基于 Cascade 的增量式组合分类算法研究

欧吉顺[1],朱玉全[1],陈 耿[2],刘 晟[1]
OU Ji-shun[1],ZHU Yu-quan[1],CHEN Geng[2],LIU Sheng[1]

1.江苏大学 计算机科学与通信工程学院,江苏 镇江 212013
2.南京审计学院 省级审计信息工程重点实验室,南京 210029
1.School of Computer Science and Telecommunications Engineering,Jiangsu University,Zhenjiang,Jiangsu 212013,China
2.Jiangsu Key Laboratory of Audit Information Engineering,Nanjing Audit University,Nanjing 210029,China

OU Ji-shun,ZHU Yu-quan,CHEN Geng,et al.Research on combined classifier algorithm based on cascade.Computer Engineering and Applications,2009,45(31):165-167.

Abstract: The Learn ++ method is applied to improve the Cascade combined classifier,and is applied to lung images classification.The experiment results show that the incremental combined classification method can obviously improve the efficiency at the precondition that assure the accuracy compared with combined classifier.

Key words: multiple classifiers combination;incremental updating;Learn++;Boosting

摘 要:利用 Learn++思想对 Cascade 组合分类器进行了改进,提出了一种基于 Cascade 的增量式组合分类算法,并将之应用到肝脏图像的分类中。实验结果表明,与原有组合分类器相比,该增量式组合分类方法可以在保证分类准确度的前提下有效地提高新增样本的学习效率。

关键词:多分类器组合;增量式更新;Learn++;Boosting
DOI:10.3778/j.issn.1002-8331.2009.31.049 **文章编号**:1002-8331(2009)31-0165-03 **文献标识码**:A **中图分类号**:TP391

1 引言

目前,常用的分类方法有:基于决策树的分类方法、贝叶斯分类方法、基于神经网络的分类方法、K-最邻近分类方法、基于关联规则的分类方法等。这些方法的性能与其所采用的样本数据种类和复杂性等因素有关,各种方法都有其相应的优点,同时也存在着一些算法自身无法克服的缺陷,如:基于决策树的分类算法具有速度快、分类规则易于理解、准确率相对较高等优点,但由于算法本身的不稳定性,不同的样本初值可能会得出不同的结论,另外,ID3 算法偏向于属性值较多的属性,对噪声也较为敏感;朴素贝叶斯分类方法要求属性值之间是相互独立的,在许多场合,该方法可以与决策树和神经网络等分类方法相媲美,但贝叶斯定理假设一个属性对给定类的影响独立于其他属性,在某些情况下是很难做到的;神经网络方法虽然具有非线性学习、联想记忆的优点,但存在的问题是神经网络系统是一个黑盒子,不能观察中间的学习过程,最后的输出结果也较难解释,影响结果的可信度及可接受程度;K-最邻近分类是一种懒散的学习方法,对未知和非正态分布的数据可取得较高的分类准确率,具有鲁棒性强、概念清晰等优点,但对于高维的数据,该方法的缺陷也得以凸显;基于关联规则的分类方法主要是通过发现训练集中的关联规则来构造分类器,其优点是分类准确度较高,其缺点是为了获得所有的分类规则,最小支持度通常设置得很小或干脆为 0,有可能产生过多的频繁项目集,从而使得程序无法继续运行,实际操作时必需考虑精度和时间的平衡问题。由此可见,设计一个能达到理想分类精度的分类器仍是一个十分困难的问题,一个简单的解决方案是采用多个分类器分别进行分类,然后选择其中最好的一个作为最终解决方案,但不同分类器所产生的错误具有不同的偏向,被弃分类器中往往隐含着被选分类器中所没有的识别信息,信息利用率不高。为了扬长避短,许多学者提出了分类器组合思想及其相应的技术解决方案,组合分类器利用了各个分类器之间存在的信息互补性,充分发挥了各个分类器的优势,弥补了各分类器的缺陷。理论和大量的实践表明,组合分类器不但可以进一步提高分类器的识别率,而且可以增强分类系统的鲁棒性,已成为国内外知识发现与知识工程等领域中的一个研究热点,并已提出了许多组合分类算法[1-6]。

现有的组合分类算法(如 Cascade)可以有效地构造静态样本数据库中的组合分类器,然而,在实际应用中,样本数据是随时间的变化而变化的(包括增加新的属性和新的类别),当前已发现的某些分类规则可能不再生效,而可能存在新的有效分类规则有待于进一步去发现,因此,不仅设计高效的算法来构造

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60572112);江苏省高技术项目(No.BG2007028);江苏省六大人才高峰项目(No.07-E-025);江苏省教育厅项目(No.06KJB120051)。

作者简介:欧吉顺(1983-),硕士研究生,主要研究方向:人工智能和知识发现等;朱玉全(1965-),博士,副教授,主要研究方向:数据库系统及其应用、人工智能和知识发现等。

收稿日期:2008-06-18 **修回日期**:2008-10-23

组合分类器,而且也迫切需要设计高效的算法来更新、维护和管理已发现的分类规则,能否有效地构造出动态样本数据库中的分类器是衡量一个算法好坏的关键因素。

该文主要考虑一组新样本数据集 $\bigcup_{i=2}^{k} D_i$ 添加到样本数据库 D_1 中去时,如何构造最新样本数据库 $D=\bigcup_{i=1}^{k} D_i$ 中的组合分类器。对于该问题,目前可用的方法是在 D 上重新运行组合方法一次,可想而知,其运算量是非常大的。在深入研究组合分类算法 Cascade 的基础上,提出了一种基于 Cascade 的增量式组合分类算法,该算法将充分利用已有信息来发现最新的分类规则。

2 问题描述

2.1 Cascade 算法思想

Cascade 组合算法的主要思想是:每层包含一个分类器,各个成员分类器顺序执行,第一层分类器的输入为初始训练集,以后每层都在上一步基础上对初始训练集属性进行扩充,将上一层得出的各类相应的类别概率估计作为新的属性,最后一层输出为最终决策结果。在 Cascade 组合算法中,各阶段都对初始数据集进行了属性扩充,高层分类器的样本信息量得到了增加,误差偏置明显降低,从而可以取得较好的分类效果。

2.2 分类规则的更新

设待学习的样本是未知分布 D,$D=\bigcup_{i=1}^{k} D_i(k\geqslant 2)$,其中来自 D_1 的样本作为训练初始组合分类系统的旧样本集,来自于 D_i($i=2,\cdots,k$)的样本是作为需要增量学习的新样本集合。分类规则的更新问题是指给定 D 以及样本数据库 D_1 上的分类规则,如何确定最新训练集 D 上的分类规则。

3 增量式 Cascade 组合分类模型

3.1 Cascade 组合增量模型

利用 Learn++[7]增量分类思想对 Cascade 组合模型进行了改进,改进后的增量式 Cascade 组合分类模型如图 1 所示。为了便于阐述,图 1 中仅给出了基于 Bayes 和 BP 算法的 Cascade 组合模型,其他可作同样处理。

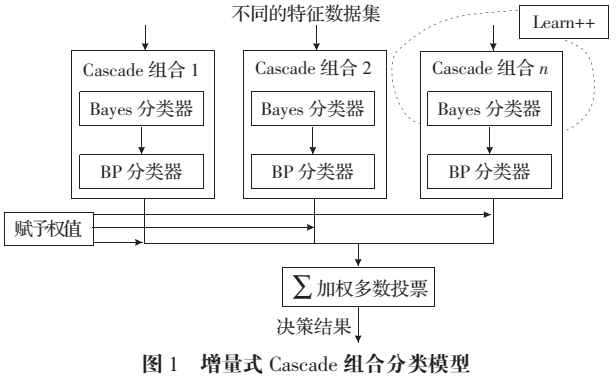

图 1 增量式 Cascade 组合分类模型

3.2 算法描述及分析

算法 1 基于 Cascade 的增量式组合分类算法。

输入:样本数据集 D_i,$i=1,2,\cdots,k$

输出:分类规则

方法:

(1)$D_{novel}=\Phi$;//用于存放边界样本的集合,初始化为空

(2) for ($i=1;i<=k;k++$) do begin

(3) $D_i=D_i\cup D_{novel}$;

(4) 从 D_i 中随机选择训练集 TR_i 和测试集 TE_i;

(5) 利用组合分类算法 Cascade 进行训练,训练集来自于分布 D_i 的 TR_i;

(6) 产生一个假设 $h_i:X\to Y$,计算在 TR_i+TE_i 上的误差 $h:\varepsilon_i=\sum_{i:h_i(x_j)\neq y_j} D_i(j)$及比例误差 $\beta_i=\varepsilon_i/(1-\varepsilon_i)$;

(7) 修改分布 D_i:$D_{i+1}(j)=\dfrac{D_i(j)}{\sum_j D_i(j)}\times\begin{cases}\beta_i, \text{当 } h_i(x_j)=y_j\\ 1, \text{否则}\end{cases}$,其中 $\sum_j D_i(j)$是一个正交化常量;

(8) 把权值为 1 的样本保存至 D_{novel} 中,去除冗余的;

(9) end;

(10) 调用加权多数算法,输出最后的假设:$H_{final}(x)=\arg\max_{y\in Y}\sum_{h_i(x)=y}\log\frac{1}{\beta_i}$;

与 Learn++算法相比,增量式 Cascade 组合分类算法有如下两大优点:

(1)算法的第(4)步。一种学习算法要达到一定的分类效果,需要一定数量的样本集及合适的正反例分布,如果增量样本集太少或者不能提供一定数量的不同类别的样本时,会影响分类算法的学习效果。特别是当增量式学习所提供的样本都属于某一类样本时,即使有足够的样本数,但并不能真正学习到所要的知识。为了解决该问题,在从 D_i 中随机选择训练集 TR_i 和测试集 TE_i 时,采用在小样本或单类数据集中添加一些合适的样本,以便能够得到更好的学习效果,又能使计算和存储方面增加的开销很小。由于分类主要是寻找类与类之间的最佳区分边界,只要能够给出与新样本临界的反例就可以形成新样本的表示区域。在分类中,被错分的样本一般都是处于决策边界的样本,当多类相互都有邻接时,只要某个决策边界发生变化,就可能使临界的样本划分到不同区域中,错分为不同的类别。考虑在 Boosting 方法中,新增弱分类器主要加强对错分的样本的学习,通过对前一个分类器的性能评价,以错分样本为主形成构造新分类器时所采用的样本集,恰好能够过滤错分的样本。这样可以借助于 Boosting 生成新样本的过程来选择需要的反例,只要开辟一个存储区域,保存每个组合分类器错分的样本形成反例集合。这样用新增样本和错分样本作为样本集对于小样本和单分类样本集学习,能够得到更好的效果。由于反例集数目远小于学习样本的总数,所需的开销也不会很大,不会造成存储容量过高的结果。通过集中学习“较难学习”的样本,达到提高整个增量学习系统泛化能力的目的;

(2)算法的第(5)和(6)步。利用从第(5)步选出的训练集进行学习及预测时,采用了基于 Bayes 和 BP 算法的 Cascade 组合模型来代替原有的单分类算法进行训练。在 Learn++算法中,唯一的要求就是每个单分类器至少应当达到 50%以上的识别率,否则则放弃该分类器,重新选取训练集进行训练。利用 Cascade 组合分类器取代原有单分类器,可以提高分类准确度,满足 Learn++算法的要求,有效避免了重新训练所带来的时间等开销,这样该组合分类器的计算误差即是全局误差,显著降

低了模型的复杂度。

4　实验结果及分析

为了验证增量式 Cascade 组合分类算法的有效性，采用了 VC++6.0 开发工具，在方正工作站（运行环境为 2.93 G CPU、256 M 内存、Windows XP）上进行了测试。实验数据采用的是大小为 512×512 象素的二维肝脏 CT 图像。在实验中，从 10 000 幅肝脏图像中选出 400 幅，其中正常图像 200 幅，异常图像 200 幅。分别提取每张图像的灰度直方图特征（均值、方差、倾斜度、峰态、能量、熵）、灰度共生矩阵特征（能量、熵、惯性、局部平稳）、小波变换特征（均值、方差、倾斜度、峰态、能量）共 15 维进行实验，类别标识包括病例正常和异常两种。把 400 幅图像分成 S_1,S_2,S_3,S_4,S_5,S_6 六个训练数据集，其中 S_1 含有 100 个样本，作为旧数据集，其余数据集作为新增数据集，每个新增数据集中分别含有 50 个样本，各个数据集之间没有重复的样本。其余样本作为测试集 *Test*，用于增量系统的性能测试。

组合增量分类系统从 S_1 开始学习，产生一系列 Cascade 组合分类器，然后用测试集分别进行测试，在不更改新增样本分布的情况下，得到的学习结果如表 1 所示。

表 1　医学图像数据的组合增量分类

数据集	S_1	S_2	S_3	S_4	S_5	S_6
Test/(%)	85.6	86.4	87.1	89.2	88.7	90.6

每行代表新加入一个 Cascade 组合分类器后测试样本集的分类正确率。由表 1 可以看出，测试集的分类正确率随组合分类器的增加逐渐增加，充分说明这种多个 Cascade 组合分类器的增量学习是有效的，随着新增数据集的增多，参与增量学习的组合分类器随之增多，有效提高了组合增量系统的分类准确度。

按照上面所提出的样本分布改进方法，对该数据集进行了同样的测试，由于实验中采用的图像类别数目较少，在此仅考虑边界样本，即将前一个组合分类器中错误分类的样本加入到后一个训练集中，从结果上可以看出，组合分类系统的增量学习能力有所提高，如表 2 所示。

表 2　改进的组合增量分类

数据集	S_1	S_2	S_3	S_4	S_5	S_6
Test/(%)	85.6	87.2	89.5	90.6	92.4	94.3

实验结果表明，当学习新的数据集时利用了原样本集中的一些边界样本，使得在后面的组合分类系统中，对大多数原有样本可以正确识别，因此组合增量分类系统对原训练样本的分类正确率不会降低得太多。由于提高了整个组合分类系统对原有样本的识别率，因此在测试集上的分类正确率有较大的提高，既达到了学习新样本的目的，又使整个分类系统的泛化能力保持在一个较高的水平上。

另外，从训练所需时间上进行分析，进行增量式学习的目的就是有效利用已学习到的知识，在保证精度的前提下尽量避免重复计算，从而节省时间开销。文中所设计的组合增量分类系统的时间开销主要包括三个部分：将边界样本加入到新增数据集所需时间、对更改分布后的数据集进行学习的时间及进行预测时多数投票所需时间。对于边界样本可以在前一次进行分类学习保存结果时自动保存出来，这部分的时间基本可以忽略不计，在此主要对训练时间进行比较。

表 3　非增量式组合分类系统时间开销

	S_1	S_2	S_3	S_4	S_5	S_6
Time/s	16.3	18.3	26.8	33.1	35.8	43.8

表 4　组合增量分类系统时间开销

	S_1	S_2	S_3	S_4	S_5	S_6
Time/s	16.3	9.7	9.9	10.5	11.3	12.6

表 3 和表 4 分别给出了在加入新数据集之后利用非组合增量分类系统和增量式组合分类系统进行训练所需的时间开销对比。由于文中所采用的增量学习方法是在新增数据集中添加部分边界样本，所以新增样本进行训练时间与原有组合分类系统训练时间之和大于全部重新进行训练所需时间，但对于新增样本的学习时间远远小于全部重新学习的时间开销，达到了增量学习的目的。

图 2 给出了增量式和非增量式组合分类系统时间开销趋势对比。在增量式组合分类系统中，当有新数据集到来时，主要是对新增样本及边界样本的学习，所以时间开销相对较小，但随着新增数据集的不断增多，参与投票的组合分类器数增多，会引起时间开销的增加，不过由图可以看出其增加趋势明显小于非增量式学习系统。

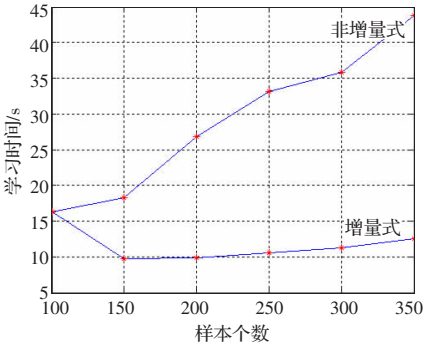

图 2　时间开销对比图

5　结束语

增量学习是各种分类算法所必须具有的能力，对于组合分类模型也不例外，研究组合分类的合理增量学习算法可以使得组合分类模型更好地应用于各个领域。该文提出了基于 Cascade 的增量式组合分类算法，该算法通过把新知识的样本作为重点学习的目标，对新增训练样本的分布进行改进，在对原有知识结构不加任何修改的基础上实现了知识的增量学习。实验结果证明这种增量式组合分类技术对于医学图像分类是有效的，在保证分类准确度的前提下有效提高了分类效率。

参考文献：

[1] Yizhak I，Chevalllier R C.Handwritten digit recognition by a supervised kohonen-like learning algorithm[C]//Proc of 1991 IEEE Joint Conf Neural Networks IJCNN，1991：2576-2581.

[2] Xu I，Krzyzak A，Suen C Y.Methods of combining multiple classifiers and their applications to handwriting recognition[J].IEEE Trans Systems，Man，and Cybernetics，1992，22(3)：418-435.

[3] Polikar R，Krause S，Burd L.Ensemble of classifiers based on incremental learning with dynamic voting weight ipdate[C]//Proceedings of the International Joint Conference on Neural Networks，2003，4：2770-2775.

（下转 189 页）