数据中挖掘出高质量的聚类结果，已经在众多领域取得了成功的应用[19,20]。

然而，IB方法在处理非平衡数据时，可能会将大簇中的数据对象划分到其他簇规模较小的若干个簇中。实验发现，传统的聚类算法比如IB算法、K-means算法，在处理非平衡数据时，趋向于将属于大簇中的数据对象划分到小簇中，往往会得到规模大小相当的聚类结果[4,6]，这一现象被称为"均匀效应"。为了更加形象地说明这个问题，本文做了如下两个实验：现有一个文档数据集是由4034个数据对象组成的，且每个对象有2000个特征属性，本数据集由两个簇组成，其中一类有3713个数据对象，另一类有321个数据对象，其数据倾斜率是11.57。采用基本IB聚类算法聚类后的结果如表1所列。从表中可以看出，基本IB算法明显将一部分属于"大簇"中的数据(933)划分到了"小簇"(321)中，产生了"均匀效应"。

表1　基本IB处理非平衡数据集的混淆矩阵

T \ C	c1	c2
t1	2780	0
t2	933	321



(a)真实数据分布

(b)IB算法聚类结果

图1

本文在人工数据点集上利用基本IB算法运行的结果如图1所示，其更清晰地说明了IB算法在处理非平衡数据时的"均匀效应"，从中可以看出部分属于大簇的数据被划分到了小簇中，得到的聚类结果中簇心有相互靠近的趋势。

针对上述问题，本文提出了一种面向非平衡数据集的多簇信息瓶颈算法(Multi-clusters Information Bottleneck，McIB)。相较于向下取样方法，McIB算法可以有效地避免大簇中的重要信息丢失；相较于信息瓶颈算法，McIB算法能够有效降低数据类分布引起的"均匀效应"对聚类效果的影响。在McIB算法中，首先提出了一种抽样标准用于确定抽样比例参数，并将其应用于把整个数据划分成多个小簇，再对数据进行初步的聚类生成多个可信赖的簇，最后对簇进行聚类合并，组织多个簇代表每个真实的簇，从而确定最终的聚类数目。由于部分数据对象并不能充分地反映大簇的特征，导致大簇中的部分对象将被划分到小簇中，因此McIB算法利用多个簇代替单个簇来代表每个实际的簇，此时多个簇能够将一个规模较大的簇划分成若干个规模相对较小的子集。这样做能够重新平衡大簇和小簇之间的规模，从而缓解了因非平衡数据的分布特性引起的聚类有效性降低。最后，本文针对非平衡共现数据的特性，使用簇合并后的互信息损失程度[21]作为簇之间的相似性度量测度，对这些簇重新组合，将代表同一簇的数据对象合并为一个簇，从而确定最终的聚类数目，得出最终的聚类结果。

本文的主要工作如下：

(1)在IB方法中，一种分离标准被提出以确定向下抽样方法中的抽样比例参数，并将其应用于将大簇划分成多个小簇，将源变量压缩为可信赖的多个规模较小的簇，重新平衡大簇和小簇之间的规模，从而减少因非平衡数据的分布特性引起的聚类有效性降低；

(2)采用簇合并之后特征信息的损失量来度量簇之间的相似度，利用簇之间的相似度对这些簇重新组合，将代表同一簇的数据对象合并为一个簇，经过反复合并融合来优化聚类效果。

2　相关知识

2.1　互信息

互信息(Mutual Information，MI)[17-20]用于描述离散随机变量之间相互包含信息的多少，它是对变量之间相互关联程度的度量，其定义如下。

定义1　若已知两个离散随机变量(X,Y)服从于联合概率分布$p(x,y)$，即$(X,Y)\sim p(x,y)$，那么离散随机变量X与Y间的互信息$I(X;Y)$定义为：

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

2.2　非平衡数据倾斜率

非平衡数据倾斜率(Imbalance Ratio，IR)[22,23]反映了数据内部的类分布情况，它是非平衡数据集类分布的一个重要参考指标，其定义如下。

定义2　若非平衡数据集X中属于最大簇的数据对象的个数为p，属于最小簇的数据对象的个数为q，则本文中定义该数据集的倾斜率为p/q。

2.3　IB方法

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，如图2所示，其中X表示带分析的数据对象，Y表示描述数据对象的特征变量，T为压缩"瓶颈"变量。变

量 X 到 T 的压缩编码 $p(t|x)$ 即为 IB 方法所获得的压缩模式，若一些数据对象被压缩到同一个簇 T 中，则它们被视为具有相同的特征模式。为使压缩编码 $p(t|x)$ 尽可能如实地反映数据中所蕴含的内在模式，IB 方法在对数据进行压缩的同时，要求“瓶颈”变量 T 尽可能最大化地保存特征变量 Y 中所载有的信息量。变量 Y 客观地描述了数据的特征，是 IB 方法数据压缩的依据。

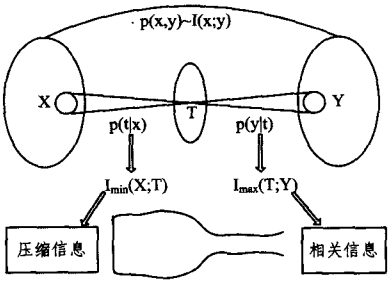


图 2　IB 方法的核心思想

为有效地提取数据中数据模式，文献[19,20]给出了如式(2)所示的 IB 目标函数，其中 $I(X;T)$ 度量了 X 到 T 的压缩程度，$I(T;Y)$ 用来度量相关信息的保存量，$\beta\in[0,\infty)$ 是平衡参数，用于平衡 X 的压缩与相关信息 Y 的保存。在实验中，由于 $|T|\ll|X|$，T 本身在很大程度上就对 X 进行了压缩，因此，β 取值为 ∞，将 IB 的重点放到相关信息 $I(T;Y)$ 的保存上，此时 IB 方法的目标函数可以简化为式(3)中的目标函数。

$$L_{max}=I(T;Y)-\beta^{-1}I(X;T) \tag{2}$$

$$L_{max}=I(T;Y) \tag{3}$$

3　多簇 IB 算法

本节以两类问题说明算法的流程：假设有数据源 $X(X_1,X_2)$，其中 X_1 和 X_2 是数据 X 中蕴藏的聚类模式，初始随机划分为 $X(Z_1,Z_2,\cdots,Z_i)$，其中 $|Z_i|\approx|X_2|(1\leqslant i\leqslant m)$。最特殊的一种划分是 $X(X_{11},X_{12},\cdots,X_{1m},X_2)$，它将 X_1 划分为若干个小簇，$X_{1i}(1\leqslant i\leqslant m)$ 的规模大小和 X_2 的规模大小相当，即 $|X_{1i}|\approx|X_2|$。以小簇 X_2 为划分基准是为了使非平衡数据中的小规模簇能被准确地识别出来，因为在现实生活中，人们关心的对象往往是那些规模较小的数据对象，这有利于有效地发现数据中的隐藏模式，例如在体检数据中，体检结果异常的数据对象往往只占很少的比例，而有效识别这些数据具有显著意义。此时，经过初步地聚类得到 $X(X_{11},X_{12},\cdots,X_{1m},X_2)$，再将源数据 X 压缩到 T_0 中，得到进一步的划分 $T_0=\{t_1,t_2,t_3,\cdots,t_m\}$，最后根据 t_i 之间的相似性度量标准，计算 $\{t_i,t_j\}$ 两两之间的合并代价，将合并代价最小的两簇合并为同一个簇，如此反复循环直到簇数目迭代到目标簇数目。算法流程如图 3 所示。

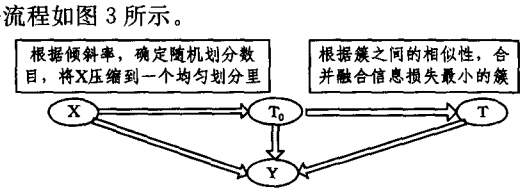


图 3　算法流程图

从图 3 中可以看出，该算法主要核心包括 3 个步骤：1)多簇数目的确定，并得到一个准确可信的均匀划分；2)初步的高质量聚类结果，生成可信赖的多个簇；3)得到初步的可信多个簇后，将归属于大簇的数据对象进行合并，避免“均匀效应”。

3.1　多簇数目的确定

一个有效的簇数目可以将非平衡数据集进行一个均匀的划分，这对于算法整体性能的提升具有极大的意义。本节在此给出簇数目 m 的一种确定方法：m 是一个关于实际簇数目 k 和非平衡数据集倾斜率 λ 的变量，其计算方法可以通过式(4)得到，其中 $\lfloor\lambda\rfloor$ 表示向下取整。

$$m=\begin{cases}\lfloor\lambda\rfloor+k-1, & \lambda-\lfloor\lambda\rfloor<\alpha\\ \lfloor\lambda\rfloor+k, & \lambda-\lfloor\lambda\rfloor\geqslant\alpha\end{cases} \tag{4}$$

其中，$k=|T|$，α 是一个经验值，本文中取值为 0.7，这样能够保证数据划分尽量均匀，经过后续的聚类分析后得到一个初步的平衡结果。

3.2　相似性度量

本文用簇合并之后特征信息的损失量来度量簇之间的相似性。显然，归属于同一簇的数据对象在合并时，因为簇之间的特征信息相似度较高，所以合并之后的特征信息损失程度是最小的。簇之间的合并代价越小，则簇之间的相似度越高；相反，簇之间的合并代价越大，则它们之间的相似度越小。对于 T 中任意两个对象 t_i,t_j，合并 $\{t_i,t_j\}\Rightarrow t^{new}$ 所产生的特征信息损失程度称为合并代价，根据式(3)将其定义为：

$$\Delta L_{max}(t_i,t_j)=I(T^{bef},Y)-I(T^{aft},Y) \tag{5}$$

其中，$I(T^{bef},Y)$ 和 $I(T^{aft},Y)$ 分别代表合并 $\{t_i,t_j\}$ 前后的 T 和 Y 之间特征信息的保存程度。由于 X,T 和 Y 形成 IB Markov 链：$X\leftrightarrow T\leftrightarrow Y$，因此有：

$$p(t^{new})=p(t_i)+(t_j) \tag{6}$$

$$p(y|t^{new})=\frac{p(t_i)}{p(t^{new})}p(y|t_i)+\frac{p(t_j)}{p(t^{new})}p(y|t_j) \tag{7}$$

将式(6)和式(7)代入式(5)得：

$$\begin{aligned}&\Delta L_{max}(t_i,t_j)\\&=I(T^{bef},Y)-I(T^{aft},Y)\\&=p(t_i)\sum_y p(y|t_i)\log\frac{p(y|t_i)}{p(y)}+p(t_j)\sum_y p(y|t_j)\log\\&\quad\frac{p(y|t_j)}{p(y)}-p(t^{new})\sum_y p(y|t^{new})\log\frac{p(y|t^{new})}{p(y)}\\&=p(t_i)\sum_y p(y|t_i)\log\frac{p(y|t_i)}{p(y)}+p(t_j)\sum_y p(y|t_j)\log\\&\quad\frac{p(y|t_j)}{p(y)}-\sum_y p(t^{new})p(y|t^{new})\log\frac{p(y|t^{new})}{p(y)}\\&=p(t_i)\sum_y p(y|t_i)\log\frac{p(y|t_i)}{p(y)}+p(t_j)\sum_y p(y|t_j)\log\\&\quad\frac{p(y|t_j)}{p(y)}-\sum_y p(t_i)p(y|t_i)\log\frac{p(y|t^{new})}{p(y)}-\sum_y p(t_j)\\&\quad p(y|t_j)\log\frac{p(y|t^{new})}{p(y)}\\&=p(t_i)\sum_y p(y|t_i)\log\frac{p(y|t_i)}{p(y|t^{new})}+p(t_j)\sum_y p(y|t_j)\log\\&\quad\frac{p(y|t_j)}{p(y|t^{new})}\end{aligned} \tag{8}$$

故有，

$$\Delta L_{max}(t_i,t_j)=(p(t_i)+p(t_j))\cdot\bar{d}(t_i,t_j) \tag{9}$$

其中，$\bar{d}(t_i,t_j)\equiv JS_\Pi[p(y|t_i),p(y|t_j)]$，$JS_\Pi[p(y|t_i),p(y|t_j)]$ 是概率分布 $p(y|t_i)$ 和 $p(y|t_j)$ 之间的 JS 散度，$\Pi=\{\pi_1,\pi_2\}=\{\frac{p(t_i)}{p(t_i)+p(t_j)},\frac{p(t_j)}{p(t_i)+p(t_j)}\}$。

其中当 t_i 为只有一个元素的单元素簇时，$\Delta L\max(\{x\},t)$ 是式(9)的一种特殊情况。

3.3 McIB算法

McIB算法的具体描述如算法1所示。从该算法的步骤中可以看出：McIB算法的1－2步是对算法的输入数据进行预处理操作；第3－19步是对预处理数据进行初步的聚类分析，得到一个均匀的划分结果；第21－25步通过一个循环迭代过程，将属于一个大簇的数据对象进行重新组合，直到得到数据中的聚类模式。

算法1 McIB算法

输入：数据集 $X=\{x_1,x_2,x_3,\cdots,x_n\}$；聚类划分 T 的簇数目 k；数据倾斜度 λ

输出：X 的聚类划分模式 T

1. 将参数 k 与 λ 代入式(4)求得 m；
2. 根据公式 $p(x_i,y_j)=\frac{y_j^{x_i}}{\sum_i\sum_j y_j^{x_i}}$ 将共现数据 X 转换成相应的联合概率分布，其中 $y_j^{x_i}$ 为 y_j 在 x_i 中出现的次数；
3. $T_0\leftarrow X$ 随机初始划分为 m 个数据模式；
4. $\beta=\infty$
5. Flag=False；
6. While ! Flag
7. Flag=TRUE；
8. For $\forall x\in X$
9. 将 x 从其当前所属的簇 t 中取出，形成单元素簇{x}；
10. For $\forall t\in T_0$
11. 根据式(9)计算 $\Delta L_{max}(\{x\},t)$
12. End for
13. 将 x 合并到新簇 $t^{new}\leftarrow\arg\min_{t\in T}\Delta L_{max}(\{x\},t)$ 中；
14. 更新 $T_0=\{T_0-\{\{x\},t\}\}\cup\{t^{new}\}$
15. If $t^{new}\neq t$
16. Flag=FALSE；
17. 将 x 合并到 t^{new} 中；
18. End if
19. End For
20. End While
21. $T\leftarrow T_0$，$\forall t_i,t_j\in T$，根据式(9)计算簇两两合并代价；
22. While $|T|>k$
23. $\{i,j\}=\arg\min_{i,j}\Delta L_{max}(t_i,t_j)$；
24. 合并$\{t_i,t_j\}$为 t^{new}，并更新 $T=\{T-\{t_i,t_j\}\}\cup\{t^{new}\}$；
25. $\forall t_i\in T$ 根据式(9)计算 $\Delta L_{max}(t_i,t^{new})$
26. End

3.4 算法时间复杂度分析

在McIB算法中，首先需要对数据集进行预处理。针对数据规模为 n 的数据集来说，将数据集转换为对应的联合概率分布的时间复杂度为 $O(n)$。核心算法中根据聚类划分 T 的簇数目 k 和数据倾斜度 λ 参数求最优划分 m 的过程的时间复杂度为 $O(1)$；循环体核心部分中，第9步“抽取”的时间复杂度为 $O(1)$；第10－12步计算当前划分的合并代价矩阵的时间复杂度为 $O(m)$；第13步的数据“合并”过程的时间复杂度为 $O(1)$；第14－19步的时间复杂度为 $O(1)$；第21步的时间复杂度为 $O(m^2)$；第22－25步的时间复杂度为 $O(km^2)$。故整个循环体的时间复杂度为 $O(tmn+km^2)$，其中 t 为算法收敛时所迭代的次数，m 为多簇数目，k 为簇的数目，n 为数据规模。对于已知的 X，簇数目 k 和参数 m 都是确定的常数，所以综合上面的分析可知，McIB算法的时间复杂度为 $O(tmn)$，可见该算法的时间复杂度与数据集的规模呈线性相关。

4 实验与评估

本文实验将分别在Reuters的6个子数据集上进行，以验证McIB算法的有效性。

4.1 实验设计

为了验证McIB算法的有效性，将该算法与以下4种经典聚类算法进行对比实验。

(1)IB算法：基于信息论的基本IB方法；

(2)K-means：一种经典的非监督学习算法；

(3) DSIB (Data Selection Information Bottleneck) 算法[21]：具有选择簇结构特征较为明确的数据对象分析性能的IB算法；

(4)NCuts(Normalized Cuts)算法：基于图分割的经典聚类算法[24]。

IB方法和K-means算法都是基于随机划分的经典非监督学习算法，对于非平衡数据具有其自身的局限性；DSIB算法是一种改进的IB算法，它在做数据分析时具有选择性功能；NCuts算法对于数据中的图结构的构造具有很高的敏感性，对于非平衡数据集具有较好的分析效果。

4.2 非平衡数据集

4.2.1 Reuters数据集

该数据集是路透社的新闻数据集，与20Newsgroup纯文本数据不同，它是SGML格式的文件，需要进行预处理。实验中，从原始数据集中去除含有多类标签的数据集，得到一个有65个类别的8293篇文档，选取其中最大的10类数据集，最终得到7285篇文档作为最终的实验数据。由于最大的2个类别分别含有3713篇文档和2055篇文档，而其他的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。针对这10个类，将2个大类与余下的8个小类随机组合，形成 $Reuters2_{13}$，$Reuters2_{24}$，$Reuters3_{134}$，$Reuters3_{245}$，$Reuters3_{167}$，$Reuters3_{289}$ 等非平衡数据，其中下标表示组成非平衡数据的类标号，详细说明如表2所列。实验中，针对这些数据集，利用特征提取算法提取其中对共现矩阵贡献度最大的2000个属性特征词作为描述文档的特征单词，形成本文的实验数据[20]。

表2 Reuters数据集

名称	类别	特征数	文档类分布	规模
$Reuters2_{13}$	2	2000	3713,321	4034
$Reuters2_{24}$	2	2000	2055,298	2353
$Reuters3_{134}$	3	2000	3713,321,298	4332
$Reuters3_{245}$	3	2000	2055,298,245	2598
$Reuters3_{167}$	3	2000	3713,197,142	4052
$Reuters3_{289}$	3	2000	2055,114,110	2279

4.3 评估方法

在非平衡数据集的处理之后，采用什么样的评估方法也是一个很重要的问题。一般的聚类问题中采用的评估标准有：聚类精度、召回率、F1度量、标准化互信息、兰德指数等。Xiong等人[13]研究发现基于信息熵的度量方法对于非平衡数据的聚类结果的评估方法具有局限性，故本文在实验中不采用与信息熵有关的度量方法。兰德指数是关于聚类准确率的度量方法，它在聚类中代表的是两个聚类结果之间的相似性，是聚类算法中常用的评估标准。然而由于非平衡数据集的独特构造性，只使用某一种度量方法是不合理的。因此，为了评

估聚类算法的有效性，本实验分析中将应用 3 个聚类有效性评测指标：1）聚类精度 Precision(P)；2） F1 度量 F1 measure；3）兰德指数 Rand Index(RI)[25,26]。

假定聚类结果为：$T=\{T_1,T_2,T_3,\cdots,T_k\}$，真实类标签为：$C=\{C_1,C_2,C_3,\cdots,C_k\}$，其中 k 为聚类结果中簇心的个数；对于聚类后的划分模式 $T_i\in T$，聚类结果类标号为该划分模式下占据显著地位的类标号，更加形象化地描述为：真实类标签 $C_i=\text{argmax}(C_i(find(T_i)))$；对于每一个真实类 $C_i\in C$，$a_i(C_i,T)$ 代表被正确分类到 C_i 中的对象个数，$b_i(C_i,T)$ 代表错误分到 C_i 的对象个数，$c_i(C_i,T)$ 代表应当正确分到 C_i 中但是被错误分到其他类中的对象个数。本文采用的评估方法的定义如下：

$$P=\frac{1}{k}\sum_{i=1}^{k}\frac{a_i(C_i,T)}{a_i(C_i,T)+b_i(C_i,T)} \tag{10}$$

$$F1=\frac{1}{k}\sum_{i=1}^{k}\frac{2a_i(C_i,T)}{2a_i(C_i,T)+b_i(C_i,T)+c_i(C_i,T)} \tag{11}$$

$$RI=\frac{\sum_{i=1}^{k}a_i(C_i,T)}{|X|} \tag{12}$$

4.4 实验结果及相关分析

图 4 给出了 McIB 算法与 IB 算法、K-means 算法、NCuts 算法分别在 Reuters 数据集上的性能对比。图 4 展示的是对比算法在 6 个 Reuters 子数据集上的 10 次运行的平均结果，其中横坐标为数据集，纵坐标分别为精度、F1 度量值、兰德指数。从图 4 可以看出：1）McIB 算法在 Reuters 的 6 个子数据集上的精度和兰德指数要优于其他聚类算法，个别算法在 F1 度量上的偏差较大造成 McIB 算法在平均结果上的表现一般；2）IB 算法在这 6 个子数据集上的精度和 F1 与 K-means 算法相比较差，但是在兰德指数上却有极大的优势；3）在第 4 个和第 6 个数据集上，虽然 McIB 算法分别在 F1 和兰德指数上的性能不及 DSIB 算法和 NCuts 算法，但是 McIB 算法在其他非平衡数据上的精度和兰德指数均有明显的优势。综上所述，McIB 算法在实验数据集上的数据分析性能整体要优于基本 IB、K-means、DSIB 和 NCuts 算法。

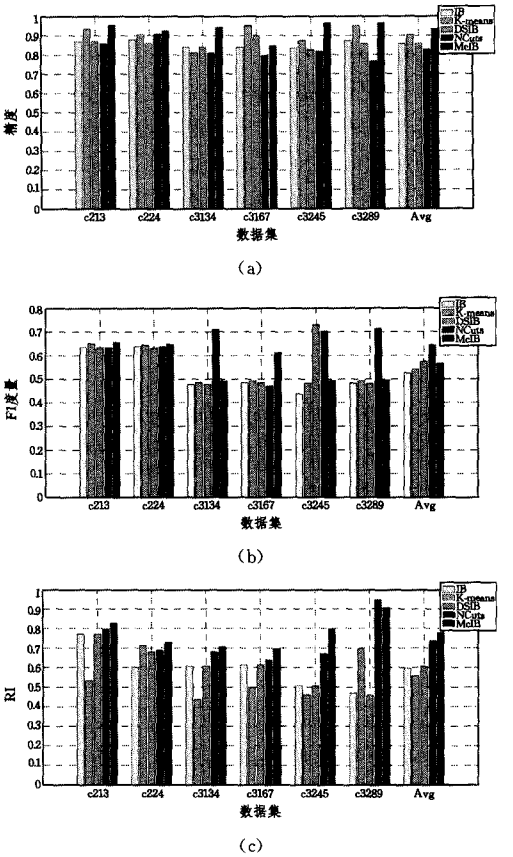


图 4 McIB 算法与 IB 算法、K-means 算法、DSIB 算法、NCuts 算法的对比实验

表 3 给出了不同算法在 $Reuters3_{289}$ 数据集上的数据分析结果，由表 3 可以看到由于“均匀效应”的出现，IB 算法、K-means 算法、DSIB 算法没有带来好的聚类结果；NCuts 算法虽然对大簇的划分较为准确，但是对规模较小的簇未能正确识别；从表 3 中还可以看出，McIB 算法所得的聚类结果非常接近非平衡数据的真实类分布。本算法在其他数据上也有类似的结果，这里不再一一列举。

表 3 在数据集 $Reuters3_{289}$ 上不同算法聚类结果的混淆矩阵

	IB			K-means			DSIB			Ncuts			McIB		
T \ C	c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3
t1	985	0	0	1574	0	0	929	0		2052	3	1	1840	0	0
t2	738	1	1	437	104	106	825	114	0	0	1	4	115	114	0
t3	359	113	109	44	10	4	301	0	110	3	110	105	100	0	110

下面给出 K-means 算法和 McIB 算法在实验数据集 $Reuters2_{13}$ 上的聚类分析结果，实验结果如表 4 所列。

表 4 K-means、McIB 处理非平衡数据集的混淆矩阵

	K-means		McIB	
T \ C	c1	c2	c1	c2
t1	1194	172	3349	0
t2	1773	149	364	321

从上述实验结果中可以还发现，K-means 算法在精度和 F 度量上整体要优于 IB 算法，但是在兰德指数上却表现不佳，对比表 1 和表 4 可以看出，相比于传统聚类算法，基本 IB 算法能够有效发现非平衡数据中的小规模数据，这在现实数据中往往是人们关心的数据模式，所以对实验结果仅使用一种度量标准是不科学的，应该予以综合考虑。从表 1 和表 3 中还可以得到如下结论来验证思想：1）传统的聚类学习方法如 IB 方法、K-means 算法，在处理非平衡数据时倾向于均匀划分，从而导致“均匀效应”；2）McIB 算法在非平衡数据上的实验效果明显优于基本 IB 算法，能有效地解决“均匀效应”问题。

结束语 由于非平衡数据的独特构造特性，非平衡数据集的研究是机器学习和模式识别领域一个新兴的且充满挑战的领域。本文根据基本 IB 算法来处理非平衡数据集的局限性，实验展示关于非平衡数据集对 IB 聚类算法影响的分析——IB 聚类算法在处理非平衡数据时有着明显的“均匀效应”。为了减小聚类结果受“均匀效应”的影响，本文提出了一种针对非平衡数据特性的多簇信息瓶颈算法（McIB）。在该算法中，首先基于非平衡数据的倾斜率，提出了一种确定多簇

数目的方法；然后对数据进行初步的聚类分析；最后，一个相似性度量测度被提出并被应用于组织多个较小的簇去代表每个实际的簇。实验分析展示本文提出的 McIB 算法能够有效解决“均匀效应”的影响，从而有效地挖掘非平衡数据集中的聚类模式；同时相比于其他聚类算法，McIB 算法在非平衡数据集上的数据分析性能在整体上表现更优。

参 考 文 献

[1] He H, Garcia E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284

[2] Longadge R, Dongre S. Class Imbalance Problem in Data Mining: Review[J]. International Journal of Computer Science and Network, 2013, 2(1): 83-87

[3] Chawla Nitesh V. Data mining for imbalanced datasets: An overview[M]// Data Mining and Knowledge Discovery Handbook, 2005. US: Springer, 2005: 853-867

[4] Provost F. Machine learning from imbalanced data sets 101 [C]// Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets, 2000. 2000: 1-3

[5] Zhi Wei-mei, Guo Hua-ping, Fan Ming, et al. Disscussion on classfication for imbalanced Data sets[J]. Computer Science, 2012, 39(S1): 304-308(in Chinese)
职为梅，郭华平，范明，等. 非平衡数据集分类方法探讨[J]. 计算机科学，2012，39(S1)：304-308

[6] Kumar C N S, Rao K N, Govardhan A, et al. Imbalanced K-Means: An algorithm to cluster imbalanced-distributed data[J]. International Journal of Engineering and Technical Research, 2014, 2(2): 114-122

[7] Jain A K, Dubes R C. Algorithms for clustering data[M]. Englewood Cliffs: Prentice hall, 1988

[8] Nguyen C H, Ho T B. An imbalanced data rule learner [C]// Knowledge Discovery in Databases: PKDD 2005. Berlin: Springer, 2005: 617-624

[9] MacQueen J. Some methods for classification and analysis of multivariate observations [C]// Proceedings of Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967. Oakland: CA, 1967: 281-297

[10] Abolkarlou N A, Niknafs A A, Ebrahimpour M K. Ensemble imbalance classification: Using data preprocessing, clustering algorithm and genetic algorithm[C]// 2014 4th International Conference on Computer and Knowledge Engineering (ICCKE). IEEE, 2014: 171-176

[11] Yen S-J, Lee Y-S. Cluster-based under-sampling approaches for imbalanced data distributions[J]. Expert Systems with Applications, 2009, 36(3): 5718-5727

[12] Zhang Y P, Zhang L N, Wang Y C. Cluster-based majority under-sampling approaches for class imbalance learning[C]// 2010 2nd IEEE International Conference on Information and Financial Engineering (ICIFE). IEEE, 2010: 400-404

[13] Xiong H, Wu J, Chen J. K-means clustering versus validation measures: a data-distribution perspective[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2): 318-331

[14] Liang J, Bai L, Dang C, et al. The-Means-Type Algorithms Versus Imbalanced Data Distributions[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(4): 728-745

[15] Qian J, Saligrama V. Spectral clustering with imbalanced data [C]// 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 3057-3061

[16] Prachuabsupakij W, Soonthornphisaj N. Cluster-based sampling of multiclass imbalanced data [J]. Intelligent Data Analysis, 2014, 18(6): 1109-1135

[17] Tishby N, Pereira F C, Bialek W. The information bottleneck method[C]// Proceedings of 37th Allerton Conference on Communication, Control and Computering, 1999. 1999: 368-377

[18] Cover T M, Thomas J A. Elements of information theory[M]. New York: John Wiley & Sons, 2012

[19] Slonim N. The information bottleneck: Theory and applications [D]. Jerusalem: The Hebrew University of Jerusalem, 2002

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

[21] Lou Zheng-zheng, Yang Chen, Ye Yang-dong. An IB algorithm based on data selection model[J]. Acta Electronica Sinica, 2014, 42(9): 1839-1846(in Chinese)
娄铮铮，杨晨，叶阳东. 基于数据选择模型的 IB 算法[J]. 电子学报，2014，42(9)：1839-1846

[22] DeGroot M H, Schervish M J, Fang X, et al. Probability and statistics[M]. MA: Addison-Wesley Reading, 1986

[23] Alcala-Fdez J, Sanchez L, Garcia S, et al. KEEL: a software tool to assess evolutionary algorithms for data mining problems[J]. Soft Computing, 2009, 13(3): 307-318

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

[25] Yang Y. An evaluation of statistical approaches to text categorization[J]. Information Retrieval, 1999, 1(1/2): 69-90

[26] Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. Cambridge: Cambridge University Press, 2008

(上接第 233 页)

[10] Mariano F, Zheng Yuan. Grammatical error correction using hybrid systems and type filtering[C]// Proceedings of the 18th Conference on Computational Natural Language Learning. Baltimore, July 2014: 15-24

[11] Collins M. Modeling L. Course notes for NLP by Michael Collins [D]. Columbia University, Spring 2013

[12] Pauls A, Klein D. Faster and Smaller N-Gram Language Models [C]// Proceeding of 49th Annual Meeting of the ACL. Portland, Oregon, June 2011: 258-267

[13] Kneser R, Ney H. Improved backing-off for M-gram language modeling[C]// Proc. ICASSP. 1995: 181-184

[14] Craswell N. Mean Reciprocal Rank[M]. Springer US: Encyclopedia of Database Systems, 2009: 1703