DOI:10.7544/issn1000-1239.2014.20130572
51(10):2148-2159,2014

基于主动学习的模式类别挖掘模型

郭虎升[1]　　王文剑[1,2]
[1](山西大学计算机与信息技术学院　太原　030006)
[2](计算智能与中文信息处理教育部重点实验室(山西大学)　太原　030006)
(guohusheng@sxu.edu.cn)

A Pattern Class Mining Model Based on Active Learning

Guo Husheng[1] and Wang Wenjian[1,2]
[1](*School of Computer and Information Technology*, *Shanxi University*, *Taiyuan* 030006)
[2](*Key Laboratory of Computational Intelligence and Chinese Information Processing* (*Shanxi University*), *Ministry of Education*, *Taiyuan* 030006)

Abstract　In practical applications, there are a lot of data mining problems with unknown class information for the diversity, fuzziness and complexity of objective world. However, traditional methods are generally based on the categories of data which is known before mining, while there are no effective methods for solving this kind of problems. To solve this kind of problems, this paper presents a pattern class mining model based on active learning, namely PM_AL. Firstly, by the difference measurements between the unlabeled samples and labeled samples, some samples are selected as the most valuable samples according to active learning technique. Then these valuable samples are labeled by experts, and the model quickly mines pattern classes implicated in unlabeled samples. Hence, the most valuable samples will be extracted, and the model can quickly mine pattern categories implicated in unlabeled samples. Therefore, a non-labeling multi-class problem can be transferred into a labeling multi-class problem with the very low labeling cost. Through active learning during initial classes mining, the proposed PM_AL model can obtain high learning efficiency, low labeling cost and good generalization performance. The experiment results demonstrate that PM_AL model can effectively find categories as many as possible and solve the large scale multiple classification problems with unknown categories.

Key words　pattern class mining; active learning; PM_AL model; discrepancy; labeling cost

摘　要　在实际应用问题中,由于客观世界物质的多样性、模糊性和复杂性,经常会遇到大量未知样本类别信息的数据挖掘问题,而传统方法往往都依赖于已知样本类别信息才能对数据进行有效挖掘,对于未知模式类别信息的多类数据目前还没有有效的处理方法.针对未知类别信息的多类样本挖掘问题,提出了一种基于主动学习的模式类别挖掘模型(pattern class mining model based on active learning, PM_AL)来解决未知类别信息的模式类别挖掘问题.该模型通过衡量已得到的模式类别与未标记样本间的关系,引入样本差异度的方法来抽取最有价值样本,通过主动学习方式以较小的标记代价快速挖掘无标记样本所蕴含的可能模式类别,从而有助于将无类别标记的多分类问题转化成有类别标记的多分类问

收稿日期:2013-04-15;**修回日期**:2013-06-19
基金项目:国家自然科学基金项目(60975035,61273291);山西省回国留学人员科研资助项目(2012-008);中国民航大学省部级科研机构开放基金项目(CAAC-ITRB-201305)
通信作者:王文剑(wjwang@sxu.edu.cn)

题.实验结果表明,PM_AL 算法能够以较小的标记代价处理无类别信息的模式类别挖掘问题.

关键词 模式类别挖掘;主动学习;PM_AL 模型;差异度;标记代价

中图法分类号 TP18

在实际应用问题中,由于客观世界物质的多样性、模糊性和复杂性,经常会遇到大量未知类别信息的多类数据挖掘问题.例如,据中国互联网络信息中心(China Internet Network Information Center, CNNIC)发布的《第 31 次中国互联网络发展状况统计报告》显示,截止 2012 年底,中国网页数量达到了 1227 亿个[1].要从大规模的网页中找到用户所需信息,自然需要对网页进行自动识别和分类,但考虑到经济成本及操作可行性的问题,只有很小一部分网页能够被权威专家进行正确的类别标记而建立小规模的数据库,大多数网页都无法获得其类别标签而游离于带标签的网页数据库之外,但是,由于网页规模的海量性和动态性,往往无法确定已标记网页的模式类别是否包含了全部的网页模式类别,如通过专家标记得到的网页数据库中包含有政治、经济和娱乐 3 个类别,但无法确认未标记网页中是否包含有教育甚至其他类别的网页.此外,疾病检测本质上也是无类别信息的多类数据挖掘问题,即可通过各种仪器检测到病人体征,从而根据已知的医学信息和知识确定病人病情,但若一个新病人测量所得体征不同于以往任何病例统计体征时,就需要考虑该地区是否有新的疫情而导致其他类别的疾病,即进行新疾病的及时检测.在实际问题中,将无类别信息的多类样本让专家进行一一标注是不现实的,主要因为:1)专家标注的经济代价往往较大,对于大规模问题无法进行有效标注;2)有些问题本身不适合进行大规模标注,如故障检测(特别是军事上的诸如哑炮检测等)问题,不能通过实际操作进行大规模标注;3)如果抽样进行标注,专家很难直接选择重要的标注样本;4)在线标注任务中,专家无法直接选择标注时机.因此在无类别的多类实际问题中,根据样本本身的特征设计高效的自动模式类别挖掘算法对于解决无类别信息的多类问题是非常必要的.

尽管已经有学者提出了多种解决已知样本类别信息的多分类模型[2-5],如一对多模型、一对一模型、有向无环图模型、全局最优化模型等.这些方法虽然能够有效解决传统有标记的多分类问题,但对于未知样本类别信息的多类数据挖掘问题,目前还没有很好的解决方式.因此,进行有效的样本模式类别挖掘,有望将未知样本类别信息的多类数据挖掘问题转化为已解决的已知样本类别的多分类问题.尽管聚类分析技术也可以用来获得数据本身类别,即实现数据集划分后类别内部具有较高的相似性,不同类别间样本具有较高的相异性,但聚类有效性的评价以及对于多类非平衡样本的处理都是困难问题,因此聚类技术更适用于数据分析,其学习过程本质上是实现数据集本身的划分.对于未知样本类别的多类数据挖掘问题,人工干预将是一个更好的选择.主动学习方法[6]正是一类采用专家干预的方式,选择最有价值样本进行标记的学习方法,因此,采用主动学习方法,可以自动地选择含有信息量大的最有价值样本加入到训练集中,有望以较小的标记代价得到较好的模式类别挖掘结果.在得到样本模式类别的信息后,无类别信息的数据挖掘问题就转换成传统已知模式类别的数据挖掘问题.

针对未知类别信息的多类样本数据挖掘问题,本文提出了一种新的基于主动学习的模式类别挖掘模型(pattern class mining model based on active learning, PM_AL).该模型采用主动学习思想来解决未知类别信息的多类问题,即首先定义样本的差异度,选择差异度最高的重要信息样本(即最有可能含有不同类别信息的样本)交由专家进行标记,有效处理无类别信息的多类问题,能够较快地挖掘到无类别信息样本集中蕴含的模式类别,以将无标记的多类问题通过较小的标记代价转换成有标记的多类问题.

1 主动学习

主动学习(active learning)是针对大规模无标记样本的一类专家干预的半监督高效机器学习方法[6],目前主要用于解决无类别信息的二分类问题,其基本思想是依据某种准则来选择含有分类信息较多的最有价值样本,并将其交由专家标记,然后将标记过的样本放入训练集从而得到新的分类器,重复该过程直到满足某种停止条件为止.目前,主动学习模型已成为机器学习研究领域的一个热点模型,并在图像分类[7]、目标检测[8]、文本分类[9]等领域得到

了成功的应用.

主动学习[10]的关键问题是最有价值样本的抽取,因为最有价值样本决定了主动学习模型的泛化性能和学习效率.目前,已有学者设计了多种最有价值样本的抽取策略. Tong[11]提出了基于主动学习的支持向量机方法,即每次迭代选取距离当前近似最优超平面最近的样本作为最有价值样本进行标记;Beygelzimer 等人[12]通过构造样本重要性权值来进行最有价值样本的提取;Schohn 等人[13-14]认为当前分类器中最不能确定类别的样本最有价值,对这些样本进行采样来构造信息样本集,设计了 Uncertainty Sampling 等采样方法.在多数数据集上,这些智能采样方法所得到的学习器都优于随机采样方法,但算法收敛的速度较慢,容易产生孤立点;Seung 等人[15-17]提出了委员会投票方法,该方法利用多个分类器组成委员会,对样本进行投票,认为票数最不一致的样本为最有价值样本;Freund 等人[18]研究证明评委会投票方法的收敛速度比 Uncertainty Sampling 方法快,但也存在孤立点问题,泛化性能不稳定;Saad 等人[19]结合神经网络提出了主动神经网络委员会投票模型,通过无监督学习与有监督学习的结合,较好地解决了信息样本标记问题;Dwyer 等人[20-21]结合集成学习理论,将决策树思想应用于主动学习问题,有效地提高了泛化能力和学习效率.此外,文献[22]结合概率模型,提出了主动贝叶斯网络分类器,在少量标记样本的情况下获得了较好的分类精度;文献[23]结合未标记样本与超平面的距离及未标记样本与已标记样本的关系,设计了置信度和平衡度来选择信息样本,从而进一步增强了主动学习器的收敛速率.

虽然以上各种主动学习模型在一定程度上提高了学习效率,改进了学习器的泛化性能.但上述主动学习模型大多是针对二分类问题的.对于网页标注、图像识别等无类别信息的多类数据挖掘问题,如何能够以最小的标记代价获得近似精确的类别标签,目前未见公开的报道.将主动学习思想应用到无类别信息的样本多分类问题,通过抽取最有价值样本进行半监督学习,有望以较小的标记代价高效地挖掘得到样本的模式类别.

2 基于主动学习的模式类别挖掘模型

由于传统的数据挖掘方法大多依赖于数据已知的类别信息,对于未知类别信息的数据挖掘问题,类别个数及类别性质在数据挖掘之前无法确定,因此需要一种能够挖掘模式类别的方法.当然,解决这种问题最简单、最直接的办法可以通过随机取样的方法进行样本标记,得到样本的类别信息,但随机取样的方法往往挖掘效率低、标记代价大,特别地,对于各类别样本规模分布不均匀的数据,标记代价会急剧增加;其次也可以通过无监督的聚类技术进行初始聚类,然后抽取每个聚类中的重要样本点进行标记,但由于聚类算法本身的瓶颈,依然无法以最小的标记代价获得最优的类别效果.本文提出了一种基于主动学习的模式类别挖掘模型,即在定义样本差异度的基础上,采用主动学习思想,根据样本差异度提取未知类别信息样本集中最有价值样本并交由专家标记,采用半监督专家干预的方式挖掘得到样本所蕴含的类别信息,在给出 PM_AL 算法之前,首先给出无标记样本差异度的定义.

定义 1. 样本差异度.假设已标记样本构成的样本集合为 $Label\_set$,无标记样本构成的样本集合为 $Unlabel\_set$,则无标记样本 x_j 的差异度定义如下:

$$Discrepancy(x_j)=\min_{x_s\in Label\_set} d(x_j,x_s),\quad x_j\in Unlabel\_set, \tag{1}$$

其中,$d(x_j,x_s)$为样本 x_j 与 x_s 间的距离.

如果未标记样本集中的样本差异度较大(超出了给定的样本差异度阈值)时,表明该样本与已标记样本集中所有标记样本的模式类别差异均较大,因此该样本越可能产生新的模式类别(如图 1 所示).假设图 1 中已标记样本集为 $Label\_set=\{L_i,L_j,L_k\}$,未标记样本集为 $Unlabel\_set=\{u_m,u_n\}$.由定义 1 可知,无标记样本 u_m 的差异度 $Discrepancy(u_m)=d(u_m,L_i)$,无标记样本 u_n 的差异度 $Discrepancy(u_n)=d(u_n,L_k)$,由于 $d(u_m,l_i)<d(u_n,l_k)$,因此,与样本 u_m 相比,在样本 u_n 上更容易挖掘得到新的模式类别.

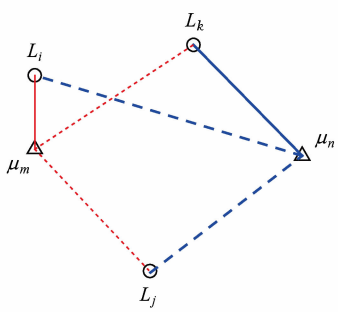

Fig. 1 The relationship between Discrepancy and class.

图 1 差异度与模式类别关系

由于对未标记样本求解差异度时需要用到至少2个已标记的样本,所以本文将PM_AL模型分为2部分:第1部分用于选择初始的2个标记样本并由专家标记;第2部分用于挖掘蕴含在未标记样本集中的其他模式类别.

PM_AL模型的具体执行过程如算法1所示.算法中 $Label\_set$ 表示已标记的样本集,$Unlabel\_set$ 为未标记的样本集,$Class\_set$ 表示已得到的模式类别集合,$\overline{Step}$ 为标记样本间隔阈值参数,以控制模式类别挖掘的进程.显然,算法中 $\overline{Step}$ 值的设置直接影响着挖掘得到的模式类别个数,也影响到最终的学习效果,在实验部分,将对标记样本间隔阈值参数 $\overline{Step}$ 作详细讨论.PM_AL具体算法如下.

算法1. PM_AL算法.

初始化:给定样本集 $X=\{x_i\}_{i=1}^{l}$,设置初始标记样本集 $Label\_set=\varnothing$,初始未标记样本集 $Unlabel\_set=X$ 及初始模式类别集 $Class\_set=\varnothing$,标记样本间隔阈值 $\overline{Step}$ 的值.

1) 初始信息样本提取

Step1. 计算未标记样本集 $Unlabel\_set$ 中所有样本的中心 μ:

$$\mu=\frac{1}{l}\sum_{p=1}^{l}x_p.$$

Step2. 提取第1个样本:

Step2.1. $x_1^*=\arg\max_{x_s\in Unlabel\_set} d(x_s,\mu)$;

Step2.2. 把样本 x_1^* 交由专家进行标记,并获得其模式类别 y_1^*;

Step2.3. $Label\_set=Label\_set\cup\{x_1^*\}$, $Unlabel\_set=Unlabel\_set\backslash\{x_1^*\}$;

Step2.4. $Class\_set=Class\_set\cup\{y_1^*\}$.

Step3. 提取第2个样本:

Step3.1. $x_2^*=\arg\max_{x_s\in Unlabel\_set} d(x_s,x_1^*)$;

Step3.2. 把样本 x_2^* 交由专家进行标记,并获得其模式类别 y_2^*;

Step3.3. $Label\_set=Label\_set\cup\{x_2^*\}$, $Unlabel\_set=Unlabel\_set\backslash\{x_2^*\}$;

Step3.4. 若 $y_2^*\neq y_1^*$ 成立,则 $Class\_set=Class\_set\cup\{y_2^*\}$.

2) 其他模式类别挖掘

Step4. 挖掘其他模式类别,得到新的标记样本集 $Label\_set$、未标记样本集 $Unlabel\_set$ 和模式类别集 $Class\_set$:

Step4.1. $Step=0$;

Step4.2. $x_o^*=\arg\max_{x_s\in Unlabel\_set} Discrepancy(x_s)$;

Step4.3. 把样本 x_o^* 交由专家进行标记,并获得其模式类别 y_o^*:

$$Label\_set=Label\_set\cup\{x_o^*\};$$
$$Unlabel\_set=Unlabel\_set\backslash\{x_o^*\}.$$

Step4.4. 若 $y_o^*\in Class\_set$

$Step=Step+1$;

否则

$\{Step=0, Class\_set=Class\_set\cup\{y_o^*\}\}$.

Step4.5. 若 $Step\leqslant\overline{Step}$

返回 Step4.2.

Step5. 算法结束并得到最终的标记样本集 $Label\_set$,未标记样本集 $Unlabel\_set$ 和模式类别集 $Class\_set$.

为测试本文提出的PM_AL算法的性能,将其与目前最常用的基于随机选择的模式类别挖掘算法(pattern class mining based on random method, PM_R)和基于聚类的模式类别挖掘算法(pattern class mining based on clustering method, PM_C)进行了对比(为便于比较,对聚类之后距离类中心最近的样本点进行标注).基于随机样本选择的模式类别挖掘方法如下.

算法2. PM_R算法.

初始化:给定初始样本集 $X=\{x_i\}_{i=1}^{l}$,设置初始标记样本集 $Label\_set=\varnothing$,初始未标记样本集 $Unlabel\_set=X$ 及初始模式类别集 $Class\_set=\varnothing$,标记样本间隔阈值 $\overline{Step}$ 的值.

Step1. 从初始未标记样本集 $Unlabel\_set$ 中随机选择一个样本 x_i.

Step2. 对样本 x_i 进行标记,并获得其模式类别 y_i:

$$Label\_set=Label\_set\cup\{x_i\};$$
$$Unlabel\_set=Unlabel\_set\backslash\{x_i\}.$$

Step3. 若 $y_i\in Class\_set$

$Step=Step+1$;

否则

$\{Step=0, Class\_set=Class\_set\cup\{y_i\}\}$.

Step4. 若 $Step\leqslant\overline{Step}$

返回 Step1.

Step5. 算法结束并得到最终的标记样本集 $Label\_set$,未标记样本集 $Unlabel\_set$ 和模式类别集 $Class\_set$.

基于聚类的模式类别挖掘算法如下.

算法 3. PM_C 算法.

初始化:给定样本集 $X=\{x_i\}_{i=1}^{l}$,初始模式类别集 $Class\_set=\varnothing$,并设定算法初始聚类个数 k,标记样本间隔阈值$\overline{Step}$的值.

Step1. 对为标记样本集 X 进行初始聚类,聚类个数为 k,即 $X\rightarrow\{X_1,\cdots,X_i,\cdots,X_k\}$,为简化问题,本文采用传统的 k 均值聚类.

Step2. 计算聚类结果中每个类 X_i 的类中心 μ_i.

Step3. 计算样本集 X 第 i 类的每个样本点 x_{i_j} 与其所属类中心 μ_i 的距离 $d(x_{i_j},\mu_i)$,并选择信息样本 x_i,使得 $x_i^*=\arg\min_{x_{i_j}\in X_i} Discrepancy(x_{i_j},\mu_i)$,构成信息样本集 $X^*=\{x_i\}_{i=1}^{k}$.

Step4. 对样本集 X^* 进行标记,并获得其中每个信息样本的模式类别 y_i^*,构成第 t 次的类别标记集合 $Y^{*,t}$.

Step5. 若 $Y^{*,t}=Y^{*,t-1}$,则 $Step=Step+1$;否则$\{Step=0,Class_set=Class_set\bigcup Y^{*,t}\}$.

Step6. 若 $Step\leqslant\overline{Step}$,$k+1\rightarrow k$,返回 Step1,否则算法结束,输出模式类别集合 Y^*.

假设样本集的规模为 l,初始信息样本提取的复杂度仅为 $O(l)$,而在其他模式类别挖掘过程中,假设需要迭代的次数为 $m(m<l)$,其中第 i 次迭代的复杂度为 $O(l-i-1)$,因此,整个算法的复杂度为 $O\left(\frac{(2lm-m^2-3)}{2}\right)+O(l)=O(lm)$;而 PM_C 算法仅聚类预处理过程的复杂度就为 $O(l^2)$,因此,PM_AL 算法的复杂度要低于 PM_C 算法.尽管由于 PM_R 算法直接进行随机取样和打标记而具有更低的复杂度,但其标记代价过高,无法满足无类别标签的实际应用问题.

3 实验及结果分析

为验证本文提出的 PM_AL 算法的有效性,本文从类别挖掘标记代价、标记样本间隔阈值参数影响及算法鲁棒性 3 个方面进行了验证.由于无类别标记的多分类问题的模拟本身存在一定的困难,本文采用实验采用 UCI 上的带标签的多类别数据集[24]进行模拟,其中通过主动学习过程挑选的最有价值样本的真实标记作为专家标签,实验数据集如表 1 所示.实验在 1 台 PC 机(2.66 GHz CPU,1 GB 内存)上进行测试,实验平台是 Matlab7.0.

Table 1 Benchmark Datasets Used in Experiments

表 1 实验采用的标准数据集

Datasets	Size	# Features	# Classes
Balance_scale	125	4	5
Glass	100	10	6
Iris	50	4	3
Letter	2 000	16	26
Machine	100	7	7
Page_blocks	1 000	10	5
Segment	1 000	19	6
Vehicle	300	18	4
Vowel	220	10	15
Wine	78	13	3

3.1 模式类别挖掘代价实验

为测试本文提出的 PM_AL 算法能否快速地找到无类别信息样本中所蕴含的模式类别,实验中将本文提出的 PM_AL 算法与基于聚类的 PM_C 算法和基于随机采样的 PM_R 算法进行对比.由于 2 个样本的差异度是定值,样本的聚类结果也是定值,所以 PM_AL 和 PM_C 算法的结果都是确定的,而基于随机采样的模式类别挖掘结果是不确定的,所以每组随机模式类别挖掘实验进行 5 次,并取其中挖掘全部模式类别所需要采样样本个数介于中间的实验结果作为参照进行对比.图 2 为在标准数据集上得到的挖掘结果,横轴表示当前挖掘过程的标记代价(标记样本个数),纵轴表示当前挖掘得到的模式类别个数.

从图 2 可以看出,随着标记代价的增加,PM_AL 和 PM_R 算法挖掘得到的模式类别个数也在增加,而在个别数据集上,PM_C 算法存在微小震荡,这是因为 PM_C 算法采用的聚类算法本身的不稳定性而形成的.除数据集 Letter 外,PM_AL 算法抽取最大类别的标记代价均低于或等于 PM_C 算法,且在数据集 Letter 上,PM_AL 算法抽取最大类别的标记代价仅略大于 PM_C 算法,在所有数据集上,PM_AL 算法抽取最大类别的标记代价均低于或等于 PM_R 算法,可通过计算得到 PM_AL 算法在不同数据集上的平均标记代价为 18.9,PM_C 算法的平均标记代价为 28.9,PM_R 算法的平均标记代价为 48.7.此外,从图 2 也不难看出,在模式类别挖掘的过程中,即挖掘得到的类别数小于各数据集的最大类别个数的过程中,PM_AL 算法也比其他两种方法在标记代价方面具有明显的优势.该结果

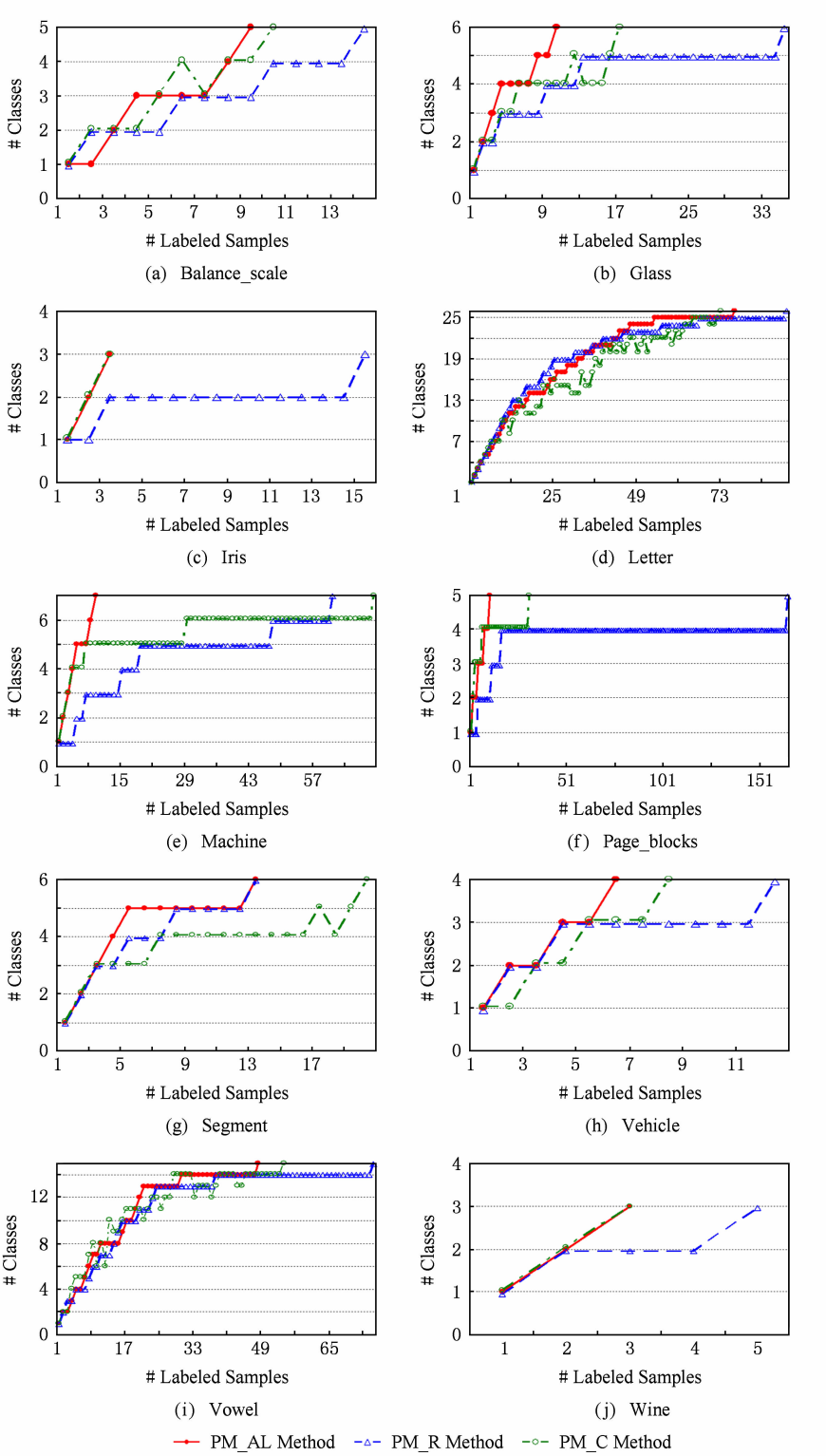

Fig. 2 Class mining results on benchmark datasets.

图 2 标准数据集上的类别挖掘结果

充分说明 PM_AL 算法能够有效地根据样本空间中样本分布的情况进行差异度的测量,从而选择差异度大的样本进行模式类别挖掘,以快速得到蕴含在无标记数据集中的模式类别.

本文还对挖掘每个模式类别所需要的标记代价(标记样本个数)进行了统计,3 种算法的对比实验结果如图 3 所示,横轴表示当前挖掘到的模式类别个数,纵轴表示挖掘该模式类别过程中所需要标注的样本个数.可以看出,挖掘每个新的模式类别所需要标记的样本数是不同的.多数情况下,模式类别挖

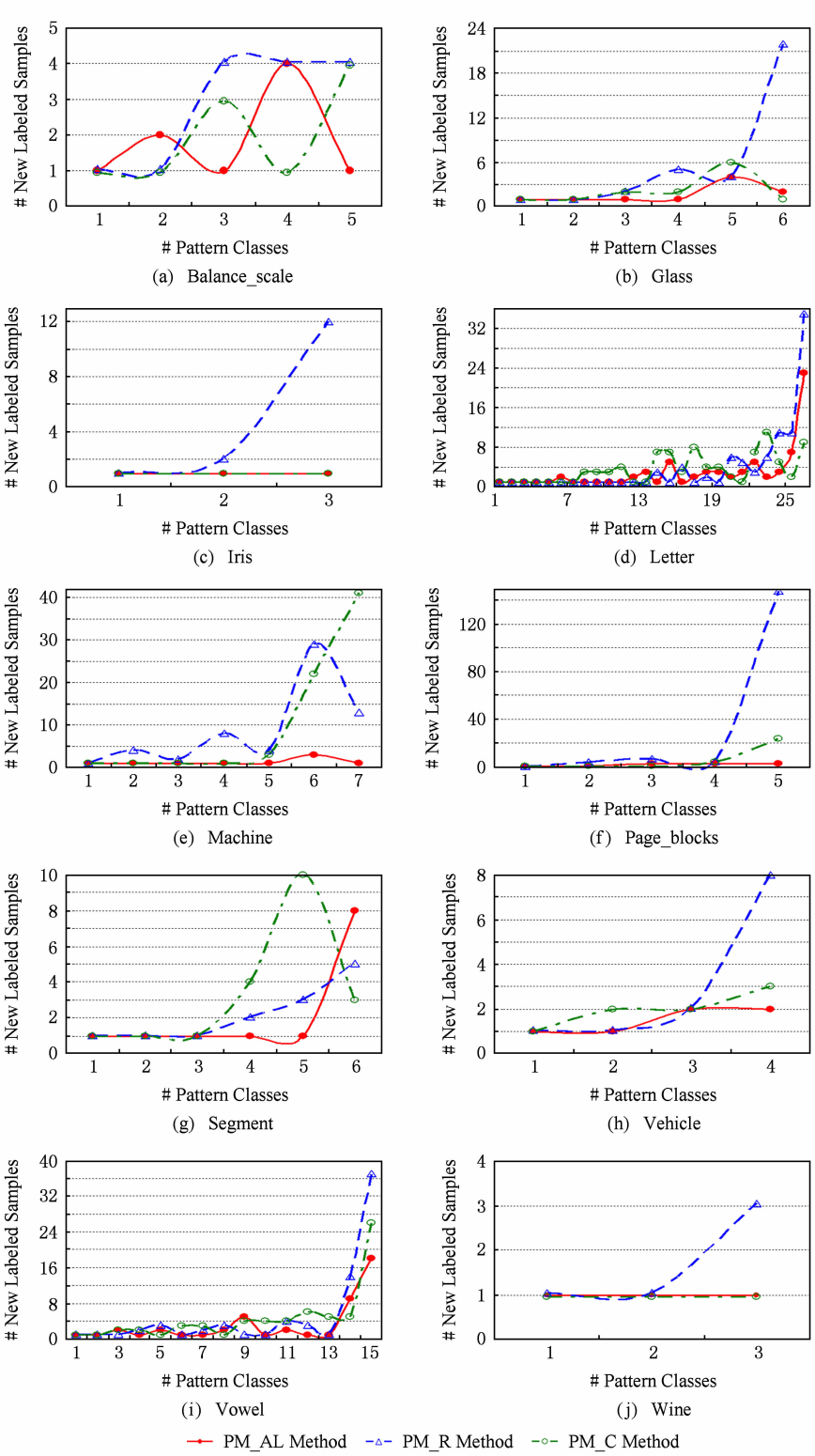

Fig. 3 New labeled samples for each class mining.

图 3 每个模式类别挖掘所需要的标记样本数

掘的起始挖掘阶段需要的标记样本较少,而到了模式类别挖掘过程的后期,尤其是当剩余未标记样本只含有少数未挖掘到的模式类别时,反而需要大量的样本标记来得到剩余模式类别. 多数情况下,PM_AL 算法的变化趋势要比 PM_C 和 PM_R 算法平缓得多. 在其中的 7 个数据集上,PM_AL 在每个类别挖掘的标记样本数都没有超过 4,且最大的标记样本数也仅仅达到 23(数据集 Letter 的最后一个类

别).而 PM_R 算法只有 2 个数据集上每个类别挖掘的标记样本数没有超过 4,且其挖掘过程中个别类别挖掘的最大标记样本数达到了 150(数据集 Page_blocks 的最后一类);PM_C 算法只有 4 个数据集上每个类别挖掘的标记样本数没有超过 4,且其挖掘过程中个别类别挖掘的最大标记样本数达到了 41(数据集 Machine 的最后一类),且明显可看出 PM_C 算法挖掘过程中每个类别所需要的标记代价变化不稳定.但是,在数据集 Segment 上,PM_AL 算法的标记样本数反而比 PM_R 还要多,这可能是由于该数据集样本类别的分布不能采用欧氏距离来衡量,即存在大量样本的欧氏距离很大但却属于同一类别的特殊现象,因此该数据集不适用于基于欧氏距离的的样本差异度度量.

由于 PM_AL 算法充分考虑了未标记样本与已挖掘得到的模式类别间的差异程度,选择差异程度大的样本进行模式类别挖掘,更容易得到剩余模式类别,提高了模式类别挖掘效率;基于随机采样的模式类别挖掘方法具有随机性,不能够有效地利用未标记样本与已挖掘模式类别间的有用信息,挖掘效率较低;基于聚类的方法尽管在一定程度上有了提高,但其挖掘模式类别所需要的标记代价依然较高,且挖掘效果不稳定.

3.2 阈值参数$\overline{Step}$优化

从算法 1 可以看出,标记样本间隔阈值参数 $\overline{Step}$ 对于模式类别挖掘结果有重要影响.因此,如何设置合理的标记样本间隔阈值参数是一件非常重要也存在困难的事情.如果 $\overline{Step}$ 值太小,那么模式类别挖掘过程可能不彻底,一些未知的样本模式类别可能得不到,若 $\overline{Step}$ 太大,模式类别挖掘效率下降,同时挖掘所需要的标记代价也大幅度上升,这与主动学习的基本目标相违背.因此,本文专门对标记样本间隔阈值参数 $\overline{Step}$ 的影响进行了实验.

图 4 给出了阈值参数 $\overline{Step}$ 对算法的影响比较,其中横轴为标记样本间隔阈值参数值,纵轴为所能挖掘得到的样本个数.可以看出,只要 $\overline{Step}$ 值设置得足够大,2 种方法均能得到所有蕴含的模式类别.在其中 8 个数据集上(Letter 和 Segment 除外),PM_AL 算法挖掘样本所蕴含的全部模式类别所需要设置的 $\overline{Step}$ 值均小于或等于 PM_R 和 PM_C 算法.从另一角度看,在相同的 $\overline{Step}$ 参数下,PM_AL 算法在多数数据集上所得到的模式类别个数均大于其他 2 种算法.因此,实验结果充分说明,PM_AL 算法在较小的 $\overline{Step}$ 值下就可以得到较好的模式类别挖掘结果.

3.3 模式类别挖掘的鲁棒性

由于在大多数数据集上,不同类样本的规模大小通常是不一样的,样本分布也往往是不均匀的,如在数据集 Machine 上只有一个样本属于第 6 类.因此,若采用随机方法挖掘模式类别,那些包含样本规模较大的类别可能会被多次标记,而含有极少数样本的类别则很难被提取,因此类别挖掘所需要的标记代价较高,挖掘效率较低.PM_AL 算法充分考虑到已标记样本与未标记样本的差异性,而不是考虑每类样本的规模,因此那些含有极少数样本的类别也有与含有多数样本的类别同样的概率被挖掘到.本文在 3 个典型的非平衡多类数据集 Glass,Machine 和 Page_blocks 上进行了验证.图 5 是以上 3 个数据集每类样本规模的统计结果,其中横轴为类别标号,纵轴为每类样本的统计个数.图 6~8 分别为采用 PM_AL 算法、PM_R 算法和 PM_C 算法在模式类别挖掘过程中所提取各类样本个数的分布图,图 6 至图 8 中横轴为类别标号,纵轴为各方法所标记每类样本的个数.由于聚类算法本身具有不确定性,因此这里 PM_C 算法统计的是每次迭代过程中各个类别的标记样本个数之和.

从图 5 可以看出,数据集 Glass 的第 1,2,6 类数据要明显多余其他类数据,而数据集 Machine 和 Page_blocks 中,均存在其中一类远多于其他类的非平衡情况.若采用随机挖掘方法,抽取的样本分布完全取决于样本各类别的个数分布,这一点在图 7 的 Machine 和 Page_blocks 两个数据集上体现得尤为明显.在这 3 个数据集上,采用随机挖掘方法所挖掘得到的多类样本分别占到总样本的比例为 70.6%,68.3%和 96.3%,而原始数据集中规模较大类样本所占的比例分别为 65.5%,67%与 88.8%.这充分说明采用 PM_R 算法时,多类样本更容易被抽取,然而这些样本对于已经挖掘得到该类信息的挖掘过程没有意义,但却导致了收敛速度缓慢,标记代价上升.对于数据集 Machine 和 Page_blocks,PM_C 算法的统计结果与 PM_R 算法具有类似的特点,这说明尽管 PM_C 算法依然在较大程度上受到样本不同类别个数分布不均衡的影响,但在数据集 Glass 上,PM_C 算法得到的各类样本的个数与原始样本集中各类样本个数不一致,这可能是由于该类样本中不同类样本分布的疏密不同,从而导致聚类结果中较大规模样本分布集中(聚类个数少),而较小规模样本聚类个数反而可能多导致的.PM_AL 算法

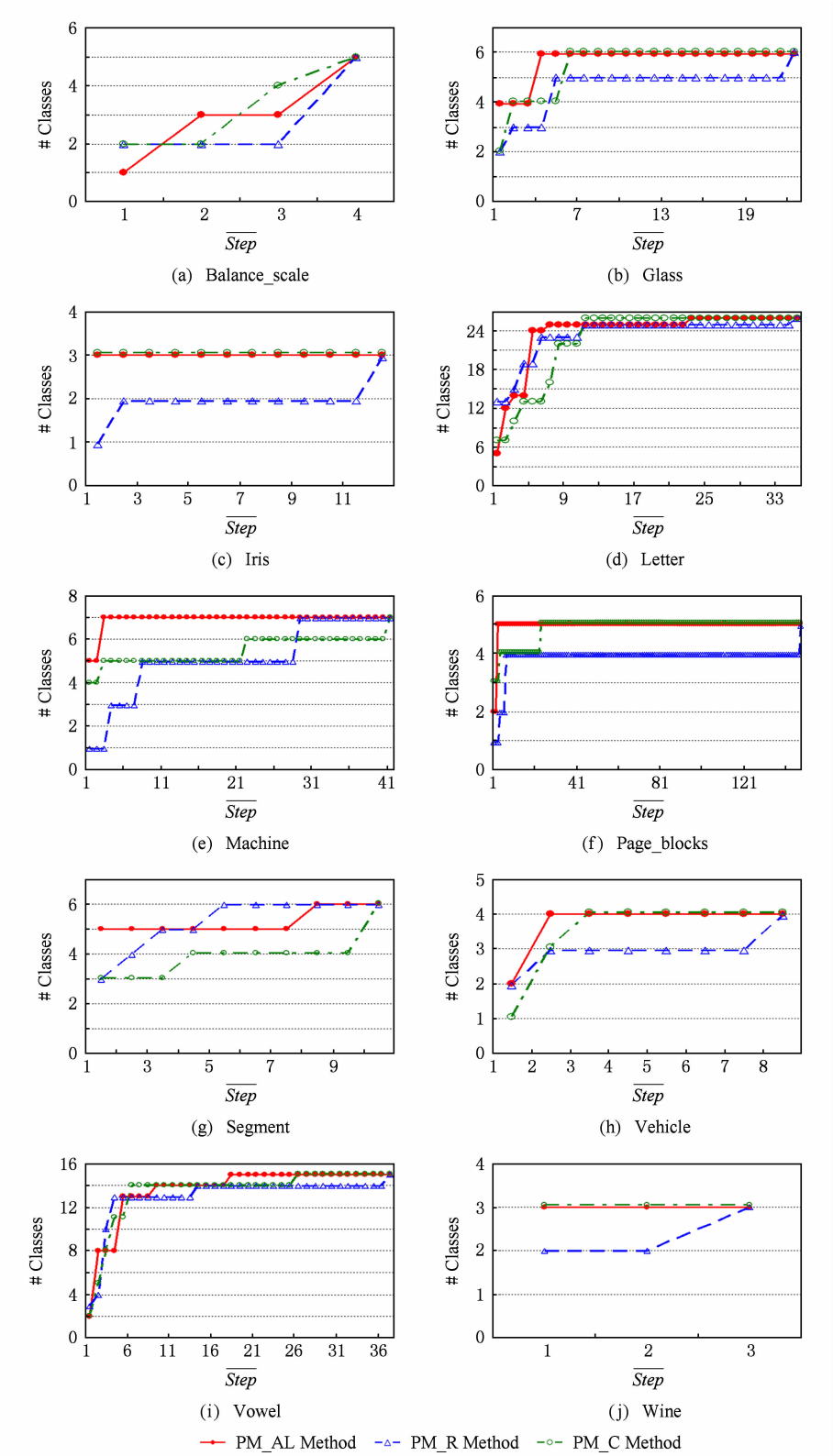

Fig. 4 Influence of parameter $\overline{Step}$ on mined classes.

图 4 标记样本间隔阈值参数$\overline{Step}$的影响

在以上 3 个多类样本规模非平衡的数据集挖掘得到不同类样本的分布几乎是均匀的.因此,PM_AL 算法进行无类别信息样本的模式类别挖掘问题的开销更小,更容易得到新的模式类别.实验结果充分说明 PM_AL 算法是高效且对数据非平衡的规模分布具有鲁棒性.

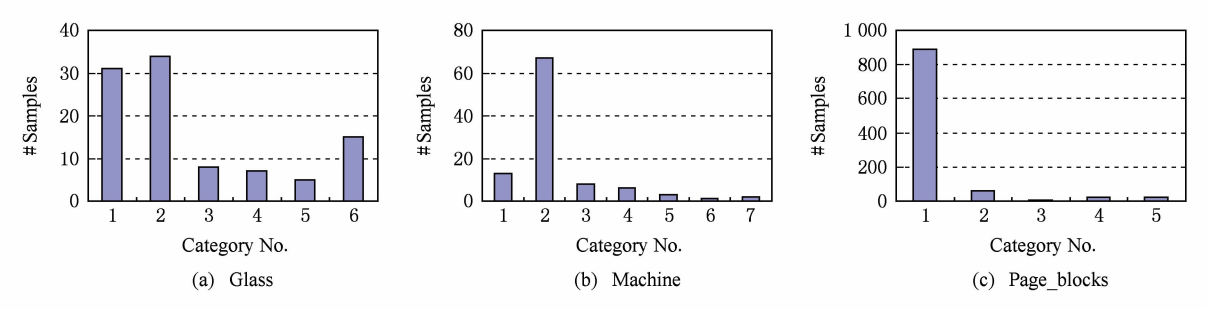

Fig. 5 Different class samples number statistics of imbalanced datasets.

图 5 非平衡数据集各类样本数目统计

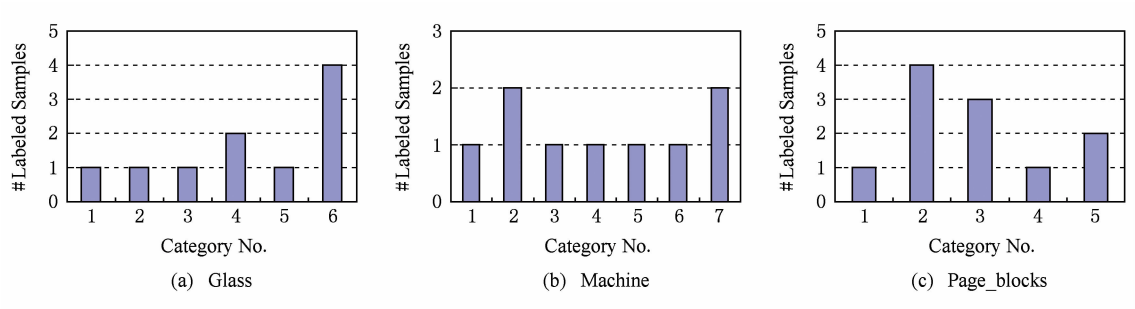

Fig. 6 Distributions of extracted sample during class mining process by PM_AL method.

图 6 PM_AL 算法模式类别挖掘过程中所提取样本的分布

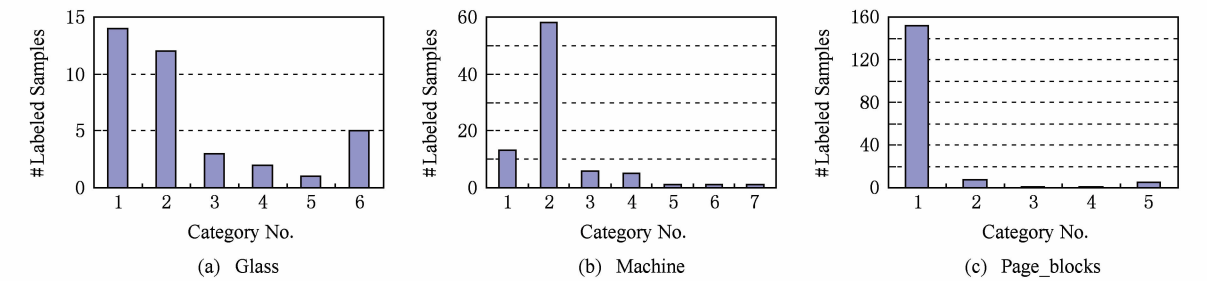

Fig. 7 Distributions of extracted sample during class mining process by PM_R method.

图 7 PM_R 算法模式类别挖掘过程中所提取样本的分布

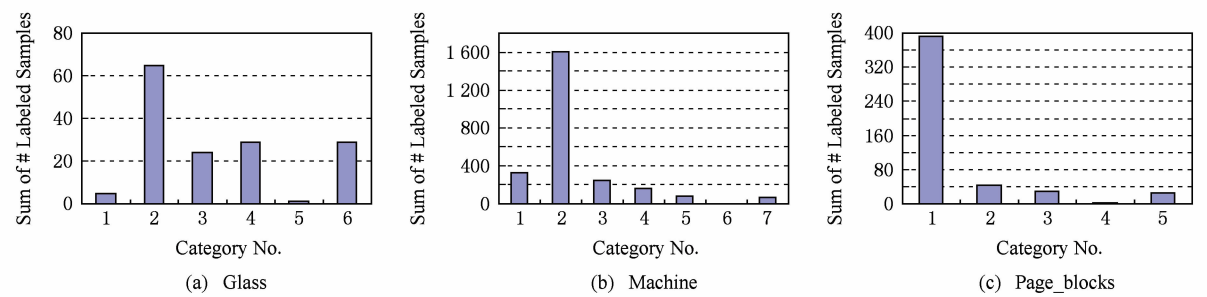

Fig. 8 Distributions of extracted sample during class mining process by PM_C method.

图 8 PM_C 算法模式类别挖掘过程中所提取样本的分布

综合上述的实验结果不难看出,PM_AL 算法能够有效地处理样本类别信息未知的多类数据集的模式类别挖掘问题. 该算法首先定义了样本间的差异度,选择差异度大的样本作为主动学习的最有价值样本进行标记,采用有人工干预的方法以较小的标记代价得到样本中所蕴含的模式类别,从而将无类别信息的多分类问题转化成传统的有类别信息的多分类问题.

4 结束语

由于传统的方法大多都是在已知样本所蕴含的模式类别信息的基础上进行数据挖掘,对于未知模式类别信息数据的挖掘目前常用的如基于聚类的抽样标记法和随机样本抽样标记法的标记代价都较高,不能够对样本模式类别进行有效挖掘.本文针对实际类别未知的数据挖掘中由于标记代价太大而无法直接全部标记的问题,提出了一种新的基于主动学习的模式类别挖掘模型.该模型通过定义无标记样本差异度,快速寻找最有价值的样本进行标记,从而以较小的标记代价快速地找到无标记样本集中所蕴含的模式类别,有效处理未知类别信息的多类问题.

由于未知类别信息的模式类别挖掘是一个不确定的问题,在实际应用中应当根据问题本身进行动态地模式类别挖掘,即根据用户的认知水平在更为抽象(挖掘较少较粗的类)或者更为具体(挖掘较多较细的类)的水平上动态调整算法执行.在未来的工作中:一方面将探讨动态的模式类别挖掘机制,使用户能够在自己认知需求的基础上动态地选择和调整类别挖掘层次;另一方面,把根据 PM_AL 模型进一步设计解决未知类别信息的主动多分类学习问题及在线多分类学习问题的方法,并广泛应用于网页分类、邮件分类、签名检测、疾病分类等实际无类别信息的大规模多类数据挖掘问题当中.此外,如何定义更为广泛的(基于非欧氏距离的)样本差异度,将模型扩展应用于复杂分布的数据集也是值得探讨的一个问题.

参 考 文 献

[1] China Internet Network Information Center (CNNIC). Statistical report of thirty-first China Internet development [DB/OL]. (2013-01-15)[2013-03-15]. http://www.cnnic.net.cn (in Chinese)
(中国互联网络信息中心. 第31次中国互联网络发展状况统计报告[DB/OL]. (2013-01-15)[2013-03-15]. http://www.cnnic.net.cn)

[2] Wu T F, Lin C J, Weng R C. Probability estimates for multi-class classification by pairwise coupling [J]. Journal of Machine Learning Research, 2004, 5, 975-1005

[3] Lin H Y. Efficient classifiers for multi-class classification problems [J]. Decision Support Systems, 2012, 53(3): 473-481

[4] Chang C C, Chien L J. A novel framework for multi-class classification via ternary smooth support vector machine [J]. Pattern Recognition, 2011, 44(6): 1235-1244

[5] Bourke C, Deng K, Scott S D, et al. On reoptimizing multi-class classifiers [J]. Machine Learning, 2008, 71(2/3): 219-242

[6] Dasgupta S. Two faces of active learning [J]. Theoretical Computer Science, 2011, 412(19): 1767-1781

[7] Qi G J, Hua X S, Rui Y, et al. Two-dimensional multi-label active learning with an efficient online adaptation model for image classification [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(10): 1880-1897

[8] Han Guang, Zhao Chunxia, Hu Xuelei. An SVM active learning algorithm and its application in obstacle detection [J]. Journal of Computer Research and Development, 2009, 46(11): 15-20 (in Chinese)
(韩光, 赵春霞, 胡雪蕾. 一种新的SVM主动学习算法及其在障碍物检测中的应用[J]. 计算机研究与发展, 2009, 46(11): 15-20)

[9] Esuli A, Sebastiani F. Active learning strategies for multi-label text classification [G] //LNCS 5478: Proc of the 31th European Conf on IR Research (ECIR). Berlin: Springer, 2009: 102-113

[10] Vlachos A. A stopping criterion for active learning [J]. Computer Speech and Language, 2008, 22(3): 295-312

[11] Tong S. Active learning: Theory and applications [D]. Stanford: Stanford University, 2001

[12] Beygelzimer A, Dasgupta S, Langford J. Importance weighted active learning [C] //Proc of the 26th Int Conf on Machine Learning. New York: ACM, 2009: 49-56

[13] Schohn G, Cohn D. Less is more: Active learning with support vector machines [C] //Proc of the 17th Int Conf on Machine Learning. San Francisco: Morgan Kaufmann, 2000: 839-846

[14] Zhu J B, Wang H Z, Yao T S, et al. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification [C] //Proc of the 22nd Int Conf on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2008: 1137-1144

[15] Seung H S, Opper M, Sompolinsky H. Query by committee [C] //Proc of the 15th Annual ACM Workshop on Computational Learning Theory (COLT). New York: ACM, 1992: 287-294

[16] Ho S S, Wechsler H. Query by transduction [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2008, 30(9): 1557-1571

[17] Abdel Hady M F, Schwenker F. Combining committee-based semi-supervised and active learning [J]. Journal of Computer Science and Technology, 2010, 25(4): 681-698

[18] Freund Y, Seung H S, Samir E, et al. Selective sampling using the query by committee algorithm [J]. Machine Learning, 1997, 28(23): 133-168

[19] Saad E W, Choi J J, Vian J L, et al. Query-based learning for acrospace applications [J]. IEEE Trans on Neural Networks, 2003, 14(6): 1437-1448

[20] Dwyer K, Holte R. Decision tree instability and active learning [G] //LNCS4701: Proc of the European Conf on Machine Learning (ECML). Berlin: Springer, 2007: 128-139

[21] Baezayates R, Hurtado C, Mendoza M. Query clustering for boosting web page ranking [G] //LNCS 3034: Proc of the Int Atlantic Web Intelligence Conf (AWIC). Berlin: Springer, 2004: 164-175

[22] Gong Xiujun, Sun Jianping, Shi Zhongzhi. An active Bayesian network classifier [J]. Journal of Computer Research and Development, 2002, 39(5): 574-579 (in Chinese)
(宫秀军, 孙建平, 史忠植. 主动贝叶斯网络分类器[J]. 计算机研究与发展, 2002, 39(5): 574-579)

[23] Bai Longfei, Wang Wenjian, Guo Husheng. A novel support vector machine active learning strategy [J]. Journal of Nanjing University, 2012, 48(2): 60-67 (in Chinese)
(白龙飞, 王文剑, 郭虎升. 一种新的 SVM 主动学习策略[J]. 南京大学学报 2012, 48(2): 60-67)

[24] Bache K, Lichman M. UCI Machine Learning Repository [EB/OL]. (2009-10-16) [2013-03-21]. http://archive.ics.uci.edu/ml

Guo Husheng, born in 1986. PhD. His main research interests include machine learning and data mining.

Wang Wenjian, born in 1968. PhD, professor. Senior member of China Computer Federation. Her main research interests include machine learning, computing intelligence, etc.