改进的 AdaBoost 算法与 SVM 的组合分类器

李亚军,刘晓霞,陈 平
LI Ya-jun,LIU Xiao-xia,CHEN Ping

西北大学 信息科学与技术学院,西安 710127
Institute of Information Science and Technology,Northwest University,Xi'an 710127,China
E-mail:liyajun22@163.com

LI Ya-jun,LIU Xiao-xia,CHEN Ping.Combined classification algorithm based on improved AdaBoost and SVM.Computer Engineering and Applications,2008,44(32):140-142.

Abstract: A combined classification algorithm based on improved AdaBoost and Support Vector Machine,is proposed in order to deal with the problems of multiclass classification.Adopt a rule sampling to solve the unbalance of samples in the SVM.Improving the AdaBoost makes it consider the importance of sparse sample distribution at the beginning,this is advantageous to the right demarcation of rare sample.Experiment proves this algorithm can raise the generalization ability compared with the standard SVM.
Key words: AdaBoost;Support Vector Machine(SVM);combined classification;rule sampling

摘 要:提出了一种改进的 AdaBoost 算法与支持向量机组合的分类方法,用来处理多类别分类。采用规则抽样来解决支持向量机分类中正负样本的不平衡性,改进 AdaBoost 算法,使其在初始化时考虑样本分布稀疏的重要性,有利于稀有类样本的正确划分。实验结果表明,此方法与标准支持向量机分类器相比,泛化性能有一定程度的提高。
关键词:AdaBoost;支持向量机;组合分类器;规则抽样
DOI:10.3778/j.issn.1002-8331.2008.32.042 **文章编号**:1002-8331(2008)32-0140-03 **文献标识码**:A **中图分类号**:TP391

1 引言

支持向量机(Support Vector Machine,SVM)已经成为一种备受关注的分类器。支持向量机的训练问题,实质上是一个凸二次规划(Convex Programming)问题。这种技术具有坚实的统计学理论基础,并在许多实际应用(如手写数字的识别、文本分类等)中展示了大有可为的实践效用,另外,SVM 可以很好地应用于高维数据,避免了维灾难问题。

组合方法通过分类学习产生多个基分类器,并依照某种策略组装合成,所得组合分类器的判定结果依赖于单个基分类器的判定结果,因为利用了多个基分类器之间的多样性,所以能够降低分类误差。

AdaBoost[1]算法是组合方法中的代表算法,它是一个迭代的过程,用来自适应地改变训练样本的分布,使得基分类器聚焦在那些很难区分的样本上。

2 问题的分析及新算法的提出

2.1 分析问题

目前,利用 SVM 来处理多类分类的方法主要有:一对一(1-1)、一对其他(1-r)和纠错输出编码(ECOC)[2]。这些方法本质上都是扩展二元分类器的技术。

本文讨论使用 SVM(1-r)技术作基分类器的情况,这种方法产生的分类器数量少,且算法简单易行,复杂度小。但这种方法也存在一些缺点:

(1)将其中的一类作为正类,其他类作为负类,导致了训练样本类别不平衡;这种偏差尤其不利于小类别的分类[3]。

(2)负样本覆盖多个类别,在特征空间分布过于离散,不利于最优超平面的划分。

(3)每次学习基分类器时要训练所有样本,训练速度慢。

(4)对稀有类处理精度弱,因为稀有样本的处理容易导致超平面向一侧偏移。

针对这些问题本文采用抽样训练的方法,对训练数据抽样的处理可以使不平衡问题平衡化,抽样是解决不平衡学习问题的有效方法,用子集来训练还可以避免每次学习基分类器时都要训练所有的样本,除此之外,给抽样制定某种规则来解决特征空间中负样本过于离散的问题。本文将 AdaBoost 应用于 SVM 多类分类中[4],因为 AdaBoost 的本质就是抽样处理,从原始数据集中提取出自助样本集,进行自适应的多轮迭代,不同的是本文将改进 AdaBoost 算法中的随机抽样方法,采用规则抽样,以提高分类器泛化能力,这种改进的方法称为 IAdaBoost 算法。

基金项目:陕西省自然科学基础基金项目(the Natural Science Foundation of Shaanxi Province of China under Grant No.2006F50);航空科学基金项目(No.06ZC31001)。
作者简介:李亚军(1983-),男,硕士研究生,主要研究方向:信息处理;刘晓霞(1965-),女,硕士生导师,主要研究方向:信息处理、图形图像处理;陈平(1982-),女,硕士研究生,主要研究方向:信息处理、Web 挖掘。
收稿日期:2007-12-11 修回日期:2008-03-19

2.2 AdaBoost 算法

令$\{(x_i,y_i)|i=1,2,\cdots,N\}$表示包含 N 个训练样本的集合。在 AdaBoost 算法中,基分类器 C_i 的重要性依赖于它的错误率。初始时,样本的权重相等,在每一轮迭代中 AdaBoost 在每个样本上调整这些权重,基分类器在训练样本上的错误率被计算出来,并以此在训练样本上调整概率分布。调整样本概率分布的作用是在被误分的样本上设置更多的权重,在分类正确的样本上减少权重。组合分类器的最终结果通过取每个基分类器预测的加权平均得到。

算法 1 输入 N 个带标记实例的序列$<(x_1,y_1),\cdots,(x_n,y_n)>$,$N$ 个实例上的分布 D,训练基分类器的算法,迭代次数 T。

(1)初始化:对每个样本初始化相同的权值$\frac{1}{N}$;

(2)调整分布 $P_t=\frac{w_t}{\sum_{i=1}^{D} w_t}$;

(3)传递分布 P_t 给基分类器训练模型,返回预测 $x\rightarrow[0,1]$;

(4)计算预测差错率 $\varepsilon_t=\sum_{i=1}^{N} P_t|h_t x_t-y_t|$;

(5)计算基分类器的重要性 $\alpha_t=\varepsilon_t/(1-\varepsilon_t)$;

(6)计算新的权重向量 $w_{t+1}=w_t\alpha_t^{1-|h_t x_t-y_t|}$。

2.3 抽样规则

从上面可以看出,AdaBoost 没有考虑问题类别的不平衡性,AdaBoost 为样本维持着权重,根据权重大小进行抽样,不容易在迭代时改进,所以本方法在迭代之前为 AdaBoost 增加一个规则抽样过程。将所有样本放在欧氏空间中,距离少数类样本点较近的点通常是噪声的可能性较大,而较远的点则认为是较可靠的多数类样本。比如,在多类分类的 SVM 的 1-r 中,少数样本为正类,多数样本为负类。通过抽样得到可靠的与正类样本数目相当的负类样本集,这样有利于使样本集变得平衡。在多类分类的 SVM 的 1-r 中,负类样本比较多,且特征分散,会分布到正类样本周围,导致分类面向正类偏移,使得正类空间缩小,负类空间扩大,结果是将更多的样本分类为负类。所以本文通过规则抽样将分布在正类样本周围的负样本点也就是噪声去除,保留可靠的样本点。在数据挖掘中,常采用两个样本点之间的欧氏距离 $D(x,y)$来去除噪声。

$$D(x,y)=\|x-y\|^2 \tag{1}$$

在 SVM 多分类中,各类别中训练样本数多,不适宜此方法,本文采用类中心向量来近似某个类中样本点的分布。求取类中心,对于第 C_i 类,其类中心向量 $\boldsymbol{Center}_i$ 的计算公式为:

$$\boldsymbol{Center}_i=\frac{1}{N}\sum_{j=1}^{N_i}\boldsymbol{D}_{ij} \tag{2}$$

其中 N_i 是第 C_i 类中文本的数目,$\boldsymbol{D}_{ij}$ 是类别为 C_i 的第 j 个文本向量。

设正类样本集为 Y,负类样本集为 N,其中样本用 n_i 表示。定义负类中样本到正类集的距离为 $d(n_i,\boldsymbol{CenterY})$。

$$d(n_i,\boldsymbol{CenterY})=\|n_i-\frac{1}{|N|}\sum_{i=1}^{N}D_i\|^2 \tag{3}$$

$\boldsymbol{CenterY}$ 表示正类的类中心向量。计算负类集中每个样本到正类集的距离,从中选出距离最大的样本,也就是可靠样本,用这些样本和正类组成新的训练集,这样样本就可以变得平衡。

2.4 IAdaBoost 算法

IAdaBoost 算法使用 AdaBoost 算法的迭代思想来训练 SVM 基分类器。但是 AdaBoost 算法对稀有类建立分类器模型有一定的局限性[5],所以 IAdaBoost 算法相对 AdaBoost 算法作出了一些改进。除了使用规则抽样之外,还将样本的初始权重用样本所在类的规模来标记,表示为:$w_i=1/C_n$,这样稀有类中的样本具有较高的权值,在规则抽样中被选中的概率较大,在迭代过程中较容易被抽到,避免了分类器忽略稀有类的现象。

对多类 SVM 的 1-r,轮换将其中任意一类作为正类,其他类作为负类,若有 m 个类,进行 m 次以下步骤:

用公式计算正类的类中心向量,类中心向量为各个样本向量的平均;计算负类样本到正类集的欧氏距离;根据欧氏距离选取与正类数目相当的负类样本,做以下步骤:

1.$w=\{w_j=1/C_n|j=1,2,\cdots,N\}${初始化 N 个样本的权值,C_n 是样本所属类中样本的个数}

2.令 k 表示迭代的轮数

3.for i=1 to k do

4.根据 w,通过对 D 进行抽样(有放回)产生训练集 D_i

5.在 D_i 上训练基分类器 C_i

6.用 C_i 对原训练集 D 中的所有样本分类

7.根据 $\varepsilon_i=\frac{1}{C_n}\sum_{j=1}^{N} w_j\delta(C_i(x_j)\neq y_j)$ 计算加权误差

8.if $\varepsilon_i>0.5$ then

9.$w=\{w_j=1/C_n|j=1,2,\cdots,N\}${重新设置 N 个样本的权值}

10.返回 4

11.end if $\alpha_i=\frac{1}{2}\ln\frac{1-\varepsilon_i}{\varepsilon_i}$

12.根据 $w_i^{j+1}=\frac{w_i^j}{Z_j}\times\begin{cases}e^{-\alpha_j}\\ e^{\alpha_j}\end{cases}$ 调整每个样本的权值

13.end for

$C^*(x)=\arg\max\sum_{i=1}^{k}\alpha_i\delta(C_i(x)=y)$,{根据基分类器的预测加权得到组合分类器的最终结果}

3 实验及结果

采用 SoGou 上提供的中文语料集[5],从中选取 6 个类别,分别是经济、旅游、医药、体育、新闻和教育。训练样本和测试样本没有交集,各个类别的样本也互不交叉。文本分布如表 1 所示。

表 1 各个类所对应的标号

类标号	1	2	3	4	5	6
类别	经济	旅游	医药	体育	新闻	教育
(样本数)	200	300	102	175	75	125

实验主要对 IAdaBoost 的性能与 SVM 及 AdaBoostSVM 的性能进行比较。分类均采用固定参数 σ 值的 RBFSVM 作为 AdaBoost 基分类器,参数 σ 取值为 12,惩罚参数 C 取值为 1 000,样本维数取值为 1 000,IAdaBoost 算法和 AdaBoost 算法迭代次数为 10。

本文使用的评测指标为准确率 Precision 和综合评价指标 F1[6]。由于本文所选类别样本数目都不大,在规则抽样时,不对正例进行抽样,利用欧氏距离选择负类样本时,选择的样本数和正例类数目相当即可。

图 1 描述了改进前后的 AdaBoost 算法与 SVM 的组合分

类器以及标准 SVM 对各个类别分类的 F1 值：

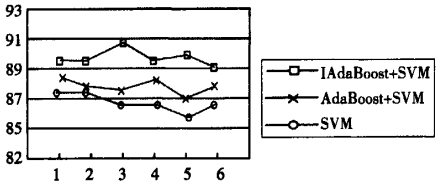

图 1 对各个类别的分类结果的比较

从三种方法处理这 6 个类别的整体性能可以看出，IAdaBoost 在处理新闻和医药这两个类别时表现了明显的优势，分析其原因，当以新闻类为正类，其余类为负类时，出现了明显的样本分布不均衡（正类 75 个，负类 902 个），使得其余两种方法的分类精度减弱，而 IAdaBoost 的规则抽样却可以很好地避免精度的降低。

为了避免训练样本数对训练结果的偏差，从经济类中选取训练样本集的大小分别为 50、80、100、130、150。图 2 是三种方法在样本大小不同的情况下分类的性能。

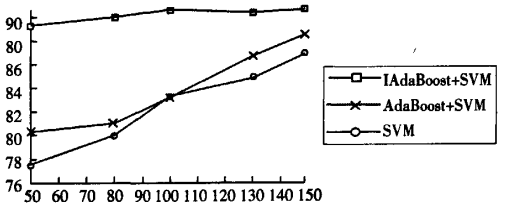

图 2 对不同样本的分类结果

从图中可以看出 IAdaBoost 算法对提高稀有类数据集的效果更有效。

4 结论

SVM 多类分类具有样本规模大、样本类别不平衡且等同地对待所有样本的特点，训练集中存在的噪音不但增加了训练负担，而且试分类器的泛化能力下降。另外，随着训练规模的增大，任一类别的反例数目远远大于正例，大多数反例对分类不起作用，降低了分类效果，还造成训练器过于复杂。利用本文提出的规则抽样不但可以减少训练样本，降低训练规模，解决样本类别不平衡的问题，还可以去除一部分噪音，选择可靠样本点进行训练。另外改进的 AdaBoost 算法的初始化又可以提高稀有类样本的权值，通过自适应地调整权值在一定程度上有利于稀有类样本的正确分类。

参考文献：

[1] 琚旭，王浩，姚宏亮.基于 Boosting 的支持向量机组合分类器[J].合肥工业大学学报，2006(10)：1220-1222.

[2] Tan Pang-Ning，Steinbach M，Kumar V.Introduction to data mining[M].[S.l.]：Posts & Telecom Publishers Inc，2006.

[3] Chew Hong-Gunn，Crisp D J，Bogner R E，et al.Target detection in radar imagery using support vector machines with training size biasing[C]//Proceedings of the Sixth International Conference on Control，Automation，Robotics and Vision，Singapore，2000.

[4] 王元珍，乐树彬.基于 MultiBoost 的最小分类误差算法[J].小型微型计算机系统，2005(11)：1948-1950.

[5] Joshi M V，Agarwal R C，Kumar V.Predicting rare classes：Can boosting make any weak learner strong?[C]//Proceedings of the Eighth ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD2002)，Edmonton，Canada，2002.

[6] 董乐红，耿国华，周明全.基于 Boosting 算法的文本自动分类器设计[J].计算机应用，2007(2)：384-386.

（上接 115 页）

DYMO 两种协议的平均端到端时延稳定下来。仿真实验表明：当移动速度大于 10 m/s 后，链路中断造成的中间节点分组丢失主要是受到经过较长路径的分组影响。也就是说，已交付分组的平均跳数快速下降。因此，当高速交付分组几乎全部沿着最短路径传递过来之后，端到端的平均时延对移动性增强就变得反应迟钝。另外从图 4 中所示，随着移动性的变化，路由载荷的性能表现几乎与端到端时延类似。MDYMO 协议路由载荷性能较 DYMO 提高达 20%以上。

6 结论

通过对 DYMO 协议的适当改进，本文得到了多路径 MDYMO 协议。MDYMO 协议对 DYMO 的主要改进在于，MDYMO 能在一次路由发现过程中找到源、目的节点间的多条链路不相交路由，并把这些路由作为后备路由；在节点间链路断开后，调用这些后备路由，直到所有路由均断开，而不是象 DYMO 一样，一条链路断开直接启动新一轮路由发现过程。

通过在 NS 下仿真表明，得到的 MDYMO 协议较 DYMO，显著提高了网络性能。MDYMO 对 PDF 的提高达到 3%以上，对平均端到端时延缩短超过 100%，标准化路由载荷降低达到 20%以上，这表明改进后的 MDYMO 协议明显优于 DYMO 协议，这也证实了多路径路由在移动 Ad Hoc 环境下性能要显然优于单路径路由协议。

参考文献：

[1] Cidon I，Rom R，Shavitt Y.Analysis of multi-path routing[J].IEEE/ACM Transactions on Networking，1999，7(6)：885-896.

[2] Marina M K，Das S R.Ad hoc on-demand multipath distance vector routing[J].Wireless Communications and Mobile Computing，2006，6(7)：969-988.

[3] Lee S，Gerla M.Split multipath routing with maximally disjoint paths in ad hoc networks[C]//Proceedings of the IEEE ICC，2001：3201-3205.

[4] Chakeres I，Perkins C.Dynamic MANET On-demand(DYMO) routing[S/OL].2006.INTERNET-DRAFT draft-ietf-manetdymo-06.txt. http://www.ietf.org/internet-drafts/draft-ietf-manet-dymo-06.txt.

[5] 陈林新，曾曦，曹毅.移动 Ad Hoc 网络——自组织分组无线网络技术[M].北京：电子工业出版社，2006：350-351.

[6] Fall K，Varadham K.The ns manual[R/OL].http://www.isi.edu/nsnam/ns/ns-documentation.html.

[7] Xu D，Chen Y，Xiong Y，et al.On the complexity of and algorithms for finding the shortest path with a disjoint counterpart[J].IEEE/ACM Transactions on Networking，2006，14(1)：147-158.

改进的AdaBoost算法与SVM的组合分类器

作者： 李亚军， 刘晓霞， 陈平， LI Ya-jun， LIU Xiao-xia， CHEN Ping
作者单位： 西北大学,信息科学与技术学院,西安,710127
刊名： 计算机工程与应用 ISTIC PKU
英文刊名： COMPUTER ENGINEERING AND APPLICATIONS
年，卷(期)： 2008,44(32)
被引用次数： 5次

参考文献(6条)

1. 琚旭;王浩;姚宏亮 基于Boosting的支持向量机组合分类器[期刊论文]-合肥工业大学学报 2006(10)

2. Tan Pang-Ning;Steinbach M;Kumar V Introduction to data mining 2006

3. Chew Hong-Gunn;Crisp D J;Bogner R E Target detection in radar imagery using support vector machines with training size biasing 2000

4. 王元珍;乐树彬 基于MuttiBoost的最小分类误差算法[期刊论文]-小型微型计算机系统 2005(11)

5. Joshi M V;Agarwal R C;Kumar V Predicting rare classes:Can boosting make any weak learner strong? 2002

6. 董乐红;耿国华;周明全 基于Boosting算法的文本自动分类器设计[期刊论文]-计算机应用 2007(02)

本文读者也读过(8条)

1. 张晓龙.任芳.ZHANG Xiao-long.REN Fang 支持向量机与AdaBoost的结合算法研究[期刊论文]-计算机应用研究 2009,26(1)

2. 王晓丹.孙东延.郑春颖.张宏达.赵学军.WANG Xiao-dan.SUN Dong-yan.ZHENG Chun-ying.ZHANG Hong-da.ZHAO Xue-jun 一种基于AdaBoost的SVM分类器[期刊论文]-空军工程大学学报（自然科学版） 2006,7(6)

3. 周维柏.李蓉.ZHOU Wei-bai.LI Rong 基于改进的AdaBoost和支持向量机的行人检测[期刊论文]-昆明理工大学学报（理工版） 2010,35(6)

4. 朱信忠.唐金良.徐慧英.赵建民.ZHU Xin-zhong.TANG Jin-liang.XU Hui-ying.ZHAO Jian-min 基于SVM和AdaBoost的人脸检测算法[期刊论文]-微型电脑应用 2009,25(9)

5. 李广群.王志海.田凤占.LI Guang-qun.WANG Zhi-hai.TIAN Feng-zhan 一种基于AdaBoost方法的树形HNB组合分类器[期刊论文]-广西师范大学学报（自然科学版） 2007,25(4)

6. 梁竞敏.UANG Jing-min 集成学习SVM在图像检索中的应用[期刊论文]-计算机工程与应用 2009,45(18)

7. Chang Tiantian.Liu Hongwei.Zhou Shuisheng Large scale classification with local diversity AdaBoost SVM algorithm[期刊论文]-系统工程与电子技术（英文版） 2009,20(6)

8. 朱信忠.唐金良.徐慧英.赵建民.ZHU Xin-zhong.TAN Jin-Liang.XU Hui-ying.ZHAO Jian-min 基于SVM和AdaBoost的人脸检测算法[期刊论文]-微型电脑应用 2009,25(4)

引证文献(5条)

1. 董璇.蔡立军 一种改进的少数类样本识别方法[期刊论文]-微型机与应用 2012(18)

2. 秦富童.杜静.杨奇才.刘迎龙 组合分类模型在雷达干扰效果评估中的应用[期刊论文]-计算机应用与软件 2014(1)

3. 霍亮.杨柳.张俊芝 贝叶斯与k-近邻相结合的文本分类方法[期刊论文]-河北大学学报（自然科学版） 2012(3)

4. 周维柏.李蓉 基于改进的AdaBoost和支持向量机的行人检测[期刊论文]-昆明理工大学学报（理工版） 2010(6)

5. 张磊.赵晓安.于明 改进的AdaBoost在表情识别中的应用[期刊论文]-微型机与应用 2012(21)