计算机研究与发展
Journal of Computer Research and Development

ISSN 1000-1239/CN 11-1777/TP
42(增刊):243~247,2005

基于最小最大模块化分类器的自重组学习算法

赵 海 吕宝粮
(上海交通大学计算机科学与工程系 上海 200030)
(zhaohai@cs.sjtu.edu.cn)

A Self Recombination Learning Algorithm for Min-Max Combining Classifier

Zhao Hai and Lü Baoliang
(*Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai* 200030)

Abstract Min-max modular classifier is a kind of combining classifier framework for elastic task decomposition and simple result combination. The previous work shows that clustering or anti-clustering method may be the more reliable decomposition method for training set under this frame. However, clustering methods are sensitive to initial partition of data sets and often lead to unbalanced partitions. Based on the task decomposition of min-max modular classifier, a symmetrical selection combination is proposed to replace traditional min-max combination, and then according to the new combination strategy, a self recombination learning algorithm is given to finish a more optimized decomposition of training set and training at the same time in parallel or modular way. Compared to other distributed learning methods, the proposed method can maintain elastic task decomposition and hold good parallel processing feature and higher response performance. In addition, the proposed method is independent of classification algorithm, and thus it can be realized through various algorithms. The experimental results verify that the proposed method can gain a better generalization performance and faster training effectiveness.

Key words combining classifier; self recombination algorithm; symmetrical selection combination

摘 要 最小最大模块化分类器是一种具有弹性的任务分解和简单的结果合成的组合分类器框架.已有研究表明,最小最大模块化分类器任务分解的一种有效策略是使用聚类或者逆向聚类方法.但是聚类算法依赖于初始分解,同时难以保证生成模块的均衡性.基于最小最大模块化分类器的任务分解策略,提出使用对称选择算法来替代原有的最小最大组合过程,在此基础上,进一步提出了一种自重组学习算法,能够在并行或者模块化处理模式下同步完成训练集较为优化的划分和训练过程.所提方法继承了传统的最小最大模块化分解的弹性分解特性,易于实现灵活分解和平衡化的学习.同时,和传统的分布式分类方法相比较,提出的方法具有良好的并行处理特性和更高的测试性能.此外,该方法与基分类器算法无关,可以适用于多种分类算法的分布式实现.实验结果验证了所提算法的有效性.

关键词 组合分类器;自重组学习算法;对称选择组合

中图法分类号 TP391

1 引 言

最近几年,人们提出了多种通过大规模分类问题分解来实现模块化和并行处理的技术.但是,在问题分解中,我们希望能够额外考虑两个方面的约束.一个是负载平衡约束.简单来说,就是希望各分类器的训练和学习时间能够尽可能接近,以便在并

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

行学习中保持一致的处理进度,从而防止处理上的瓶颈. 对于内存受限系统的学习,负载平衡的划分也能够充分利用单机的计算资源. 另外一个是单个分类器的学习平衡性约束. 我们希望单个基分类器的学习能够通过适当的训练集划分避免不平衡学习的出现. 而这往往就要求训练集的各个类别的样本数量尽量接近. 最小最大模块化(min-max modular, M^3)分类器的问题分解能较为容易满足上述处理要求[1],作为一种基于训练集划分的组合分类器,它将是本文的中心研究内容. 由于同时考虑了上述两点约束条件,本文研究的出发点也和以往工作稍有不同.

在基于训练集划分的分类任务分解中,如何有效划分训练集以便保证分解后产生的基分类器的组合精度是一个重要课题. 目前在组合分类器领域,通常使用两种方法来解决这个问题. 一个是使用聚类方法[2~4]. 除了聚类,另外一个办法是使用特定策略抽取的一系列样本组成新的训练子集,通过学习这些训练子集产生一系列新的基分类器,继而利用这些基分类器的某种组合来实现分类效果的改善. 这类方法的典型例子是 bagging 方法[5]和 AdaBoosting 算法[6].

我们早期的一个工作也建议使用类似的聚类方法,如超平面划分作为最小最大模块化分类器的划分策略[7]. 但是聚类分解方法有 3 个明显的缺点,一是很容易导致不均衡划分. 二是任务分解策略对于最终的组合精度的影响不是直接的,这使得聚类算法的中止条件往往只能根据经验进行. 三是聚类方法对于初始化状态往往很敏感,在不同数据集上,聚类算法的表现有时很不一致.

本文提出一种自重组学习算法,用来部分地克服以上的困难. 该算法的一个重要特点就是在维持划分规模和处理单元数量不变的情况下,通过多轮训练来优化分解过程以获得较佳的组合分类效果.

2 最小最大模块化分类器的任务分解和结果组合

最小最大模块化分类器[1]可以弹性地将一个复杂的分类问题,分解成许多规模小的、各自独立的二类子问题,对于每个二类子问题,使用一个基分类器学习,然后根据两个结合规则,即最小化规则和最大化规则,将所有基分类器的分类结果结合成原问题的结果. 下面给出一个二类分类问题的最小最大模块化分类器的描述.

对于一个二类问题 T, χ^+ 表示属于类别 C 的正的训练样本集, χ^- 表示属于类别 C 的负的训练样本集.

$$\chi^+ = \{(x_i^+, +1)\}_{i=1}^{l^+},$$
$$\chi^- = \{(x_i^-, -1)\}_{i=1}^{l^-}, \tag{1}$$

其中 $x_i \in R^n$ 表示输入样本向量, l^+ 和 l^- 分别表示训练样本中正类和负类的样本的个数. 样本集 χ^+ 和 χ^- 又可以分别被分解为 N^+ 和 N^- 个子样本集,

$$\chi_j^+ = \{(x_i^{+j}, +1)\}_{i=1}^{l_j^+}, \text{且 } j = 1, \cdots, N^+, \tag{2}$$

$$\chi_j^- = \{(x_i^{-j}, -1)\}_{i=1}^{l_j^-}, \text{且 } j = 1, \cdots, N^-, \tag{3}$$

其中, $\bigcup_{j=1}^{N^+} \chi_j^+ = \chi^+$, $1 \leqslant N^+ \leqslant l^+$ 和 $\bigcup_{j=1}^{N^-} \chi_j^- = \chi^-$, $1 \leqslant N^- \leqslant l^-$.

将正负样本集 χ^+ 和 χ^- 分别分解为 N^+ 和 N^- 个子样本集后,最初的二类问题 T 被分解为 $N^+ \times N^-$ 个相对较小、并且比较平衡的二类子问题 $T^{(i,j)}$:

$$(T^{(i,j)})^+ = \chi_i^+, (T^{(i,j)})^- = \chi_j^-, \tag{4}$$

其中, $(T^{(i,j)})^+$ 和 $(T^{(i,j)})^-$ 分别表示子问题 $T^{(i,j)}$ 的正负样本集. 可以看出,所有的二类子问题是相互独立的,因此可以并列进行学习. 在不引起误会的情况下, $T^{(i,j)}$ 也用来表示对其进行学习的基分类器.

在所有子问题训练结束后,对于一个测试样本,根据下面两个基分类器结合规则[1],利用 N^+ 个 MIN 单元和一个 MAX 单元,将 $N^+ \times N^-$ 个基分类器的分类结果组合成原始问题的分类结果.

$$T^i(x) = \min_{j=1}^{N^-} T^{(i,j)}(x), i = 1, \cdots, N^+,$$
$$T(x) = \max_{i=1}^{N^+} T^i(x), \tag{5}$$

其中, $T^{(i,j)}(x)$ 表示对应子问题 $T^{(i,j)}$ 所训练的基分类器的传递函数, $T^i(x)$ 表示将 N^- 个基分类器运用 MIN 单元集成后所代表的基分类器的传递函数.

传统的最小最大模块化分类器使用的基分类器算法有 BP 神经网络[1]、SVM[8]和 k-NN 算法[9],并成功应用于多种不同领域[10,11].

3 基于对称选择算法的结果组合

为了实现自重组学习,我们需要更为有效的组

合策略,来替代上节所提到的最小最大组合(分类任务分解还是依据上节的处理). 为描述方便,首先给出如下定义. 一组基分类器,$T^{(ij)}, j=1,\cdots,N^-$,称为一个组号为 i 的正类组. 对称地,一组基分类器,$T^{(ij)}, i=1,\cdots,N^+$称为一个组号为 j 的负类组. 容易看出,上节描述的最小最大组合过程,可以视为对一组分类结果均支持正类的正类组的搜索过程. 如果这样的一个正类组存在,那么,最小最大组合定义的组合分类结果就是正类,否则就是负类. 我们称其中的基分类器一致支持正类输出的一个正类组为获胜正类组. 对称地,可以引入负类组搜索. 称其中的基分类器一致支持负类输出的一个负类组为获胜负类组,则负类组搜索就是搜索获胜负类组的过程.

下面引入的对称选择算法将是同时对获胜正类组和获胜负类组加以考虑的组合过程. 考虑如下两点:

(1) 在一次测试中,不可能同时存在获胜正类组和获胜负类组. 因为这两组之中,必有一个分类器是公用的,它不可能同时输出两种不同的类别.

(2) 为了使得组合输出无偏,定义组合输出为负类的条件为不可能找到一个正类获胜组,定义组合输出为正类的条件为不可能找到一个负类获胜组.

基于以上考虑设计的对称选择算法描述如下:

(1) 设置初始测试的行号 $i=1$ 以及列号 $j=1$;

(2) 重复如下操作:

① 测试基分类器 $T^{(ij)}$;

② 如果 $T^{(ij)}$输出支持正类,则 $j=j+1$;

③ 如果 $T^{(ij)}$输出支持负类,则 $i=i+1$;

④ 如果 $N^+ + 1 = i$,则立即返回,组合输出为负类;

⑤ 如果 $N^- + 1 = j$,则立即返回,组合输出为正类.

容易看出,对称选择算法所要用到的基分类器个数不会超过 $N^+ + N^- - 1$ 个.

4　自重组学习算法

自重组学习算法来源于如下的简单思路. 如果将所有的基分类器的测试输出按照正类组一行,负类组一列排成一个矩阵,那么,在对称选择算法的搜索过程中,出发点是左上角的基分类器,结束点是最后一行或者最后一列上的基分类器. 越是靠近右下角,也就是越是远离出发点的基分类器将越少有机会获得调用,参与测试过程. 或者,越是靠近后方的基分类器越是对于最终的分类结果影响力越小. 因此,如果将错误率高的基分类器尽可能分配到靠后的行或者列,那么就能降低测试出错的机会,也就是能够改进组合输出的推广能力. 调整错分样本到训练集的后半部分就能实现这样的要求. 自重组学习算法描述如下:

(1) 按照指定的规模完成二类问题的 M^3 分解;

(2) 设置最大自重组学习次数 M,设置自重组学习次数计数器的初始值 $t=1$.

(3) 如果 $t<M$,重复如下操作:

① 依据训练集分解结果,并行训练各基分类器;

② 各个基分类器对于训练集中的所有样本依次进行自测试,记下各基分类器测试结果;

③ 按照上节的对称选择算法组合各个基分类器的测试结果. 记下各个训练样本自测试的组合输出结果;

④ 将所有测试错误的正类样本放到正类训练集的后半部分,将所有测试错误的负类样本放到负类训练集的后半部分;

⑤ 按照最开始指定的数量顺序抽取训练集中的样本产生所有的训练子集对,完成新的分解.

上述算法流程中,只有步骤③和步骤⑤是需要串行完成的. 步骤③就是上节的对称选择组合过程. 步骤⑤则只是涉及到简单的样本交换.

大量的实验结果表明,自重组学习的最终训练误差不是严格单调下降的,而是在一个较低的水平范围内发生连续的轻微震荡. 此时的训练误差上界一般来说远远低于初始的训练误差,从而改善了学习效果. 对于自重组学习的另一个问题是,如何确定最佳的自重组次数 M,我们建议在训练误差减少到趋于稳定时,停止自重组学习过程.

5　仿真实验

使用 UCI 数据库[12]的两个二类数据集,Census Income 和 Spambase 进行实验. 各个数据集的样本数量类别信息如表 1 所示. 针对两个类别中样本数量较小的那个类别的训练集,依次分为 2～10 个大小相当的模块,较大的类别进行对应的划分,使其分出的每个模块的样本数量和较小类别分出的各模块的样本数量相当. 所有的任务分解的初始状态均为随机划分. 采用 RBF 核函数的 SVM 算法用于实验. 两个训练参数 C 和 γ 均分别取为 8 和 0.025.

表 1 数据集的样本数量分布

数据集	训练集样本数		测试集样本数	
	正类	负类	正类	负类
Census Income	15102	5005	7552	2503
Spambase	2091	1359	697	454

5.1 最小最大组合和对称选择组合的组合精度比较

表 2 比较了不同模块规模时的两种不同组合策略的分类精度. 可以看到,对称选择几乎在所有情形都获得了和最小最大组合至少相当甚至更佳的组合精度. 因此,它能够作为后续的自重组算法的可靠基础.

表 2 最小最大组合和对称选择组合的精度比较 %

较小类别中的模块数	Census Income		Spambase	
	最小最大	对称选择	最小最大	对称选择
2	77.45	77.46	76.37	76.37
3	77.25	77.25	75.67	75.67
4	76.86	76.89	76.28	76.20
5	76.95	76.95	75.50	75.50
6	76.86	76.86	74.11	74.11
7	77.13	77.16	75.33	75.33
8	77.28	77.31	75.07	75.33
9	76.91	77.00	74.98	75.07
10	76.92	76.94	74.11	74.11

5.2 多轮自重组学习后的测试精度比较

图 1 是 Census Income 数据集在多轮自重组学习后的测试精度比较.

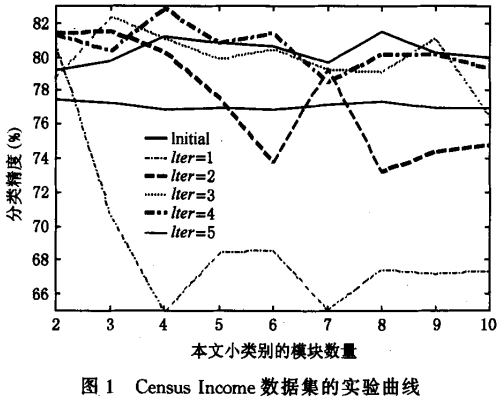

图 1 Census Income 数据集的实验曲线
(基分类器用 SVM)

图 2 是 Spambase 数据集在多轮自重组学习后的测试精度比较. 最高使用了 10 轮自重组.

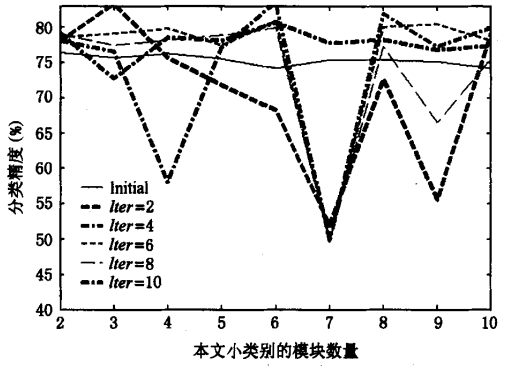

图 2 Spambase 数据集的实验曲线(基分类器用 SVM)

实验结果显示,随着自重组次数的增加,组合分类结果的精度在提升的同时,也越来越稳定,而较少地受到模块数量增加的负面影响.

5.3 自重组学习的训练精度和测试精度比较

将 Census Income 数据集的两个类都划分为 625 个样本或者相当大小的若干个子训练集(625 是将该训练集较小类别分为 8 份的结果),进行 91 轮的自重组学习. 将 Spambase 数据集的两个类都划分为 226 个样本或者相当大小的若干个子训练集(226 是将该训练集较小类别分为 6 份的结果),进行 41 轮的自重组学习.

图 3 是 Census Income 数据集各轮次的自重组后的测试精度和训练精度的比较.

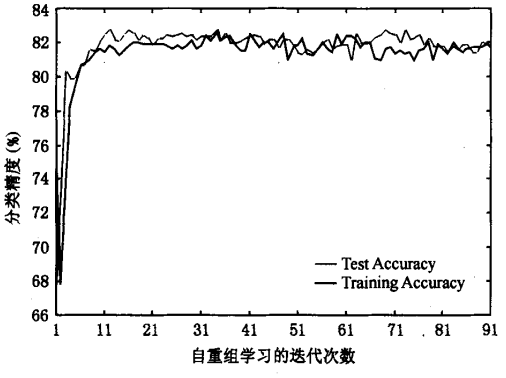

图 3 Census Income 数据集在不同次数的自重组学习后训练精度和测试精度比较(基分类器用 SVM)

图 4 是在 Spambase 数据集各个轮次的自重组后的测试精度和训练精度的比较.

从各种情形下的训练精度和测试精度的对比来看,两者基本上保持了一致的变化趋势. 这说明的确可以利用训练精度变化程度作为自重组学习的中止判据.

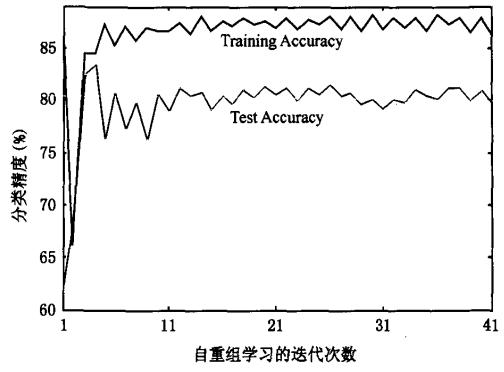

图 4 Spambase 数据集在不同次数的自重组学习后训练精度和测试精度比较(基分类器用 SVM)

6 结束语

本文在最小最大模块化分类器的任务分解策略的基础上,提出了一种基于对称选择组合的自重组学习算法. 任务分解策略继承了最小最大模块化分解的弹性特征,易于实现灵活分解和平衡化的学习. 同时,和传统的分布式分类方法相比较,提出的方法具有较好的并行处理特性和测试性能. 此外,提出的方法是适应于多种分类算法实现. 实验结果验证了提出算法的有效性.

在进一步的研究中,我们希望能够找到更快速更为稳定的交换算法,进一步探索有效的重组中止判据,并从理论上深入分析算法的动态特性,对于其收敛性给出严格的理论证明.

参 考 文 献

1 B. L. Lu, M. Ito. Task decomposition and module combination based on class relations: A modular neural network for pattern classification, IEEE Tran. Neural Networks, 1999, 10: 1244~1256

2 G. Giacinto, F. Roli. Automatic design of multiple classifier systems by unsupervised learning, The 1st Int'l Workshop, MLDM'99, Leipzig, Germany, 1999

3 D. Frosyniotis, A. Stafylopatis, A. Likas. A divide-and conquer method for multi-net classifiers. Pattern Anal. Applic, 2003, 6: 32~40

4 N. Chawla, S. Eschrich, L. O. Hall. Creating ensembles of classifiers. University of South Florida, Tech. Rep.: ISL-01-01. http://isl.ceee.usf.edu/report, 2001

5 L. Breiman. Bagging predictors. Machine Learning, 1996, 24 (2): 123~140

6 R. E. Schapire. Theoretical views of boosting. The 12th Annual Conf. Computational Learning Theory, Santa Cruz, CA, 1999

7 K. A. Wang, H. Zhao, B. L. Lu. Task decomposition using geometric relation for min-max modular SVMs. In: Advances in Neural Networks-ISNN2005, LNCS 3496. Berlin: Springer-Verlag, 2005. 887~892

8 B. L. Lu, K. A. Wang, M. Utiyama, *et al*. A part-versus-part method for massively parallel training of support vector machines. The IJCNN'04, Budapast, 2004

9 H. Zhao, B. L. Lu. A modular k-nearest neighbor classification method for massively parallel text categorization. The Int'l Symposium on Computational and Information Sciences(CIS'04), Shanghai, 2004

10 F. Y. Liu, K. Wu, H. Zhao, *et al*. Fast text categorization with min-max modular support vector machines. IEEE/INNS Int'l Joint Conf. Neural Networks (IJCNN2005), Montréal, Québec, Canada, 2005

11 B. L. Lu, J. Shin, M. Ichikawa. Massively parallel classification of single-trial EEG signals using a min-max modular neural network. IEEE Trans. Biomedical Engineering, 2004, 51(3): 551~558

12 C. L. Blake, C. J. Merz. UCI repository of machine learning databases. Irvine, CA: Department of Information and Computer Science University of California. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998

赵海,1976 年生,博士,主要研究方向为组合分类器、非监督学习和算法复杂度.

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,主要研究方向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.