基于 Cascade 组合分类器的医学图像分类方法研究

张春芬[1],朱玉全[1],陈 耿[2],王 敏[1]
ZHANG Chun-fen[1],ZHU Yu-quan[1],CHEN Geng[2],WANG Min[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,南京 270029
1.School of Computer Science and Telecommunications Engineering,Jiangsu University,Zhenjiang,Jiangsu 212013,China
2.Nanjing Audit University,Nanjing 270029,China
E-mail:zhchunfen@163.com

ZHANG Chun-fen,ZHU Yu-quan,CHEN Geng,et al.Research on medical image classification based on Cascade combined classifiers.Computer Engineering and Applications,2007,43(36):211-213.

Abstract: Based on Cascade combination algorithm,two combined classifiers constructed by naïve bayes,BP neural network and decision tree are developed and applied to lung images classification.Compared with existing medical image classifications,experiment results indicate that this method can obviously improve the accuracy and stability of medical image classification.
Key words: multiple classifiers combination;Naïve bayes;BP neural network;decision tree

摘 要:提出了利用 Cascade 组合方法生成基于贝叶斯、神经网络与决策树的组合分类器,并将之应用到肝脏图像的分类中。实验结果表明,与现有医学图像分类方法相比,该组合方法可以有效地提高医学图像分类的准确性和稳定性。
关键词:多分类器组合;朴素贝叶斯;神经网络;决策树
文章编号:1002-8331(2007)36-0211-03 **文献标识码:**A **中图分类号:**TP391

1 引言

近年来,随着计算机、图形图像及生物工程等相关技术在医院信息化建设中的广泛应用,许多医院均已收集了大量的影像数据,如 CT、MR、SPECT、PET、DSA、超声图像、电阻抗图像等。在这些医学影像数据中,绝大部分影像已经被医生确诊,即已经知道确诊影像所属类别(正常、异常甚至何种异常),如何充分利用已确诊病例影像数据信息和医生的临床诊断经验来判断未确诊医学影像所属类别,辅助医生进行临床诊断,正是计算机辅助医学诊断系统要实现的一个重要目标。因此,医学图像分类方法的研究具有广泛的应用前景。

目前,医学图像识别领域已经存在许多分类方法[1],常见的有基于决策树的分类器、基于神经网络的分类器、基于关联规则的分类器等, 这些算法的性能与其所采用的图像特征种类、特征数据的复杂性等因素有关,存在着一些算法自身无法克服的缺陷,如基于决策树算法的分类器[2],虽然具有速度快、容易转化成分类规则等优点,但由于算法本身的不稳定性,不同的样本初值或特征空间可能会得到不同的结果;朴素贝叶斯分类要求属性值之间是相互独立的[3],这在很多情况下无法满足,其准确度难以有较大的突破。为此,许多学者提出了分类器组合思想, 多分类器组合利用各个分类器之间存在的信息互补性,充分发挥各个分类器的优势,从而提高了监督分类的精度[4],成为国内外模式识别领域研究的一个热点,尤其在人脸及字符识别等领域取得了不错的效果。见诸报道的有:Brunelli R、Falavign D 等人提出的基于音频和视频数据组合分类的人脸识别[5];南京理工大学杨静宇等人提出的多距离组合分类器及其在人脸识别中的应用[6];多特征多分类器组合的手写体数字识别系统[7]等。

本文在深入研究医学图像特征及各种分类方法的基础上,提出了基于 Cascade 组合分类器的医学图像分类方法,该方法可以有效地提高医学图像分类的准确性和稳定性。

2 组合分类器

分类是指从训练样本的属性中发现个体或对象的一般分类规则,并根据这些规则对非训练样本数据对象进行分类。其中,分类器的构造是分类的重点,也是难点,目前虽然已经出了许多不同种类的分类器,但不同分类器的分类机制不同,分类识别性能有所差别,在实际应用中,单个分类器的性能往往难以达到令人满意的程度。从 20 世纪 80 年代人们就开始了组合分类系统的研究,大量的实验和应用证明,通过某种组合技术将多个单分类器的预测进行组合,能够充分利用各单分类器提供的关于被分类对象的互补信息,得到比单个分类器更好的性能。这一方法引起了学者们的广泛关注,取得了较大的进展。

2.1 组合分类器的组合方法

分类器的组合方法有很多,根据所采用的分类算法是否相

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60572112)。
作者简介:张春芬(1982-),女,硕士研究生,主要研究方向:医学图像数据挖掘;朱玉全(1966-),男,副教授,硕士生导师,主要研究方向:医学图像数据库,数据挖掘等;陈耿(1965-),男,教授,主要研究方向:数据挖掘,数据库系统等;王敏(1980-),女,硕士研究生,主要研究方向:医学图像数据挖掘。

同，可以分为同种分类算法的组合和不同分类算法的组合；根据结构形式可分成串行、并行及混合型三种；根据组合的层次，可以分为两层及更多层次的组合，目前采用的多是两层结构，对于三层及更高层次的组合方式，较之于两层组合的优劣比较还有待于进一步研究[4]。

2.2　组合分类器的组合策略

分类器的组合不是随意进行的，其所组合的单分类算法必须具有一定的特点。大量的研究与实验结果表明，一般情况下，组合分类器中所采用的分类算法必须满足以下两个规则[4]：

(1)在较低的组合层采用误差漂移较低的分类算法；

(2)在较高的组合层采用误差偏置较低的分类算法。

误差偏置与误差漂移是衡量分类器的泛化能力与稳定性的两个指标[8]。偏置量越小，表明分类器的稳定精度越高，泛化能力越强；漂移量越小，表明分类器越稳定。学习算法的稳定性是指训练集的变化导致该算法所产生的预测函数发生变化的可能性及程度。

3　Cascade 组合算法及其在医学图像组合分类中的应用

与同种分类算法的组合相比，不同分类算法的组合能充分利用各分类算法的互补性，体现组合分类的多样性，更具有实用价值。Cascade 组合模型是其中较为简单有效的一种，该方法能够充分利用初始数据信息和单分类器的决策信息，取得较好的效果。在多层 Cascade 组合结构中，每层包含一个分类器，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，最后一层输出最终决策结果。

3.1　Cascade 组合算法[4]

定义 1 向量合并 设向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_m)$，向量 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)$，则 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 合并定义为：$\boldsymbol{X}\oplus\boldsymbol{Y}=(x_1,x_2,\cdots,x_m,y_1,y_2,\cdots,y_n)$。

给定初始训练集 L，测试集 T，分类算法 δ，利用对训练集 L 进行学习得到的分类器 $\delta(L)$。$\delta(L)$将输入向量 $\boldsymbol{X}$ 映射到输出变量 $\boldsymbol{Y}$，即：对应每个输入向量 $\boldsymbol{x}_i=(x_1,\cdots,x_m)$，$\delta(\boldsymbol{x},D)$输出一个 g 维变量$(p_1,\cdots,p_g)$，其中 g 为类别数，P_i 表示 $\boldsymbol{x}$ 属于类别 c_i 的概率。

$\varphi(\delta(L),T)$表示用分类器 $\delta(L)$对 T 进行分类，对于给定的两个分类算法 δ_1、δ_2，在 Cascade 组合分类器中，第一层分类器所采用的分类算法 δ_1 的训练集和测试集分别为 L 和 T，第二层分类算法 δ_2 的训练和测试数据分别为 $Level_2train$、$Level_2test$，其产生方法如下：

$$Level_2train=L\oplus\varphi(\delta_1(L),L),Level_2test=T\oplus\varphi(\delta_1(L),T) \quad (1)$$

综上所述，基于分类算法 δ_1、δ_2 的 Cascade 组合分类器为：

$$\delta_2\nabla\delta_1=\varphi(\delta_2(Level_2train),Level_2test) \quad (2)$$

其中，符号▽表示分类器组合中算法的先后顺序，将式(1)代入式(2)后：

$$\delta_2\nabla\delta_1=\varphi(\delta_2(L\oplus\varphi(\delta_1(L),L)),(T\oplus\varphi(\delta_1(L),T))) \quad (3)$$

N 层分类器的组合可表示为：

$$\delta_n\nabla\delta_{n-1}\nabla\delta_{n-2},\cdots,\delta_1 \quad (4)$$

第 N 层分类器的训练集和测试集分别增加了$(n-1)\times g$ 个属性，详细推导公式为：

$$\delta_n\nabla[\delta_{n-1},\delta_{n-2},\cdots,\delta_1]=\varphi(\delta_n(L\oplus\varphi(\delta_1(L),L)\oplus,\cdots,\oplus\varphi(\delta_{n-1}(L),L)),(T\oplus\varphi(\delta_1(L),T)\oplus,\cdots,\oplus\varphi(\delta_{n-1}(L),T))) \quad (5)$$

3.2　基于 Cascade 组合的医学图像分类

Cascade 组合分类器采用成熟的两层框架结构，朴素贝叶斯(Naïve Bayes)作为第一层分类器，第二层分类器分别采用 BP 神经网络算法[9,10]和决策树中比较成熟的 C4.5 算法，形成两层模型结构。由于基于 Cascade 组合的分类算法要求分类数据由特征向量组成，不能直接在原始图像上进行，所以在实施分类之前必须先对图像进行预处理，主要包括对图像本身的处理，如去噪、图像增强等，以及特征提取两个子过程。抽取的特征被组织在一个数据库中，作为分类系统的输入。具体处理过程如图 1 所示。

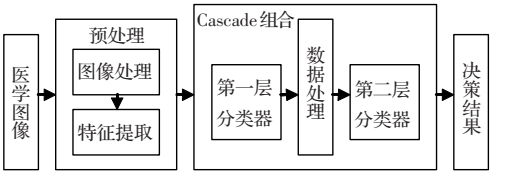

图 1　基于 Cascade 组合的医学图像分类框架

朴素贝叶斯算法假定所有的条件属性相对独立，尽管这一假定很大程度上限制了它的应用，但目前许多研究和应用都表明：即使违背这种假定，朴素贝叶斯也表现出很强的健壮性[3]，其坚实的数学基础和丰富的概率表达能力，也有利于 Cascade 组合。另外，朴素贝叶斯算法误差漂移较小，分类性能较为稳定，而误差偏置相对稍高，但 BP 神经网络和决策树算法误差漂移较高、误差偏置较低[4,8]，和贝叶斯算法可达到互补的效果，从理论上讲适于组合。

在两层分类器之间需要对数据进行处理，即将初始训练集和测试集分别与其第一层分类器预测结果进行合并(⊕)操作，这也是实现该组合方式的关键步骤之一。另外由于所采用的医学图像特征数据为数值型数据，数据值的范围较大，对于 BP 神经网络，为加快学习速度需进一步进行归一化处理。

在医学图像分析模型中，Naïve Bayes 的训练数据和测试数据均为医学图像特征，即初始输入数据集 L 和 T，第二层分类器的输入数据为处理后的数据，增加了两个新属性(该图像分别属于正常和异常的概率)，根据式(1)、(2)可知，第二层分类器的训练样本集为：

$$Level_2train=L\oplus\varphi(NaiveBayes(L),L)$$

测试样本集为：

$$Level_2test=T\oplus\varphi(NaiveBayes(L),T)$$

综上所述，基于 Cascade 组合的医学图像分类算法描述如下：

(1)对医学图像(本文选用肝脏图像)进行去噪、增强处理，分别提取其直方图特征、共生矩阵特征、小波变换特征，并进行离散化，分为训练集 L 和测试集 T，作为整个分类器的输入；

(2)利用贝叶斯算法对 L 进行学习，并对 L 和 T 进行测试，得出每个样本 $\boldsymbol{x}_i=(x_1,\cdots,x_m)$的类别概率估计$(P_1,P_2)$，并比较得出基于贝叶斯构造的单分类器决策结果；

(3)利用 BP(加入归一化处理)和 C4.5 算法分别对 L 进行学习，并对 L 和 T 进行测试，得出每个样本 $\boldsymbol{x}_i=(x_1,\cdots,x_m)$基于 BP 和 C4.5 构造的单分类器决策结果；

(4)将 L 和 T 中样本的初始属性向量$(x_1,\cdots,x_m)$与对应的由贝叶斯分类器得出的概率估计向量(P_1,P_2)合并，得到第二层训练数据 $Level_2train$ 和测试数据 $Level_2test$；

(5)利用 BP 和 C4.5 算法分别对 $Level_2train$ 进行学习，得

到两个同种组合类型的组合分类器 *BP*▽*Bayes* 和 *C*4.5▽*Bayes*,再分别对 *Level*$_1$*train* 和 *Level*$_2$*test* 进行测试,得出组合分类器针对各训练集和测试集的决策结果,作为最终输出结果。

4　实验结果及分析

为了验证基于 Cascade 组合的医学图像分类的有效性,采用 VC++6.0 开发工具,以 Microsoft Office Access 构建数据库,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 像素的二维肝脏 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。同时在所选出的图像中随机抽取正常图像 140 幅、异常图像 140 幅,共 280 幅作为训练库,其余作为测试库。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、相关)、Gabor 小波变换特征(均值、方差、能量)及将各种特征数据综合进行实验,决策类别为正常和异常。

实验中,BP 神经网络的输入层节点数依照所采用的特征数据集维数而定,隐含层节点数等于输入层节点数和输出层节点数的和,步长初值设为 0.01,在学习过程中采用变步长算法,训练达到系统目标精度 0.006 或最大训练次数 1 000 时则终止。C4.5 算法采用信息增益率来选择最佳分裂属性。表 1、表 2 分别给出了针对各个特征数据集构造的单分类器及 *BP*▽*Bayes*、*C*4.5▽*Bayes* 组合分类器的训练集及测试集分类精度。

表 1　采用的特征数据集及单分类器和组合分类器训练集分类精度　%

特征数据集	分类器				
	单分类器及分类精度			组合分类器及分类精度	
	Bayes	*C*4.5	*BPNN*	*BP*▽*Bayes*	*C*4.5▽*Bayes*
综合特征	79.29	78.57	85.00	85.36	87.50
直方图特征	83.57	86.43	86.07	90.71	91.43
灰度共生矩阵特征	84.29	89.29	90.00	92.86	90.71
Gabor 小波特征	83.93	87.86	85.71	90.35	89.29

表 2　采用的特征数据集及单分类器和组合分类器测试集分类精度　%

特征数据集	分类器				
	单分类器及分类精度			组合分类器及分类精度	
	Bayes	*C*4.5	*BPNN*	*BP*▽*Bayes*	*C*4.5▽*Bayes*
综合特征	77.50	78.33	85.83	88.47	86.67
直方图特征	82.50	87.50	85.00	90.83	91.67
灰度共生矩阵特征	84.17	88.33	89.16	92.50	90.00
Gabor 小波特征	83.33	85.83	87.50	89.17	88.33

从表 1 和表 2 可以看出,对于各个特征数据集,组合分类器的训练集分类精度和测试集分类精度都要优于单分类器。另外,从综合特征数据集和 Gabor 小波特征数据集的分类精度差别上也可以看出,由于采用的特征数据集不同,单分类器所得出的分类精度上下浮动稍大,而组合分类器则比较集中,可见在分类器稳定性上组合分类器较单一分类器也要略胜一筹。组合分类器分类步骤较多,执行分类所用时间自然要比单分类器长,但目前医学图像自动分类作为医生诊断或历史病情分析的辅助系统,对准确度的要求远远高于对速度的要求。

表 3 给出了两种组合分类器的时间开销,由于 BP 神经网络算法本身的复杂性及需要对数据进行归一化处理,*C*4.5▽*Bayes* 组合分类器的平均计算速度要稍优于 *BP*▽*Bayes* 组合分类器。

表 3　两种组合分类器时间开销对比表　s

	BP▽*Bayes*	*C*4.5▽*Bayes*
平均分类时间	41	33

5　结束语

医学图像分类至今已有多年历史,期间也取得了许多成果,但还存在许多问题有待进一步的研究。文中尝试提出一种多分类器组合的医学图像分类,首先利用朴素贝叶斯分类器得出图像样本分别属于正常和异常两个类别的概率,再将该决策信息与初始数据合并作为下层分类器的输入,进行训练和测试,利用分类算法的互补性克服单分类器的缺陷,达到了较高的分类精度和稳定性,充分说明了这种组合方法对于医学图像分类的有效性。针对医学图像分类,为了获得更好的分类效果和分类速度,基于多特征的多分类器的组合及多类别的医学图像分类都是今后的工作重点。(收稿日期:2007 年 5 月)

参考文献:

[1] Antonie M I,Zaiane O R,Coman A.Application of data mining techniques for medical image[C]//Proceedings of the Second International Workshop on Multimedia Data Mining(MDM/KDD'2001),in Conjunction with ACM SIGKDD Conference,USA,2001.

[2] 王曙燕,耿国华,李丙春.决策树算法在医学图像数据挖掘中的应用[J].西北大学学报:自然科学版,2005,35(3):262-265.

[3] 罗可,林睦纲.数据挖掘中分类算法综述[J].计算机工程,2005,31(1):3-5.

[4] Gama J.Combining classification algorithms [D].Proto:Universidade do Porto,2000.

[5] Brunelli R,Falavigna D.Person identification using multiple cues[J].IEEE Trans Pattern Analysis and Machine Intelligence,1998,12(10):955-966.

[6] 杨余旺,杨静宇.多距离分类器组合试验在人脸识别中的应用[J].计算机工程,2005,31(2):50-53.

[7] 胡钟山,娄震,杨静宇.基于多分类器组合的手写体数字识别[J].计算机学报,1999,22(4):369-374.

[8] Dietterich T G,Kong E B.Machine learning Bias,statistical Bias,and statistical variance of decision tree algorithms[D].Dept Computer Sci,Oregon State Univ,Corvallis,OR,1995.

[9] 李丙春,耿国华,周明全,等.一个医学图像分类器的设计[J].计算机工程与应用,2004,40(17):230-232.

[10] 周志华,曹存根.神经网络及其应用[M].北京:清华大学出版社,2003.