计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
42(增刊):222~227,2005

集成学习算法在增量学习中的应用研究

文益民[1,2] 杨 旸[1] 吕宝粮[1]
[1](上海交通大学计算机科学与工程系 上海 200030)
[2](湖南工业职业技术学院 长沙 410007)
(ymwen@cs.sjtu.edu.cn)

Rescearch on the Application of Ensemble Learning Algorithms to Incremental Learning

Wen Yimin[1,2], Yang Yang[1], and Lü Baoliang[1]
[1](*Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai* 200030)
[2](*Hunan Industry Polytechnic, Changsha* 410007)

Abstract How to retain the knowledge learned before and how to acquire new knowledge are two problems for incremental learning. In this paper, some ensemble learning algorithms are applied to incremental learning and a modularized incremental learning model is proposed. The possibility of applying behavior knowledge space (BKS), dynamic classifier selection (DCS), and majority voting (MV) to incremental learning is investigated, and a new algorithm BKS based on DCS (BoD) is proposed. The simulation results indicate that DCS is a good algorithm for incremental learning, BKS and MV are not as good as DCS while BoD boosts BKS and works as efficiently as DCS does; The proposed incremental learning model can not only completely retain the knowledge learned before, but also acquire the current new knowledge that includes concept drift.

Key words incremental learning; knowledge retention and acquisition; ensemble learning; modularization; support vector machines

摘 要 如何能有效地保持原本学习过的知识,又能不断获取新知识?这是增量学习面临的难题.将集成学习算法移植应用于增量学习,建立了模块化增量学习模型,研究了 Behavior Knowledge Space (BKS)、Dynamic Classifier Selection (DCS)和 Majority Voting (MV)3 种集成学习算法应用于增量学习的可能性,并提出了算法 BKS based on DCS (BoD).仿真实验表明,DCS 表现最好,BKS 和 MV 表现次之,BoD 很好地提升了 BKS 而与 DCS 完全相当;提出的增量学习模型不但能完全保持以前学习过的知识,而且能有效地获取当前的新知识(包括概念漂移 concept drift).

关键词 增量学习;知识保持与知识获取;集成学习;模块化;支持向量机

中图法分类号 TP391

1 引 言

人脑具有渐进学习的能力,研制具有类似人脑学习能力的计算模型一直是机器学习领域的重要分支之一.在实际应用中由于采集样本的代价或时间等原因,很难一次性获得全部样本.实际问题也不允许等到获取全部样本后再进行机器学习.因此只

收稿日期:2005-06-15
基金项目:国家自然科学基金项目(60375022,60473040)

能逐步将获取样本中包含的知识纳入学习系统中,也就是进行增量学习. 然而,目前各种人工学习系统的构造算法在本质上都是基于当前学习环境而以尽量保证学习系统推广能力为目的的一个寻优过程,所以现有的各种机器学习算法本质上都不适应增量学习. 这种不适应[1]体现在:计算模型或者缺乏获取新知识的能力,或者不能保持原本获取的知识. 机器学习领域对此问题进行了大量的研究. 人工神经网络通常通过改变隐层单元数目或控制权值改变的幅度来实现新知识的获取或减少遗忘[2],这显然会使网络的学习能力或网络的知识容量受到限制. 将人工神经网络和决策树结合是解决增量学习问题的一种好方法[3]. 支持向量机的增量学习通常通过将原有的支持向量与新增加样本合并在一起重新训练[4]而得到新的支持向量机,但当存在概念漂移(concept drift)时该方法效果不佳[5].

集成学习(ensemble learning)一直是机器学习的一个研究热点[6~10]. 模块化是扩展现有机器学习算法增量学习能力的有效方法之一[11]. 已有学者尝试将两者结合而提出了多种模块化增量学习的算法[12,13]. 但是,每增加一批学习样本,这些算法都需要训练多个子学习系统,增量学习系统的规模以超线性的速度增长.

2 模块化增量学习模型

提出的模块化增量学习模型可用图1描述:

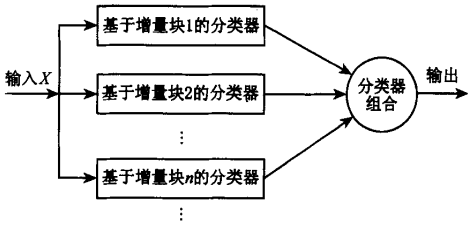


图1 模块化增量学习模型

算法1. 模块化增量学习模型的训练算法.

(1) 初始化. 训练算法相关参数;初始训练数据集:

$$TR_1=\{\{\boldsymbol{X}_i,\ y_i\}\mid \boldsymbol{X}_i\in R^n,$$
$$y_i\in\{0,1,2,\cdots,c\},\ 1\leqslant i\leqslant N_1\},$$

$\boldsymbol{X}_i$ 表示样本,y_i 表示样本类别,$\{0,1,2,\cdots,c\}$ 表示类别标志集合. 当前验证集 $V_1=TR_1$.

(2) 使用训练数据集 TR_1 得到分类器 C_1.

(3) For $t=2,3,\cdots,n$

3.1 用在时间段$[t-1,t]$采集的训练集增量块 TR_t 训练得到子分类器 C_t,其中的 t 在下文中被称做时刻或增量学习步进数(当前增量学习进行的步骤数);

3.2 更新验证集 $V_t=V_{t-1}\cup TR_t$;

3.3 将子分类器 C_t 加入增量学习系统,则当前 t 个训练好的子分类器构成当前增量学习系统.

算法2. 模块化增量学习模型的测试算法.

(1) 初始化. 当前增量步进数 t;测试样本 $\boldsymbol{X}$.

(2) 初始化特定分类器组合策略需要的相关参数.

(3) For $i=1,2,\cdots,t$

3.1 求分类器 C_t 对 $\boldsymbol{X}$ 给出的类别判断 $C_i(\boldsymbol{X})$;

3.2 计算特定分类器组合策略需要的信息.

(4) 若所有 $C_i(\boldsymbol{X})$相同,则 $C_i(\boldsymbol{X})$为 $\boldsymbol{X}$ 的类别判断.

(5) 如果不是所有的 $C_i(\boldsymbol{X})$相同,则根据分类器组合策略给出 $\boldsymbol{X}$ 的类别输出.

2.1 Behavior Knowledge Space (BKS)分类器组合方法

BKS[9]方法使用当前全部分类器组合给出分类. 设当前时刻为 t,则增量学习系统中有 t 个子分类器,则 t 个子分类器对同一个训练样本的类别判断构成 BKS 中的一个 t 维向量. 在时刻 t,将验证集 V_t 中的训练样本提交给各个子分类器测试将得到有 $|V_t|$(V_t 中元素的数目)个元素的 BKS. 根据此向量空间可以计算得到 k^t(k 表示类别数)个 BKS 单元. 测试时,各个子分类器对同一测试样本给出的类别判断构成一个 t 维的查询向量,以此查询向量去查找对应的 BKS 单元,最后根据相应 BKS 单元的信息给出该测试样本的类别判断. 为了节省计算时间,可在增量学习的同时,逐步计算并保存每个增量块的 BKS 单元信息.

2.2 Dynamic Classifier Selection (DCS)分类器组合方法

在 DCS 方法[10]中,类别判断置信度最高的一个子分类器被选中给出测试样本的类别判断. 设在时刻 t 验证集 V_t 中测试样本 $\boldsymbol{X}$ 的 q 个最近邻样本构成集合 Q_x^t. 设某个子分类器 C_i 将测试样本 $\boldsymbol{X}$ 判断为 $C_i(\boldsymbol{X})$类,则在时刻 t 该分类器对该测试样本 $\boldsymbol{X}$ 的类别判断置信度为

$$T_i^t(\boldsymbol{X})=\frac{Q_x^t\text{ 中真正属于 }C_i(\boldsymbol{X})\text{ 类的样本数目}}{Q_x^t\text{ 中被分类器 }C_i\text{ 判断为 }C_i(\boldsymbol{X})\text{ 类的样本数目}}.$$

2.3 Majority Voting (MV)分类器组合法

在时刻 t 由所有子分类器对测试样本 $\boldsymbol{X}$ 的类别分别给出判断,然后进行多数投票. 获票多数的类别作为 $\boldsymbol{X}$ 的类别判断. 如果出现票数相等情形,则进行随机猜测.

2.4 BKS based on DCS (BoD)分类器组合法

从上面对于 BKS 和 DCS 的描述可以知道,BKS 重视模块间合作,而 DCS 重视模块间竞争. 本文将 BKS 和 DCS 的这两个特点组合起来,得到了 BoD 算法.

算法 3. BoD 算法.

(1) 预先设置阈值 ε, $0\leqslant\varepsilon\leqslant1$;

(2) 在时刻 t,根据 DCS 方法计算各子分类器对测试样本 $\boldsymbol{X}$ 的类别判断置信度 $T_i^t(\boldsymbol{X})$, $1\leqslant i\leqslant t$;

(3) 寻找置信度最高的分类器:

$$imax=\arg\max_{i=1}^{t}T_i^t(\boldsymbol{X});$$

(4) 求得集合 $\nabla=\{i\mid(T_{imax}^t(\boldsymbol{X})-T_i^t(\boldsymbol{X}))\leqslant\varepsilon,\ 1\leqslant i\leqslant t\}$. 如果 $|\nabla|=1$,则采用 DCS 进行分类器组合,否则选择分类器 $C_j(j\in\nabla)$ 采用 BKS 进行分类器组合.

BoD 方法如同针对具体问题在专家库中挑选若干专家进行集体决策. 阈值 ε 的设置较为简单,通常设置为 0.001,这是感觉上认为两个分类器准确率值没有太大区别的一个数值.

3 仿真实验

本文采用的实验平台为配置 2GB RAM 和 3GHz CPU 的 PC;分类器训练算法采用 libSVM;核函数采用 RBF 核函数. 针对两类问题共 5 个数据集进行了仿真实验. 一类是当数据分布保持不变时,增量学习系统不断学习新样本,能否不断提高自己的推广能力? 另一类是当数据分布随时间渐变,也就是存在概念漂移的情况下,增量学习系统能否不断学习新概念,从而具备对新概念的推广能力?

第 1 类问题中,第 1 个数据集是机器学习领域著名的双螺旋线问题,第 2 个是棋盘数据(checkboard)问题[14]. 第 3 个数据集是来自 UCI[15] 的 Adult 数据集. 第 4 个数据集采用来自 UCI 中的共有 7 类数据的 Forest coverType 数据集,本文抽取了 Forest coverType 数据中的分别属于第 2 类和第 6 类的样本. 在第 1 类问题的每个实验中,将整个训练集随机划分成 10 等份,每一等份当作一个训练增量块,这相当于按照时间顺序进行了 10 次样本采集.在每个增量步,使用共同的测试集测试增量学习系统的性能.

第 2 类问题中,按照式(1)生成了随着时间延续有概念漂移的人工数据 Gaussian data.

$$f_{\boldsymbol{X}}^t(\boldsymbol{X}\mid P)=\frac{1}{2\pi\sigma_p^2}\exp\|\boldsymbol{X}-\boldsymbol{\mu}_p^t\|^2,$$
$$f_{\boldsymbol{X}}^t(\boldsymbol{X}\mid N)=\frac{1}{2\pi\sigma_n^2}\exp\|\boldsymbol{X}-\boldsymbol{\mu}_n^t\|^2, \tag{1}$$

其中:$\sigma_p=1$; $\sigma_n=2$; $\boldsymbol{\mu}_p^t=[0,0]^{\mathrm{T}}+(t-1)\times[8,0]^{\mathrm{T}}$, $\boldsymbol{\mu}_n^t=[2,0]^{\mathrm{T}}+(t-1)\times[8,0]^{\mathrm{T}}$, $t=1,2,\cdots$. P 表示样本 $\boldsymbol{X}$ 属于正类, N 表示样本 $\boldsymbol{X}$ 属于反类. Gaussian data 的数据分布示意图如图 2 所示:

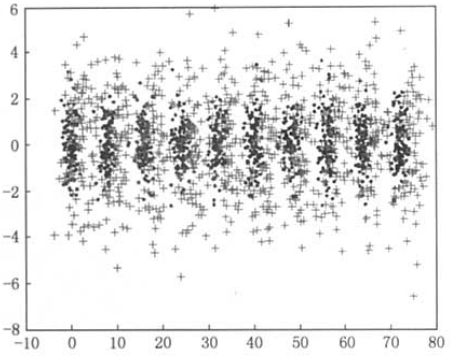


图 2 存在概念漂移的高斯数据的分布图

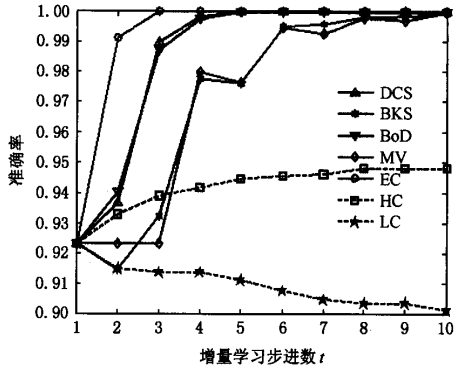
在第 2 类问题的实验中,按照式(1)在第 t 个时刻附近随机采集 2000 个样本作为训练集 TR_t, 20000 个样本作为测试集 TS_t. 用 $\bigcup_{i=1}^{t}TS_i$ 作为时刻 t 的测试集. 由于存在概念漂移,MV 显然不合适作为分类器组合规则,所以只比照研究了采用 DCS, BoD 和 BKS 作为分类器组合规则时增量学习系统学习新概念的能力问题.

以上所有实验均重复 10 次,图 3、图 4 和图 5 中是 10 次实验的平均值. 上述 5 个数据集的数据分布和相关实验参数见表 1.

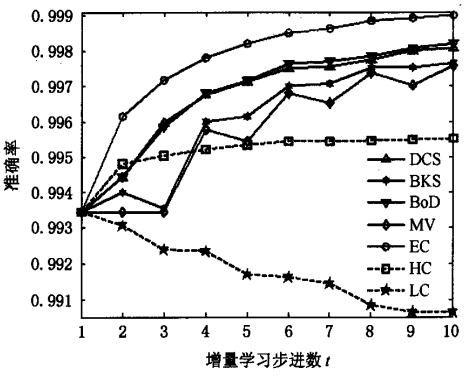
4 实验结果分析与讨论

在图 3 中将 DCS 和 BoD 的测试准确率曲线与使用当前全部训练样本训练得到的单个分类器的测试准确率曲线比较可以知道:DCS 和 BoD 不但保持了原本学习到的知识,还能不断获取新样本带来的知识. 与 DCS 和 BoD 相比,采用 BKS 和 MV 的增量学习系统的增量学习能力较差,且波动性较大.

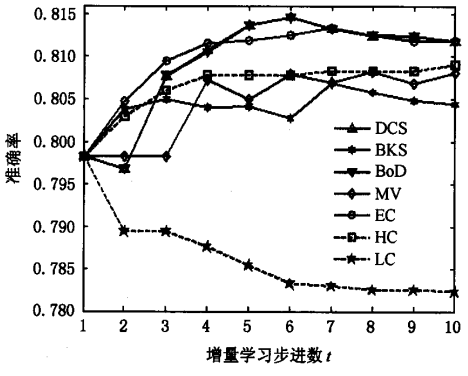
图3还说明了:DCS和BoD均能有效地整合各个子分类器的分类能力,最终使分类器的分类能力得到提升,因为它们的推广能力比它们的组成模块中推广能力最好的子分类器还要好.



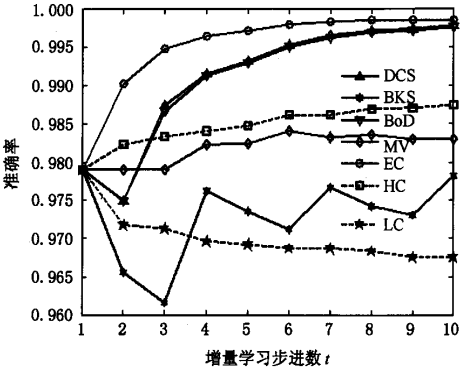
(a) 在 twospirals 数据集上进行的准确率实验



(b) 在 checkboard 数据集上进行的准确率实验



(c) 在 adult 数据集上进行的准确率实验



(d) 在 forest covertype 数据集上进行的准确率实验

EC表示使用当前采集的所有训练样本训练的单个分类器. HC表示当前推广能力最好的子分类器. LC表示当前推广能力最差的子分类器

图3 准确率比较

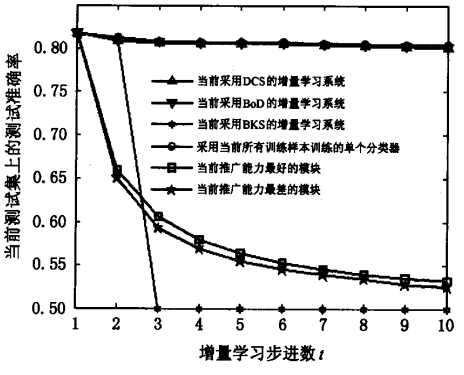


图4 存在概念漂移的高斯数据实验中使用DCS, BoD和BKS在增量测试集上的准确率比较

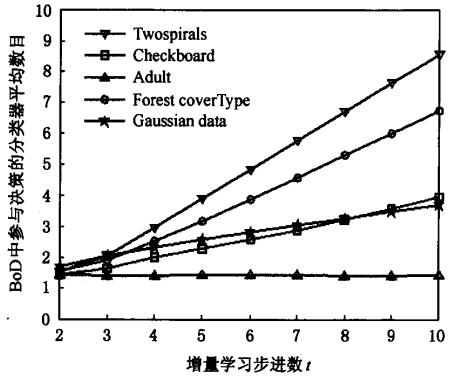


图5 进入测试算法第5步,BoD中参与分类器组合给出一个测试样本类别判断的平均分类器数目

从图4可知,BoD和DCS的测试准确率曲线与使用当前全部训练样本训练的单个分类器的测试准

确率曲线几乎完全重合,这说明 BoD 和 DCS 能非常有效地获取因概念漂移产生的新知识. 显然,如果将分布改变的数据当做产生的新类别,BoD 和 DCS 就能处理新类别信息. 在图 4 中由于存在概念漂移,增量学习系统的各个子分类器均不能掌握全部训练样本中的全部知识. 图 4 还说明 BKS 学习能力严重依赖于它的组成模块的学习能力.

通过比较 BoD 与 DCS 的测试准曲率曲线还可以看出:BoD 完全与 DCS 相当. 其中的原因可以从图 5 给出解释. 可以看到:在 BoD 中,采用 DCS 排除了一些类别判断置信度低的分类器,使参与 BKS 的分类器减少. 这完全相当于针对要解决的问题在专家群体中选择一部分权威专家进行集体决策. 所以,在硬件软件实现时,BoD 将比 DCS 更具有可靠性,因为 DCS 可以看成 BoD 中只有一个分类器没有失效时的情形.

后续的工作包括:如何选择性地"遗忘" V_t 中的部分样本;基于 BoD 和 DCS 的模块化增量学习系统与前述模块化增量学习模型[12,13]的比较;q 值的设置规则和模块化增量学习系统的增量学习能力与增量块大小的关系等.

表 1 五个数据集的数据分布和相关实验参数设置

数据集	特征维数	训练数据	测试数据	增量块大小	c	σ	邻域大小 q
Two spirals	2	3000	20000	300	128	2	5
Checkboard	2	16000	80000	600	1000	31.62	50
Adult	14	30162	15060	3016	16	1	4000
Forest(2,6)	54	28132	28132	281	128	0.25	90
Gaussian data	2	20000	200000	2000	0.1	2	100

5 结 论

本文通过建立模块化增量学习模型,探讨了几种集成学习算法应用于增量学习的可能性. 与前述模块化增量学习模型[12,13]相比,本文提出的模块化增量学习模型构造简单,已训练好的模块不需重新训练,易于软件和硬件实现. 仿真实验表明,基于 DCS 和 BoD 的增量学习系统具有很好的增量学习能力,基于 BKS 和 MV 的增量学习系统表现很不稳定. BoD 改进了 BKS 而与 DCS 相当,但其可靠性要比 DCS 好. 基于 DCS 和 BoD 的增量学习模型不但能有效保持学习系统原来的学习成果,而且能不断获取新知识(包括概念漂移).

参 考 文 献

1 S. Grossberg. Nonlinear neural networks: Principles, mechanisms and architectures. Neural Networks, 1988, 1(1): 17~61

2 L. M. Fu, H. H. Hsu, J. C. Principe. Incremental backpropagation learning networks. IEEE Trans. Neural Networks, 1996, 7(3): 757~761

3 Z. H. Zhou, Z. Q. Chen. Hybrid decision tree. Knowledge-Based Systems, 2002, 15(8): 515~528

4 李凯, 黄厚宽. 支持向量机增量学习算法研究. 北方交通大学学报, 2003, 27(5): 34~37

5 S. Rüping. Incremental learning with support vector machines. ICDM2001, San Jose, CA, 2001

6 周志华, 陈世福. 神经网络集成. 计算机学报, 2002, 25(1): 1~8

7 王珏, 石纯一. 机器学习研究. 广西师范大学学报, 2003, 21(2): 1~15

8 B. L. Lu, M. Ito. Task decomposition and module combination based on class relations: A modular neural networks for pattern classification. IEEE Trans. Neural Networks, 1999, 10(5): 1244~1256

9 Y. S. Huang, C. Y. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. IEEE Trans. Pattern Analysis and Machine Intelligence, 1995, 17(1): 90~94

10 K. Woods, W. P. Kegelmeyer, K. Bowyer. Combination of multiple classifiers using local accuracy estimates. IEEE Trans. Pattern Analysis and Machine Intelligence, 1997, 19(4): 405~410

11 罗四维, 温津伟. 神经场整体性和增殖性研究与分析. 计算机研究与发展, 2003, 40(5): 668~674

12 R. Polikar, L. Udpa, S. S. Udpa, *et al*. Learn++: An incremental learning algorithm for supervised neural networks. IEEE Trans. System, Man, and Cybernetic, 2001, 31(4): 497~508

13 B. L. Lu, M. Ichikawa. Emergent online learning in min-max modular neural networks. The IJCNN'01, Washington, DC, USA, 2001

14 Y. M. Wen, B. L. Lu. A cascade method for reducing training time and the number of support vectors. In: Lecture Notes in Computer Science 3173. Berlin: Springer-Verlag, 2004. 480～485

15 C. L. Blake, C. J. Merz. UCI. ftp: // ftp.ics.uci.edu/pub/machine-learning-database, 1998

文益民,1969 年生,博士研究生,主要研究方向为统计学习理论、集成学习、生物信息学和图像处理.

杨旸,1981 年生,博士研究生,主要研究方向为统计学习理论、生物信息学.

吕宝粮,1960 年生,日本京都大学工学博士,2002 年起任上海交通大学教授,博士生导师,目前的主要研究领域为仿脑计算理论与模型、神经网络、并列机器学习、脑-计算机接口、人脸识别、生物信息学和自然语言处理.

集成学习算法在增量学习中的应用研究

作者： 文益民， 杨旸， 吕宝粮
作者单位： 文益民(上海交通大学计算机科学与工程系,上海,200030;湖南工业职业技术学院,长沙,410007)， 杨旸,吕宝粮(上海交通大学计算机科学与工程系,上海,200030)