一种基于贝叶斯和神经网络的医学图像组合分类方法*)

陈健美[1] 宋顺林[1] 朱玉全[1] 宋余庆[1] 陈 耿[2] 程 鹏[1] 桂长青[1]
(江苏大学计算机科学与通信工程学院 镇江 212013)[1] (南京审计学院 南京 210029)[2]

摘 要 医学图像分类是当前医学图像自动诊断和模式识别领域的一个新的研究热点,其任务是从给定的医学图像训练样本中提取能反映图像内容的特征,并根据这些特征进行图像分类,实现医学图像中病变组织的自动识别,以保证临床医学诊断更客观、准确和科学。通过对医学图像分类中的一些关键问题分析和研究,提出一种基于贝叶斯和神经网络的医学图像组合分类方法,并据此构造出医学图像组合分类器。这种组合分类器能够充分发挥各个分类器的优点,获得较好的图像分类结果。

关键词 图像分类,医学图像,特征提取,多分类器

A Method of Medical Images Combining Classification Based on Bayesian and Neural Network

CHEN Jian-Mei[1] SONG Shun-Lin[1] ZHU Yu-Quan[1] SONG Yu-Qing[1] CHEN Geng[2] CHEN Pen[1] GUI Chang-Qing[1]
(School of Computer Science & Communication Engineering, Jiangsu University, Zhenjiang 212013)[1]
(Nanjing Audit University, Nanjing 210029)[2]

Abstract Medical images classification is a new research hotspot for medical image diagnosis automatically and pattern recognition. Its main tasks: firstly, to extract features, which can describe image contents, from training sample image sets; secondly, to classify the test image sets according to those features; finally, to recognize pathological tissue automatically by the classification results and can ensure a more objective and accurate result of clinic medical diagnosis scientifically. In this paper, some key problems of medical image classification are discussed and analyzed and a medical image combining classification method is proposed on the basis of Bayesian and Neural Network. The experiments results show that our medical image combining classifier can make use of the advantages of each classifier, and it can obtain a good classification results.

Keywords Image classification, Medical image, Feature extraction, Combining classifier

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B超扫描图像、彩色多普勒超声图像、核磁共振(MRI)图像、CT图像、PET图像、SPECT图像、数字X光机(DX)图像、X射线透视图像、电子内窥镜图像、病理切片图像等数据。在这些医学影像数据库中,绝大部分医学影像所属类别(正常的还是异常的,如是异常的,还包括何器官何种异常)已被医生确诊,如何充分利用这些已确诊病例影像数据的信息和医生的临床诊断经验来诊断未确诊或临床病例医学影像所属类别,确保临床诊断更客观、准确和科学,已成为医学图像自动诊断中的一个关键技术。另外,在图像指导治疗技术中,医学图像后处理及其治疗方法、手术计划与导航、可视人计划(VHP)和医学虚拟现实及其相关技术等都需要进行大量的医学图像分类工作。因此,医学图像分类方法的研究不但具有较高的学术价值,而且具有广泛的应用前景,引起了国内外许多学者的极大关注,并已提出了一些有效的医学图像分类方法。如:Antonie等人提出的基于神经网络和关联规则的医学图像分类方法[1];Zhang等人提出的基于贝叶斯网络和小波变换的医学图像分类方法[2];朱玉全等人提出的基于频繁模式树的医学图像关联分类规则挖掘算法[3]。另外,还有一些其它方法,如基于粗糙集的方法[4]、基于马尔可夫模型的方法[5,6]等。对这些方法进行分析可以发现,这些方法总体上来说仍处于理论探讨阶段,离实际应用还有一定的距离。另外,算法本身也存在着一些缺陷,如人工神经网络的训练慢和容易陷入局部最小点问题、贝叶斯方法的无限样本集问题等。

综上所述,尽管在医学图像分类这一领域方面很多学者已做了大量的研究工作,并已取得了一些成绩,但目前仍没有一个广泛使用的可用于医学图像的分类器,还有许多问题有待于进一步的研究和完善。为此,本文对医学图像分类中的一些关键问题进行了分析和研究,提出一种基于贝叶斯和神经网络的医学图像组合分类方法,并据此构造出医学图像组合分类器。这种组合分类器能够充分发挥各个分类器的优点,获得较好的图像分类结果,从而进一步提高了医学图像分类的准确性和稳定性。

1 医学图像分类器的基本框架

虽然医学图像分类还没有形成可以广泛应用的图像分类器,但医学图像分类方法可以用图1表示其基本框架,主要包括:医学图像的预处理、医学图像的特征描述、图像分割与局部特征的提取以及分类器的构造,其中特征描述、局部特征的提取以及分类器的构造是医学图像分类中的难点,也是重点。

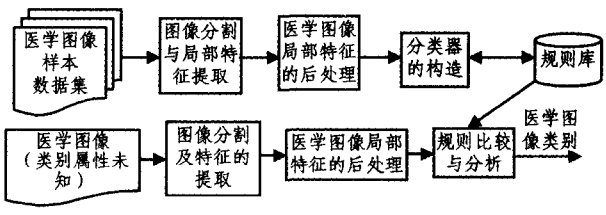


图1 医学图像分类器的基本框架

*)本文得到国家自然科学基金(60572112),江苏省软件与集成电路专项基金资助([2005]196)和江苏省教育厅自然科学基金(06KJB120051)资助。**陈健美** 副教授,博士研究生,主要从事数据挖掘和医学图像数据库研究。

2 医学图像特征提取与后处理

医学图像特征选择、提取和后处理是医学图像分类中的一个重要步骤，数据准备是否做好将直接影响到分类的效率、准确度以及最终模式的有效性。

2.1 医学图像分割

一般情况下，一幅医学图像包含许多丰富的组织，局部组织的病灶有时会与周围正常组织在灰度、形状上十分相似，整幅医学图像的特征可能无法反映出某一局部组织器官的微小病变，仅取图像的全局特征显然是不准确的，因此，实现医学图像中人体组织器官的分割将有着十分重要的现实意义。

医学图像分割就是根据某种均匀性原则，使用图像分割方法(如：阈值分割技术、微分算子边缘检测、区域增长技术、聚类分割技术等)，将一幅医学图像分成若干个有意义的部分，使得每一部分都符合某种一致性的要求，而任意两个相邻部分的合并都会破坏这种不一致性。简单地讲，通过对医学图像进行分割，可以识别出一幅医学图像中的局部组织或临床诊断医生所关心的感兴趣区域。

2.2 医学图像特征选择和提取

一幅医学图像一般有三个层次的特征：元数据、文字注释、内容特征，其中元数据特征和文字注释特征是医学图像的外在特征，与医学图像分类无关，因此，本文仅考虑医学图像的纹理、颜色、形状和空间等内容特征。

特征提取是否合适是医学图像分类结果准确与否的一个关键因素，表达人体解剖器官的局部医学图像特征比全局图像特征更加重要。目前绝大多数的特征提取都是建立在整幅医学图像基础之上，或者简单地将图像分成规则的几部分，分别提取每部分的特征，这些特征不能真正表示图像中各组织器官的真实信息。

针对每一个分割出来的局部组织或临床诊断医生所关心的感兴趣区域，分别提取其相应的内容特征，具体包括灰度直方图的均值、方差、倾斜度、能量、峰态、熵等特征，灰度共生矩阵的角二阶矩、对比度、逆差分矩、熵、惯性矩等特征；基于小波系数的均值、方差、倾斜度、能量等特征；医学图像的聚类特征[7]等等。

2.3 医学图像局部特征的后处理

不同特征(如均值、方差等)之间在数值上存在很大的悬殊，为了避免这种悬殊对分类结果的影响，需要对所提局部特征进行后处理。后处理包括特征的归一化、特征值约简和离散化等。

2.3.1 特征的归一化

为了消除量纲的影响，将数据经过标准差变换后，使每个变量的均值为0，标准差为1，但其值不一定保证在[0,1]区间内。经过极差变换后，每个变量的取值范围为[0,1]，且消除了量纲的影响。式(1)和(2)分别给出了标准差和极差的变换公式。

$$x''_{ik}=\frac{x'_{ik}-\overline{x'_k}}{S_k} \tag{1}$$

其中 $S_k^2=\frac{1}{n}\sum_{i=1}^{n}(x'_{ik}-\overline{x'_k})^2$，

$$\overline{x'_k}=\frac{1}{n}\sum_{i=1}^{n}x'_{ik}$$

$$x_{ik}=\frac{x''_{ik}-x''_{k\min}}{x''_{k\max}-x''_{k\min}} \tag{2}$$

其中，$x''_{k\min}$，$x''_{k\max}$分别表示最小和最大值。

2.3.2 特征约简和离散化

从医学图像中提取的上述特征，并不一定同等重要，甚至某些特征是冗余的，这样需要进行特征约简，在保持医学图像分类效果变化不大的情况下，删除那些不相关或不重要的特征，下面采用粗糙集的方法进行特征约简。

由于分类所需的训练样本数据是离散数据，因此要求特征数据表中的值用离散数据表达，即需进行离散化处理。本文将使用 Semi Naive Scaler 算法对特征数据进行处理，取得候选断点集，然后采用基于属性重要性的离散化算法处理候选断点集，算法如下。

第1步：根据特征的重要性由小到大对所有特征 $V_i(i=1,2,\cdots,n)$进行排序，在特征重要性相同的情况下，按特征断点个数由多到少对特征进行排序；

第2步：对每个特征 $V_i\in A$ 执行：

对特征 V_i 中的每一个断点 $C_j(j=1,2,\cdots,i_j)$，考虑它的存在性：把信息系统中与 C_j 相邻的两个属性值的较小值改为较大值，如果信息系统不引入冲突，则 $C_{Vi}=C_{Vi}\backslash\{C_j\}$；否则，把修改过的特征值还原。

在对原始数据进行离散化的基础上，粗糙集理论提取出最能反映分类本质的特征，即在保证不丢失知识库中有效信息的前提下，消除知识库中的冗余分类或冗余基本范畴，这一过程消去知识库中所有不必要的知识，保留获取正确决策的最小信息量，从而也降低了特征空间的维数[7]。

定义1 四元组 $s=(U,A,V,f)$是一个决策表，其中 U 表示对象的非空有限集合，称为论域；$A=C\cup D$，$C\cap D=\Phi$，C 称为条件属性集，D 称为决策属性集；$V=\bigcup_{a\in A}V_a$，V_a 是属性 a 的值域；f 表示 $U\times A\to V$，它是一个信息函数，即：$f(x,a)\in V_a$，$a\in A$，$x\in U$，每个属性子集 $B\subseteq A$ 决定了一个二元不可区分关系 $IND(B)$，即：$IND(B)=\{(x,y)\in U\times U\mid \forall a\in B, f(x,a)=f(y,a)\}$，易知，$IND(B)$为一等价关系。

定义2 子集 X 的下近似集$\underline{B(x)}$和上近似集$\overline{B(x)}$分别定义如下：$\underline{B(x)}=\{x_i\in U\mid B(x_i)\in X\}$；$\overline{B(x)}=\{x_i\in U\mid B(x_i)\cap X\neq\Phi\}$。如果$\overline{B(X)}-\underline{B(X)}=\Phi$，则集合 X 为 B 上的可定义集合，否则称 X 为 B 上的粗糙集；X 的 B 正域是所有根据知识 B 能确定的划入集合 X 的 U 中对象的集合，即：$POS_B(X)=\underline{B(x)}$。

定义3 记 $A=\{a_1, a_2,\cdots a_m\}$为有限元素集合，准则函数 $C(a)$定义为 A 中元素的排序函数，$\forall a\in A$，有 $C(a)>0$。按 $C(a)$大小对 A 中元素进行排序得序集 $OA=\{a'_1, a'_2,\cdots a'_m\}$，其中 $C(a'_1)\geqslant C(a'_2)\geqslant\cdots\geqslant C(a'_m)$，定义 OA 中元素权值为 $C(a'_i)=2^{m-i}$，$1\leqslant i\leqslant m$。

定义4 令 $T(OA)$为集合 OA 的幂子集，$T_1(OA)$为集合 OA 的一阶幂集，给 $T_1(OA)$中元素赋以权值，有 $\forall A'\in T_1(OA)$，$\omega(A')=\omega(a'_i)$，$a'_i\in A'$。按 $\omega(A')$大小对 $T_1(OA)$中的元素进行排序，得到一阶有序幂子集 $OT_1(OA)$，同理，可得到 i 阶有序幂子集 $OT_i(OA)$。

归纳特征约简算法为：

第1步 求取 P 的 D 核 $\mathrm{CORE}_D(P)$；

第2步 求取 P 的 D 最小属性约简 $\mathrm{mred}_D(P)$：

(1) 令 $X=\mathrm{CORE}_D(P)$，$L=P\backslash X=\{a_1, a_2,\cdots a_m\}$，$T(L)$表示 L 的幂集，$T_i(L)$为 L 的 $i(1\leqslant i\leqslant m)$阶幂子集；

(2) 如果 $POS_X(D)=POS_P(D)$，则 $\mathrm{mred}_D(P)=X$，转(10)；

(3) $i=1$，flag=0，Z,A,X；

(4) $Y=T_i(L)$；

(5) 任取 $y\in Y$，$A=X\cup\{y\}$，

如果 $POS_A(D)=POS_P(D)$，则

如果 flag=0，则 $Z=A$，flag=1；

否则，如果 card($U|Z$)>card($U|A$)，则 $Z=A$；

(6) $Y=Y-\{y\}$；

(7) 如果 $Y\neq\Phi$，转(5)；

(8) 如果 flag=1，则 $\text{mred}_D(P)=Z$，转(10)；

(9) $i=i+1$，如果 $i\leqslant m$，则转(4)；

(10) 结束。

3 医学图像分类算法

目前，可用的医学图像分类方法有很多，每一种方法均有其优点，也存在一些算法本身无法克服的缺陷，如基于决策树算法的分类器虽然具有速度快、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出不同的结果；朴素贝叶斯分类要求属性值之间是相互独立的，在很多情况下这是很难得到满足的，其准确度难以有较大幅度的突破。为了充分发挥各种分类器的优势，提出了一种基于组合分类策略的医学图像分类器，该分类器通过某种组合技术，将多个单分类器的预测进行组合而产生一个新分类器。

综上所述，我们可以得到医学图像的分类算法，算法描述见算法 1。

算法 1 医学图像分类算法 MICCA(Medical Images Combining Classification Algorithm)。

输入：训练样本医学图像集 L 和测试医学图像集 T。

输出：分类规则或分类函数。

方法：

(1) 提取各训练样本医学图像的内容特征；//具体的内容特征见 2.2 节；

(2) 特征归一化处理，包括标准差变换和极差转换；//参见公式(1)和(2)；

(3) 特征约简与离散化，采用粗糙集方法对特征进行约简；//具体内容见 2.3 节；

(4) 利用贝叶斯算法对 L 进行学习，并对 L 和 T 进行测试，得出每个训练样本的类别(假设类别有：正常、异常 1、异常 2、…、异常 k，$k\geqslant 1$)概率估计(p_0，p_1，p_2，…，p_k)，并比较得出基于贝叶斯构造的单分类器决策结果；

(5) 将 L 和 T 中训练样本的初始特征向量(x_1，x_2，…，x_m)与对应的由贝叶斯分类器得出的概率估计向量(p_0，p_1，p_2，…，p_k)合并，得到第二层的训练样本数据 Level_2train 和测试数据 Level_2test，其中：

$\text{Level}_2\text{train}=L\oplus\varphi(\text{NaiveBayes}(L),L)$。

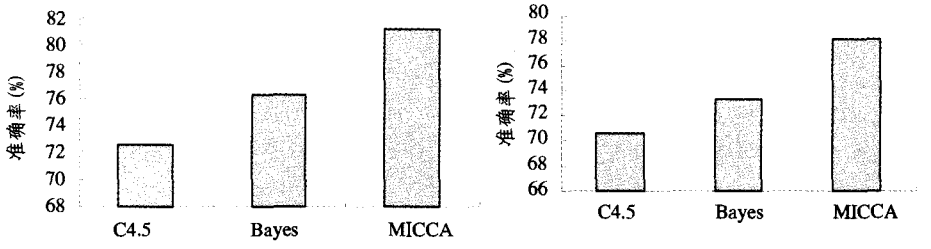
$\text{Level}_2\text{test}=T\oplus\varphi(\text{NaiveBayes}(L),T)$。

$\varphi(\text{NaiveBayes}(L),L)$、$\varphi(\text{NaiveBayes}(L),T)$分别表示用分类算法 NaiveBayes 对训练集 L 进行学习得到的分类器 NaiveBayes(L)对 L 和 T 进行分类。

(6) 利用 BP 对 Level_2train 进行学习，得到基于贝叶斯和神经网络的医学图像组合分类器 BP_Bayes，对 Level_2train 和 Level_2test 进行测试，得出该组合分类器有关训练集和测试集的决策结果，作为分类规则或分类函数的最终输出结果。

4 算法实现与分析

为了验证算法 MICCA 的有效性和可扩展性，我们用 Vc++6.0 在内存为 512M、CPU 为 PentiumIV- 1.8MHz、操作系统为 Window XP 的机上实现了算法 MICA 和单分类器 Bayes 和 C4.5 算法，算法所用数据为大小 512×512 像素的肝脏 CT 二维图像数据。在实验中，我们随机地从 10000 幅肝脏图像中选出两组训练样本集，第一组有 400 幅医学图像，其中正常和异常各占 200 幅；第二组有 2000 幅医学图像，其中正常和异常各占 1000 幅。在余下的肝脏图像中选取 200 幅医学图像作为分类器的测试集。另外，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 或最大训练次数为 1000 时则终止。实验结果如图 2 所示，图 2 表明算法 MICCA 可以有效地提高医学图像分类的准确性。



(a) 第一组训练样本　　(b) 第二组训练样本

图 2 算法分类准确率比较

结束语 计算机图像处理技术特别是图像的模式识别与模式分类技术在近三十多年中得到了飞速的发展，并已在许多领域得到了非常广泛的应用。然而目前还没有一个很好的适合于医学诊断应用领域的图像分类技术。因此，如何构建一个高效率的医学图像分类识别系统一直是计算机学者和医学专家们探讨的热门课题。

本文对医学图像分类中的一些关键问题进行了分析和研究，提出一种基于贝叶斯和神经网络的医学图像组合分类方法，并据此构造出医学图像组合分类器。这种组合分类器能够充分发挥贝叶斯和神经网络的优点，获得较好的图像分类结果，从而进一步提高了医学图像分类的准确性和稳定性。

参 考 文 献

1 Antonie M L, Zaiane O R, Coman A. Application of data mining techniques for medical image classification [C]. In: Proc. of Second Int'l Workshop on Multimedia Data Mining in Conjunction with Seventh ACM SIGKDD, San Francisco, USA, 2001. 94～101

2 Zhang X P, Desai M D. Wavelet Based Automatic Thresholding for Image Segmentation [C]. In: Proc. of the ICIP' 97 conference, Santa Barbara, CA, 1997. 26～29

3 朱玉全，宋余庆，杨鹤标，等. 基于频繁模式树的关联分类规则挖掘算法[J]. 江苏大学学报(自然科学版)，2006，27(3)：262～265

4 Brazokovic D, Neskovic M. Mammogram Screening Using Multiresolution based Image Segmentation [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2001, 7(6): 1437～1460

5 Li H. Markov Random Field for Tumor Detection in Digital Mammography [J]. IEEE Trans Medical Imaging, 2000, 14(3): 565～576

6 Bottigli U, Golosio B. Feature Extraction from Mammographic Images Using Fast Matching Methods [J]. Nuclear Instruments and Methods in Physics Research, 2002, A 487: 209～215

7 宋余庆，谢从华，朱玉全，等. 基于近似密度函数的医学图像聚类分析研究[J]. 计算机研究与发展，2006，43(11)：1947