ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 1,Jan. 2016,pp. 152－159
DOI:10.16337/j.1004-9037.2016.01.016

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

一种动态的主动多分类方法

郭金玲[1]　樊东燕[1]　郭虎升[2]

(1.山西大学商务学院信息学院,太原,030031; 2.山西大学计算机与信息技术学院,太原,030006)

摘　要: 在面向大数据问题的应用领域中,由于现实世界的多样性和复杂性,经常会遇到大规模的多类别数据挖掘问题,传统的多分类方法一方面存在着超平面不平衡更新的问题,另一方面学习效率较低,对于复杂的多类别数据无法进行高效分类。针对这个问题,本文提出了一种改进的动态主动多分类(Dynamical active multiple classification, DYA)方法,该方法通过将死锁、激活等概念引入到主动多分类过程,在主动多分类过程中随着分类器的不断更新,动态地控制样本是否参与主动学习的过程;同时,采用分位计数、轮换学习方式的主动多分类方法,使得多类别的分类器能够得到平衡的学习和更新。实验结果表明,本文提出的动态主动多分类方法有效提高了模型的学习效率和泛化性能。

关键词: 主动学习;多分类;动态主动多分类;分位计数;轮换学习

中图分类号: TP181　　**文献标志码:** A

Dynamical Active Multiple Classification Method

Guo Jinling[1], Fan Dongyan[1], Guo Husheng[2]

(1. School of Information, Business College of Shanxi University, Taiyuan, 030031, China; 2. Computer and Information Technology Department, Shanxi University, Taiyuan, 030006, China)

Abstract: In the application of big data theory, there are many large scale multiple classification problems for the diversity and complexity of real world. However, the hyperplane updating of traditional multiple classification methods are not balanced. And the learning efficiency of them are low, and they are not efficient for the complex multiple classification data. To solve this problem, this paper presents an improved dynamical active multiple classification method (DYA). By combining the definitions of deadlock and activation with the active multiple classification process, the proposed method controls dynamically the status whether the sample is to be involved in the active learning process with the updating of classifier in it. Meanwhile, the active learning method with sub-bit counter and rotation learning approach is used to the balance learning and updating of classifier. The experiment results demonstrate that the proposed DYA method can improve both the learning efficiency and generalization performance.

Key words: active learning; multiple classification; dynamical active multiple classification; sub-bit counter; rotation learning

---

基金项目: 国家自然科学基金(61503229)资助项目;山西省自然科学基金(2015021096,2014011018-1)资助项目;山西省高等学校科技创新(2015110)资助项目;山西大学商务学院科研基金(2015009)资助项目。
收稿日期: 2015-07-11;**修订日期:** 2015-10-22

引 言

随着科技进步及人们管理和知识水平的提高,现实世界中需要存储和处理的数据规模越来越大,数据种类越发繁多。2008 年 9 月,《Nature》杂志推出了大数据存储、管理和分析问题讨论的专刊[1],之后《Science》杂志也相继出版了关于大数据的研究报告[2],并围绕科学研究和实际应用领域中的大数据问题进行了深入讨论,阐述了大数据问题处理的重要性和必要性。目前,大数据问题的研究已经成为人工智能乃至整个计算机学科领域的研究热点问题[3]。

在大数据应用领域,由于客观现实世界的多样性和复杂性,经常会遇到许多大规模的多类别数据分类问题,如文本分类[4]、图像分类[5]等。尽管目前已经提出了多种解决多类别数据分类问题的经典方法[6],如一对多方法通过构造分类器将每一类样本与其他类样本分开,这种多分类方法模型简单,但由于所有样本都用于每个子分类器分类,因此学习速度较慢;一对一方法则在每两个类之间都要建立分类器,因此子分类器数目比较多,模型相对复杂;有向无环图方法中分类器构建的过程与一对一方法类似,构造一棵二叉树,并每次从根结点出发寻找一条路径到达叶结点,该叶结点所代表的类别即为对样本的类别预测;全局最优化方法通过构建一个整体的最优化问题进行求解,但当训练数据规模较大时,求解过程会变得非常复杂。尽管目前已经提出了一些成熟的解决多类数据挖掘问题的优化方法,但这些方法往往都存在学习效率较低、泛化性能不够好的问题[7]。

主动学习(Active learning, AL)[8-10]是一种典型的解决大规模数据学习的策略,已经在图像分类[11]、生物信息学[12]许多领域得到成功应用。与传统采用所有训练样本构建分类器的方法不同,基于主动学习思想的机器学习方法往往是根据某些启发式的规则或信息,选择训练样本中少量最有价值的样本参与训练,以循环迭代更新的方式逼近最优分类器。由于在主动学习过程中只有少数最有价值样本参与训练,缩小训练样本规模,减小噪声影响,因此可以提高分类器的学习效率和泛化性能。目前,已经构建了多种基于主动学习的分类方法,如 Tong 等[13]选取与当前超平面最近的样本作为最有价值样本参与训练;Huang 等[14]通过衡量样本不确定性来提取最有价值样本;Abdel 等[15]通过多个子分类器构成投票委员会,对样本进行投票以选取最有价值样本;Saad 等[16]提出了基于神经网络的委员会投票主动学习方法;此外采用概率统计的主动贝叶斯网络分类器[17]近年来也得到广泛研究。

针对传统多分类方法超平面更新不平衡且学习效率较低的问题,本文将主动学习技术应用于多分类问题,提出一种改进的动态主动多分类方法,通过将死锁、复位等概念引入到主动多分类方法,在主动多分类过程中随着分类器的不断更新,动态地控制样本是否参与主动学习的过程;同时,采用分位计数、轮换学习方式的主动多分类方法,使得多类别的分类器能够得到平衡的学习和更新,有效提高了主动多分类方法的性能。

1 动态主动多分类方法

在多类别的分类任务中,假设训练集为 $Tr=\{X,Y\}=\{(x_i,y_i)\}_{i=1}^{l}$,其中 $x_i\in\mathbf{R}^d$,Y 为标签且 $y_i\in\{1,2,\cdots,c\}$,假设通过多分类方法得到的多分类器集合表示为 $F=\{f_1,f_2,\cdots,f_c\}$。本文采用支持向量机作为基准分类器,采用一对多的多分类思想构造多类学习模型。

1.1 主动多分类器的死锁和复位

假设 f_i 为一对多的多分类任务中第 i 类样本与其他类样本之间构造的分类器,如果对于训练样本集 X 中的任意样本 $x_j(j=1,\cdots,l)$,均有 $|f_i(x_j)|>1$ 成立,即所有训练样本 x_j 位于分类器 f_i 的分类间隔外,则这个超平面进行硬间隔的主动学习没有意义。

定义 1(分类器死锁) 假设对于训练集 X 中的任意样本 x_j,均有 $|f_i(x_j)|\geqslant 1$ 成立,即所有训练样本 x_j 位于分类器 f_i 的分类间隔外时,称分类器 f_i 当前处于死锁状态。

定义 2(分类器激活) 假设训练集 X 中的存在样本 x_j,使得 $|f_i(x_j)|<1$ 成立,即存在样本 x_j 位

于分类器 f_i 的分类间隔内时，称分类器 f_i 当前处于激活状态。

定理 1 在基于主动学习的硬间隔支持向量机多分类问题中，当前分类器集合 F^{old} 中存在某个分类器处于激活状态，则在经过主动学习方法执行之后，一定可以获得更新的分类器集合 F^{new}，即 $F^{\text{new}} \neq F^{\text{old}}$。

证明：根据分类器激活的概念，假设当前分类器集合 F^{old} 中至少存在某个分类器 f_i^{old} 处于激活状态，即存在样本 x_j，使得 $|f_i^{\text{old}}(x_j)|<1$ 成立，不妨假设样本 x_j 到分类器 f_i^{old} 的距离最近，即存在样本 x_j 位于分类器 f_i^{old} 的分类间隔内且在本轮主动学习中被选为最有价值样本，分为如下 4 种情况。

(1)若 $f_i^{\text{old}}(x_j)>0$ 且样本 x_j 的标记 y_j 为 i，由于分类器 f_i^{old} 处于激活状态，因此 $0<f_i^{\text{old}}(x_j)<1$ 成立，但由于 $y_j=i$，经过本轮主动学习之后 $f_i^{\text{new}}(x_j)\geqslant 1$ 成立，因此 $f_i^{\text{new}}\neq f_i^{\text{old}}$。

(2)若 $f_i^{\text{old}}(x_j)>0$ 且样本 x_j 的标记 y_j 不为 i，由于分类器 f_i^{old} 处于激活状态，因此 $0<f_i^{\text{old}}(x_j)<1$ 成立，但由于 $y_j\neq i$，经过主动学习之后 $f_i^{\text{new}}(x_j)\leqslant -1$ 成立，因此 $f_i^{\text{new}}\neq f_i^{\text{old}}$。

(3)若 $f_i^{\text{old}}(x_j)<0$ 且样本 x_j 的标记 y_j 为 i，由于分类器 f_i^{old} 处于激活状态，因此 $-1<f_i^{\text{old}}(x_j)<0$ 成立，但由于 $y_j=i$，经过本轮主动学习之后 $f_i^{\text{new}}(x_j)\geqslant 1$ 成立，因此 $f_i^{\text{new}}\neq f_i^{\text{old}}$。

(4)若 $f_i^{\text{old}}(x_j)<0$ 且样本 x_j 的标记 y_j 不为 i，由于分类器 f_i^{old} 处于激活状态，因此 $-1<f_i^{\text{old}}(x_j)<0$ 成立，但由于 $y_j\neq i$，经过本轮主动学习之后 $f_i^{\text{new}}(x_j)\leqslant -1$ 成立，因此 $f_i^{\text{new}}\neq f_i^{\text{old}}$。

在基于主动学习的硬间隔的 SVM 多分类问题中，当前分类器集合 F^{old} 中存在某个分类器处于激活状态，则在经过主动学习方法执行之后，一定可以获得更新的分类器集合 F^{new}，即 $F^{\text{new}}\neq F^{\text{old}}$。

证毕。

从定理 1 的证明过程可以看出，在证明中的情形(2)当中，由于对于旧的分类器 f_i^{old} 处于激活状态，根据分类器激活的定义，说明训练集 X 中的存在样本 x_j，使得 $|f_i(x_j)|<1$ 成立，不妨假设 x_j 位于分类器 f_i^{old} 判断为正类样本的一侧，即 $0<f_i^{\text{old}}(x_j)<1$ 成立。对于证明中的第二种情形，即样本 x_j 的标记 y_j 不为 i，即 $y_j\neq i$，因此经过主动的一对多 SVM 多分类之后，得到新的分类器 $f_i^{\text{new}}(x_j)$ 且样本 x_j 应当位于 SVM 分类器非第 i 类的一侧，即存在 $f_i^{\text{new}}(x_j)\leqslant -1$ 成立，对比 $0<f_i^{\text{old}}(x_j)<1$ 可以看出，分类器不仅有更新，而且更新幅度较大，对于证明中的情形(3)也类似，这不仅说明主动多分类器能够进行更新，而且算法具有较快的收敛速度。

1.2 轮换式的主动多分类学习

在传统支持向量机主动学习任务中，往往选择距离当前分类器最近的样本作为最有价值样本加入训练集参与训练，以更新分类器，但在多分类问题中，可能存在部分超平面附近样本点较为集中而另一部分超平面附近样本点距离都比较远的情形，这时如果仍然采用传统基于距离的主动学习方法，可能会导致部分超平面一直更新，而另一部分超平面则一直得不到更新，即分类器的更新存在不平衡性。针对这个问题，本文提出一种分位计数、轮换学习的多类别主动学习方法，以达到分类器平衡更新的目的。

由于处于死锁状态的分类器进行主动学习时是无法进行更新的，因此每轮主动学习过程仅考虑处于激活状态的分类器，假设当前处于激活状态的分类器数目为 c'，记当前处于激活状态的分类器集合为 $F'=\{f_1',f_2',\cdots,f_{c'}'\}$，则首先选择当前未参与训练的样本集中距离某个分类器最近的样本作为最有价值样本参与训练，不妨假设未参与训练的样本集中的样本 x_j 距离分类器 f_i' 的距离最近，样本 x_j 与分类器 f_i' 的距离为

$$d(x_j,f_i')=|f_i'(x_j)| \tag{1}$$

则将距离最近的样本 x_j 加入训练集参与训练，更新分类器集合，并进行分位计数，即用于区分第 i 类与其他类的分类器 f_i' 随带的计数器 n_i 加 1，如此循环迭代，且每轮训练过程中，都通过上述启发式的方法加入一个新的样本更新其中一个分类器，并采用分位计数方法进行标记当前分类器更新的次数，例如假设在第 t 轮的主动学习过程中，位于激活状态的分类器集合为 $F^t{}'=\{f_1^t{}',f_2^t{}',\cdots,f_{c'}^t{}'\}$，且每个分类器所对应的分位计数器分别为 $n_1^t,\cdots,n_{c'}^t$，则选择分位计数器最小值所对应的分类器进行更新，通过这

种分类计数的方式使得分类器可以进行平衡的更新,即避免了样本分布不均衡时,部分分类器一直更新而另一部分分类器确始终得不到优化的困境。

1.3 动态主动多分类算法

动态主动多分类方法首先抽取部分最有价值样本得到初始分类器,本文采用对正类样本和负类样本进行 k 均值聚类,然后取中心得到初始代表样本的方式得到初始训练样本,并进行初始的 SVM 训练,以得到初始分类器集合。根据初始分类器集合确定哪些分类器激活,哪些分类器死锁,然后对于激活的分类器进行分位计数、轮换学习的方式进行训练,在训练的过程中,死锁的分类器集合和激活的分类器集合之间会有互相转化,即随着训练分类器的改变,部分死锁的分类器会转化为激活状态,而部分激活的分类器由于当前没有样本位于分类间隔内会转为死锁状态。因此,这种动态的主动多分类学习方式一方面减少了主动多分类问题中需要考虑和训练的分类器数目,另一方面可以有效避免分类器的非平衡更新。具体地,动态主动多分类算法如下。

算法 1 动态主动多分类算法

初始化:假设训练样本集为 $Tr=\{X,Y\}=\{(x_i,y_i)\}_{i=1}^{t}$,其中 X 为训练样本集且 $x_i\in R^d$,Y 为训练样本集标签且 $y_i\in\{1,2,\cdots,c\}$。测试集为 $TE=\{TX,TY\}=\{(tx_j,ty_j)\}_{j=1}^{t}$,其中 TX 为测试样本集,TY 为测试样本集标签。

步骤 1 对训练样本集 X 进行初始划分,得到初始划分结果 $X\rightarrow\{X_p\}_{p=1}^{c}$,划分采用 k-均值聚类的方式:

(a) 从初始训练样本 X 中随机选择 c 个初始样本,计算其他样本到这 c 个初始样本的距离。

(b) 将每个样本归入距离最近的初始样本所属类别当中,并计算每个类别的中心

$$\mu_p=\sum_{i=1}^{n_p}x_{p_i}/n_p \tag{2}$$

式中:n_p 为第 p 个类 X_p 中样本的个数。

(c) 重新计算每个样本到类中心的距离,并反复迭代执行步骤(b,c),直到类中心不变化时为止。

(d) 计算每个样本到最终类中心的距离,选择距离每个类中心最近的一个样本构成训练集,采用这些类心进行训练,得到初始分类器集合 F^0。

步骤 2 根据分类器死锁和激活的定义,将分类器划分为死锁分类器集合 F^{0-S} 和激活分类器集合 F^{0-J}。

步骤 3 选择初始激活分类器集合 F^{0-J} 中距离每个分类器最近的样本,计算其与所属分类器的距离,选择距离最小的样本参与训练,并更新分类器集合及随带计数器。

步骤 4 重新构建死锁分类器集合与激活分类器集合,选择随带计数器最小的分类器构成分类器激活子集。

步骤 5 选择激活子集中距离最小的样本加入训练集训练,并更新分类器及随带计数器,并计算测试精度。

步骤 6 循环执行步骤 4~步骤 5,直到所有分类器不再更新为止,算法结束。

2 实验结果及分析

为验证本文提出的动态主动多分类方法(DYA)的有效性,本文与其他多种多分类方法进行了比较,对比方法包含 4 种基于主动学习的多分类方法以及 4 种传统的非主动多分类方法,其中主动多分类方法包括:基于距离的主动分类方法(Active classification method based on distance,DIA)[13]、基于概率分布的主动分类方法(Active classification method based on probability,PRA)[18]、基于香农熵的主动分类方法(Active classification method based on Shannon entropy,SEA)[11]以及基于信息熵的主动分类

方法(Active classification method based on information entropy, IEA)[11];对比的传统非主动多分类方法包括[6]:一对一的多分类方法(One versus one, OvO)、一对多的多分类方法(One versus most, OvM)、有向无环图多分类方法(Directed acyclic graphs, DAG)以及基于决策树(Decision tree, DT)的多分类方法。实验采用的 UCI 标准数据集见表 1。

表 1 实验数据集

Tab. 1 Datasets of experiment

数据集	#训练集	#测试集	#特征	#类别
Balance_scale	125	500	4	5
Glass	100	114	10	6
Iris	50	100	4	3
Letter	2 000	18 000	16	26
Machine	100	109	7	7
Page_blocks	1 000	4 473	10	5
Segment	1 000	1 310	19	6
Vehicle	300	546	18	4
Vowel	220	770	10	15
Wine	78	100	13	3

2.1 模型选择

对于 DYA 模型,采用高斯核作为核函数时,其涉及到的主要参数就是核参数 σ 及惩罚因子 C,表 2 反应了不同数据集取最优核参数时得到的结果。

表 2 不同参数的测试结果

Tab. 2 Testing results of various parameters

数据集	p	迭代次数	最大精度/%	平均精度/%
Balance scale	1.5	13	33.4	32.4
Glass	1.0	21	57.9	54.0
Iris	2.5	6	95.0	91.5
Letter	1.5	70	56.6	53.8
Machine	0.8	21	76.2	74.1
Page block	2.0	23	93.1	74.6
Segment	1.0	50	85.5	74.3
Vehicle	1.5	34	59.7	39.6
Vowel	1.5	70	46.2	42.6
Wine	2.0	15	98.0	91.6

以数据集 Wine 为例,当核参数取 1.0 时,图 1 给出了测试精度随着惩罚参数的变化情况。

从图 1 可以看出,当惩罚参数取值较小时,测试精度较低,当惩罚参数从 0.01 变到 10 的过程中,测试精度急剧增大,当惩罚参数在 10 到 90 的区间内增加时,测试精度缓慢增加,当惩罚参数超过 90 时,测试精度不再变化,对于数据集 Wine,惩罚参数取大于 90 的常数即可,但为了适用于所有数据集,本文中后续实验的惩罚参数统一取 200。

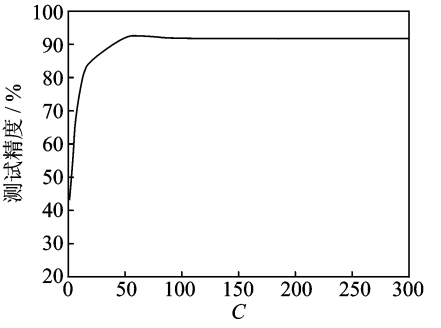

图 1 测试精度随惩罚参数的变化趋势

Fig. 1 Change tendency of testing accuracy with the penalty parameter

实验采用支持向量机作为分类器,核函数选用高斯核,参数设为 1.0,惩罚参数设置为 200。实

验中,对于训练集规模小于 500 的数据集,初始聚类划分参数取 20,否则取 50。

2.2 测试结果及分析

图 2 为 5 种基于主动学习的多分类方法在每个数据集上的主动迭代过程所得到的平均测试精度和上下标准偏差。上标准偏差和下标准偏差求法分别为

$$\mathrm{UPSD}=\sqrt{\frac{1}{n_{\mathrm{upper}}-1}\sum_{a_i\geqslant\overline{A}}(a_i-\overline{A})^2} \tag{3}$$

$$\mathrm{UNSD}=\sqrt{\frac{1}{n_{\mathrm{under}}-1}\sum_{a_i<\overline{A}}(a_i-\overline{A})^2} \tag{4}$$

式中:n_{upper}和 n_{under}分别表示测试精度值大于和小于平均测试精度$\overline{A}$的个数。从图 2 中可以看出,除数据集 Page_block 外,DYA 方法的平均测试精度都要优于其他 4 种主动学习方法,这说明本文提出的动态主动多分类方法与传统主动多分类方法相比,采用启发式的方法抽取最有价值样本,提高了主动多分类器的泛化性能;对于数据集 Page_block,由于其在初始主动学习过程中得到的分类器的测试精度明显低于其他 4 种方法,导致其对应的 UNSD 值较大且平均测试精度低于 DIA_MC 和 PA_MC,这说明本文提出的动态主动多分类方法通过分位计数的方式进行分类器更新,分类结果受到数据集分布不平衡性的影响减小,避免了分类器的非平衡更新,模型在数据集上的泛化性能也更为稳定。

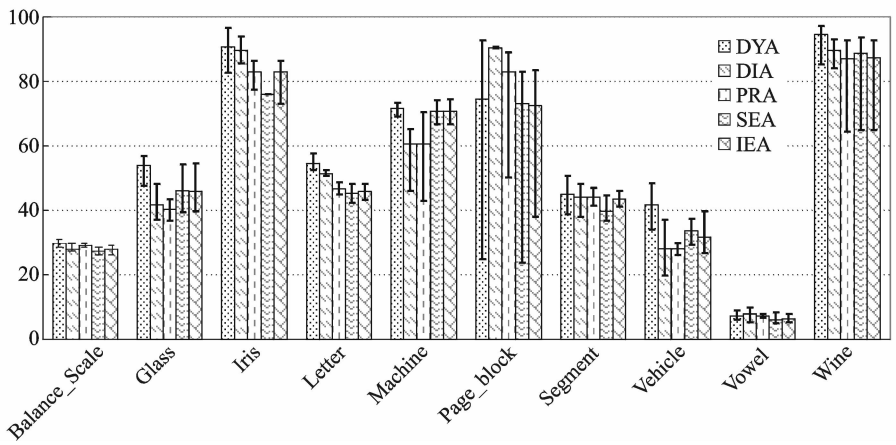

图 2 多种主动多分类方法测试精度比较

Fig. 2 Testing accuracy comparison of different active multiple classification methods

本文所提出的 DYA 方法与其他非主动多分类方法测试结果见表 3 所示。表中本文提出的 DYA 方法取主动学习的迭代子过程中最优的测试值作为对比。表 3 中带下划线的数值代表本文方法优于其他 4 种传统多分类方法的情况。从表中可以看出,在 5 个数据集 Balance_scale, Glass, Iris, Vehicle 及 Wine 上,本文提出的 DYA 方法要优于其他 4 种传统多分类方法,而在 Machine, Page_block 和数据集 Segment 上,DYA 方法与其他 4 种方法相差不大,而只有在数据集 Vowel 及 Letter 上,基于 DYA 方法与其他 4 种多分类方法有较为明显的差异。尽管 DYA 方法在数据集 Vowel 及 Letter 得到的结果较差,但结合图 2 可以看出,所有基于主动学习的方法在这两个数据集上的性能都较差,且测试精度都明显低于传统多分类方法,这可能是由于数据集 Vowel 及 Letter 数据集分布较为复杂,其样本的价值量无法采用基于欧氏距离的评价指标衡量,即在主动学习过程中无论采用哪种主动最有价值样本选取方法都无法有效得到最有价值样本,因此分类性能较差。

表 4 为本文提出的 DYA 方法及传统多分类方法在各数据集上的训练时间。从表中可以看出,在多数数据集上,本文提出的 DYA 方法与传统多分类方法相比,减少了训练时间,提高了学习效率。

表 3 不同方法的测试结果

Tab. 3 Testing results of various methods %

数据集	DYA	OvR	OvO	DAG	DT
Balance_scale	33.4	27.6	28.0	28.8	26.0
Glass	57.9	47.4	55.3	51.8	45.6
Iris	95.0	91.0	91.0	91.0	91.0
Letter	56.6	68.9	81.7	82.9	74.6
Machine	76.2	79.8	83.5	81.7	75.2
Page_block	93.1	94.7	94.6	92.8	87.6
Segment	85.5	95.6	96.5	93.4	82.0
Vehicle	59.7	29.5	31.7	32.6	28.8
Vowel	46.2	69.9	73.9	71.2	69.0
Wine	98.0	97.0	97.0	97.0	95.0

表 4 各数据集上的训练时间对比

Table 4 Training time comparison on datasets s

数据集	DYA	OvR	OvO	DAG	DT
Balance_scale	0.31	25.20	11.28	12.370	7.84
Glass	0.36	8.25	3.17	4.500	2.79
Iris	0.03	1.45	0.63	0.750	0.74
Letter	914.40	8 215.02	876.39	1 238.700	596.50
Machine	0.81	8.64	4.31	6.375	3.30
Page_block	13.41	3 078.90	1 240.88	1 143.800	865.80
Segment	16.68	2 156.70	461.19	820.300	477.91
Vehicle	1.19	74.03	25.19	30.620	22.60
Vowel	11.13	223.00	86.05	128.120	64.48
Wine	0.05	2.49	0.95	1.810	0.91

从实验结果可以看出,本文提出的动态主动多分类方法结合主动学习思想解决大规模样本的分类问题,具有如下多方面的优势:(1)DYA 方法通过结合主动学习思想,引入分类器激活和死锁的概念,一方面只考虑激活类的分类器,对于死锁的分类器则不参与训练,另一方面通过提取最有价值样本,对实际参与训练的样本规模进行了压缩,从分类器和样本两个方面减小了训练模型的规模和复杂度,提高了学习效率;(2)通过引入分位计数、轮换更新的方法,使得超平面进行平衡性的更新,避免了某些超平面一直更新而另一些超平面得不到更新的问题,有效提高了模型的泛化性能;(3)DYA 方法的设计为大数据挖掘提供了一条新的思路,即尽管数据规模很大、结构很复杂,但可能往往只有其中小规模的、简单结构的样本含有重要信息,因此可以通过一些传统方法首先挖掘这些重要数据,从这些重要数据中得到的分类器逼近甚至优于在全局样本上学习得到的分类模型。

3 结束语

针对目前已有的解决大规模多分类问题方法存在的学习效率低、泛化性能差的问题,本文提出了一

种改进的动态主动多分类方法,通过将死锁、激活等概念引入到主动多分类过程,在主动多分类过程中随着分类器的不断更新,动态地控制样本是否参与主动学习的过程,从分类器和样本两个方面减小了训练样本规模和模型复杂度,提高了学习效率,并采用分位计数、轮换学习方式的主动多分类方法,使得多类别的分类器能够得到平衡的学习和更新。在以后的工作当中,应将进一步探索新的面向多分类主动学习问题的最有价值样本提取机制,并设计针对不同类型数据的相应算法,以进一步提高模型的泛化性能。

参考文献:

[1] Nature. Big data [EB/OL]. http://www.nature.com/news/specials/bigdata/index.html, 2008-09-03/2012-10-02.

[2] Science. Special online collection: Dealing with data [EB/OL]. http://www.sciencemag.org/site/special/data/, 2011-02-11/2012-10-02.

[3] Viktor M S, KEnneth C. Big data [M]. Hangzhou: Zhejiang People Press, 2013: 1-23.

[4] Wu Qihui, Qiu Junfei, Ding Guoru. Machine learning methods for big spectrum data processing [J]. Journal of Data Acquisition and Processing, 2015, 30(4): 703-713.

[5] Liu Shaoyu, Zhou Jie, Li Bicheng, et al. Entity relation extraction method based on multi-SVM-KNN classifier [J]. Journal of Data Acquistion and Processing, 2015, 30(1): 202-210.

[6] Zhang P, Li M, Wu Y, et al. Unsupervised multi-class segmentation of SAR images using fuzzy triplet Markov fields model [J]. Pattern Recognition, 2012, 45(11): 4018-4033.

[7] Krawczyk B, Wozniak M, Herrera F. On the usefulness of one-class classifier ensembles for decomposition of multi-class problems [J]. Pattern Recognition, 2015, 48(12): 3969-3982.

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

[9] 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 Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(10): 1880-1897.

[10] 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.

[11] Qi G, Hua X, Rui X, et al. Two-dimensional active learning for image classification [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Anchorage, Alaska: [s.n.], 2008: 24-26.

[12] Saito P T, Suzuki C T, Gomes J F, et al. Robust active learning for the diagnosis of parasites [J]. Pattern Recognition, 2015, 48(11): 3572-3583.

[13] Tong S. Active learning: Theory and applications [D]. Stanford: Stanford University, 2001: 1-26.

[14] Huang S J, Zhou Z H. Active leraning by querying informative and representative examples [C]// Advances in Neural Information Processing Systems 23 (NIPS2010). Cambridge, MA: MIT Press, 2010: 2388-2396.

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

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

[17] Di Peng, Duan Liguo. New naive bayes text classification algorithm [J]. Journal of Data Acquistion and Processing, 2014, 29(1): 71-75.

[18] Jain P, Kapoor A. Active learning for large multi-class problems [C]// Proceedings of the 2009 IEEE Computer Vision and Pattern Recognition. [S.l.]:IEEE,2009: 219-224.

作者简介:

郭金玲(1982-),女,讲师,研究方向:机器学习与数据挖掘。

樊东燕(1965-),女,教授,研究方向:模式识别、图形图像处理。

郭虎升(1986-),通讯作者,男,讲师,博士,研究方向:计算智能与机器学习,E-mail: guohusheng@sxu.edu.cn。