基于粗糙神经网络的医学图像分类新方法*)

蒋 芸[1,2] 李战怀[1] 王 勇[1] 张龙波[1]

(西北工业大学计算机学院 西安 710072)[1] (西北师范大学数学与信息学院计算机系 兰州 730070)[2]

摘 要 由于乳腺X光图像的复杂性,直接从图像中看出肿瘤及其良、恶性质是比较困难的,因此建立高效的肿瘤自动诊断系统是非常必要的。文章将粗糙集理论中基于信息增益的约简方法和神经网络相结合,提出了粗糙神经网络算法RNN,将其应用于乳腺X光图像分类。实验结果表明,该方法的分类精确度可达到92.37%比单独使用神经网络方法的分类精确度(81.25%)要高,同时所花费的时间也明显减少。

关键词 粗糙神经网络,粗糙集理论,乳腺X光图像

A New Medical Image Classify Approach Based on Rough Neural Network

JIANG Yun[1,2] LI Zhan-Huai[1] WANG Yong[1] ZHANG Long-Bo[1]

(College of Computer Science, Northwest Polytechnical University, Xi'an 710072)[1]

(College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070)[2]

Abstract Detecting tumor in mammography is a difficult task because of complexity in the image. This brings the necessity of creating automatic tools to find whether a mammography present tumor or not. In this paper we join neural network with information gain reduction of rough sets theory which we call the rough neural network(RNN) to classify digital mammography. The experimental results show that the rough neural network performs better than only neural network algorithm in terms of time though it can get 92.37% classify accuracy which is higher than 81.25% using neural network only.

Keywords Rough neural network, Rough sets theory, Mammography

1 引言

早期通过对乳腺X光图像的检查是发现并预防乳腺癌的最好方法之一。由于乳腺X光图像的低对比度、肿瘤组织的不同等原因,直接从图像中看出肿瘤并诊断其良、恶性质的精确度是比较困难的,因此通过计算机来辅助诊断是十分必要的[1]。文[1]用神经网络和关联规则数据挖掘方法对乳腺X光图像进行分类,他们用神经网络方法能达到81.25%的分类精确度。虽然神经网络所需要的数据训练时间要比关联规则等方法多[1],但关联规则没有神经网络的分类精确度高,所以研究高效的、数据训练时间短的神经网络分类方法是非常必要的。近年来,在乳腺X光图像分类方面所做的工作主要有:文[2]利用贝叶斯网络来寻找和分类人类专家感兴趣的区域,文章中应用了基于小波变换的分段算法,并使用域值来对应图像中感兴趣区域的最小值。文[3]中,作者依据边缘分段算法提出了一种对乳腺X光图像中病灶部位提取特征的方法。文[4]对从乳腺X光图像中感兴趣的区域内提取纹理特征的不同方法进行了评价。在文[5]中,我们能够看到作者是如何应用小波变换在乳腺X光图像中探查微钙化组织的。另外,还有一些其它的方法,如基于粗糙集的方法[6]、基于马尔可夫模型的方法[7]等。虽然已经在乳腺X光图像的分类方面做了这么多工作,但多数是基于图像处理的方法,基于数据挖掘的方法比较少。同时还没有将这些方法广泛应用于医学领域,主要原因是由于该领域需要较高的分类精确度,而且对数据的训练不能花费太多的时间。

在文章中,我们将粗糙集理论中基于信息增益的约简原理与神经网络方法相结合,减少了特征属性从而降低了数据训练时间,同时也避免了单独使用粗糙集分类的过度约简问题。将该方法应用于乳腺X光图像,在标准数据集MIAS[8](the Mammographic Image Analysis Society)上做实验并获得了92.37%的分类精确度,同时也大大降低了训练所需要的时间。

2 图像预处理和特征提取

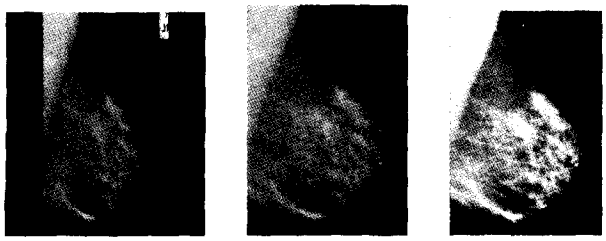
我们使用的图像预处理和特征提取方法与文[1]中的方法相类似,目的是为了比较后继挖掘算法的分类精确度以及训练时间。

2.1 图像的预处理

MIAS数据集中图像的典型尺寸是1024×1024,由于这些图像是在不同外部条件下获取的,因此一些图像的亮度很高而另一些图像却太暗,其中50%的图像在背景中含有大量噪声。去除噪声首先是用剪切操作来修剪图像;然后是图像增强。我们去除了几乎所有的背景信息和大多数噪声。图1(a)是MIAS中的一幅原图,(b)就是经过剪切和去除噪声以及背景信息后的图像。由于图像的大小不同,因此在做剪切操作时横纵坐标的x和y的取值范围规定为(0,255),我们用

*)国家自然科学基金资助项目(60373108);国家自然科学基金资助项目(60573096);甘肃省自然科学基金资助项目(3ZS051-A25-042)。蒋 芸 副教授,博士研究生,主要研究方向:数据挖掘技术、粗糙集理论及应用;李战怀 教授,博士生导师,主要研究方向:数据库理论与技术;王 勇 讲师,博士研究生,主要研究方向:数据流挖掘技术;张龙波 博士研究生,主要研究方向:数据流管理、数据流挖掘技术。

垂直剪切的方法去除了多余的部分；然后用直方图均衡法增强图像避免图像过亮或过暗影响分类的效果，图 1(c)就是经过增强后的图像效果。



(a) (b) (c)

图 1 (a)原始图像；(b)去除噪声后的图像；(c)增强后的图像

2.2 图像特征提取

预处理之后，我们将提取的特征数据放入数据库中，并加入一些 MIAS 数据集中已经存在的有关图像的信息，构成用于做数据分类的特征库。文章提取的特征是 4 个统计参数：均值(mean)、方差(variance)、偏斜度(skewness)和峰度(kurtosis)，这 4 个参数的计算公式分别如下[11]：

$$Mean:\mu=\sum_{k=1}^{N} f_k p_f(f_k) \tag{1}$$

$$Variance:\sigma^2=\sum_{k=1}^{N}(f_k-\mu)^2 p_f(f_k) \tag{2}$$

$$Skewness:\mu_3=\frac{1}{\sigma^3}\sum_{k=1}^{N}(f_k-\mu)^3 p_f(f_k) \tag{3}$$

$$Kurtosis:\mu_4=\frac{1}{\sigma^4}\sum_{k=1}^{N}(f_k-\mu)^4 p_f(f_k) \tag{4}$$

首先将图 1(c)中的图像均分成 4 块，再将其中的每一块均分成 4 块，最终将该图像均分成 16 块，在每一块中分别提取 4 个统计参数，我们获得了 64 个统计特征。

3 粗糙神经网络方法(RNN)

文章首先应用粗糙集理论中基于信息增益的属性约简原理，发现属性间的依赖关系，约简原属性集，然后将约简得到的数据集作为后向传播神经网络的输入，从而获得更高效的分类结果。以下是粗糙集理论和后向传播神经网络的基本概念，以及粗糙神经网络算法描述。

3.1 粗糙集理论[9]

对决策系统 $S=(U,A,V,f)$，$\forall B\subseteq A$ 是条件属性集合的一个子集，称二元关系 Ind(B)为 S 的不可区分关系：$Ind(B)=\{(x,y)\in U\times U \mid \forall a\in B, f(x,a)=f(y,a)\}$，它表示对象 x 和 y 关于属性集 A 的子集 B 是不可区分的。给定 $X\subseteq U$，$B(x_i)$是按等价关系 Ind(B)得到的包含 x_i 的等价类。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：

$$\underline{B(X)}=\{x_i\in U \mid B(x_i)\subseteq X\}$$

$$\overline{B(X)}=\{x_i\in U \mid B(x_i)\cap X\neq\phi\}$$

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 能确定地划入集合 X 的 U 中对象的集合，即：

$$POS_B(X)=\underline{B(X)}$$

约简是粗糙集中的一个基本概念，即去除决策系统中冗余的信息。设 A 是属性集，如果 $Ind(A-c_I)=Ind(A)$，那么 c_i 就是冗余的属性。$\forall A_i\in A$，A_i 的信息增益为：$Gain(A_i)=1-E(A_i)/E(S)$，其中 $E(S)=-\sum_{i=1}^{n} P_i \log_n P_i$，$E(A_i)=\sum_{i=1}^{m} W_i * E(S_i)$，$n$ 是 S 中的对象总数，S_i 是属于第 i 个类别的对象数，$S_i\subseteq S$，$P_i=|S_i|/|S|$，$W_i=S_i$ 的样本数/S 的样本数。

3.2 后向传播神经网络[10]

后向传播是一种神经网络学习算法，在文[1]中使用的就是该算法，后向传播神经网络算法是在多层前馈神经网络上学习。这种神经网络分为输入层、隐藏层和输出层，输入对应于对每个训练样本度量的属性，输入同时提供给称作输入层的单元层。这些单元的加权输出依次同时地提供给称作隐藏层的“类神经元”的第二层，该隐藏层的加权输出可以输入到另一个隐藏层，隐藏层的数量是任意的，最后一个隐藏层的加权输出作为构成输出层的单元的输入，输出层发布给定样本的网络预测。这种算法的优点包括其对噪声数据的高承受能力，以及它对未经训练的数据分类模式的能力；但神经网络需要很长的数据训练时间，因此用新方法缩短神经网络的训练时间是非常必要的。

后向传播通过迭代地处理一组训练样本，将每个样本的网络预测与实际知道的类标号比较，进行学习。对于每个训练样本，修改权，使得网络预测和实际类之间的均方误差最小。这种修改“后向”进行，即由输出层，经过每个隐藏层，到第一个隐藏层。一般情况下权将最终收敛，学习过程停止。具体算法见文[10]。

3.3 粗糙神经网络算法(RNN)

RNN 算法分为两部分，第一部分是应用粗糙集的信息增益方法进行属性约简；第二部分是后向传播神经网络。假设算法执行前已将连续的属性值离散化。RNN 中的 RJ 算法来自文[11]，是一种属性判断约简方法。

RNN 算法

输入：S 是决策表。其中属性集 A=C∪D，C 是条件属性集，D 是决策属性集。L 是学习率，Net 是多层前馈网络。

输出：对样本分类的神经网络。

1)在 S 上执行 RJ 算法来提取部分属性的约简，然后将它们放入约简池。

2)计算所有属性的信息增益。

3)初始化约简表 SR，设 $R_m=0$；//SR 就是最终
　　//对 S 经过约简的决策表

4)While not(约简池=empty)do

5)从约简池中选择一个约简 RD；

6)R_i=information_gain(RD)；// 计算
　　//约简 RD 的信息增益，放入 R_i 中；

7)If $R_i>R_m$ then

8)SR←R_i；//将 R_i 作为 S_i 的一个成员

9)$R_m=R_i$；

10)Endif

11)endwhile

12)back_propagation(SR,L,Net)
　　//后向传播算法，SR 就是约简后的训练样本集

4 实验结果

4.1 数据集说明

文章用于实验的数据集来自于 MIAS，是研究乳腺 X 光图像的标准数据集，文[1]中使用的也是该数据集。在 MIAS 数据集中包含 322 幅乳腺 X 光图像，所有图像都是乳腺侧面图，它们分属于三类：正常、良性和恶性，后两类又统称为非正常。其中属正常的图像 208 幅，非正常 114 幅(良性 63 幅，恶

性51幅)。所有非正常的图像都包含出现异常的位置等信息,例如肿瘤的圆区域、它的半径、乳房位置(左、右)、乳房组织的类型(密度、多脂的、多脂含腺的)以及是否存在肿瘤等。

4.2 实验结果及分析

文章用10层交叉的方法在特征库上做分类测试,将特征库随机分成10份,选择其中90%做训练,其余10%做测试,记录其分类精确度;同时还记录了在训练集上10次执行RNN算法和后向传播神经网络算法所需要的平均时间,其中神经网络中的学习率设为0.01。特征库是由从MIAS的每幅图中抽取的64个统计参数和已存在的一些数据组成,共69个属性,所有连续值属性都用算法DBCHi2[12]进行了离散化处理。

表1是我们的实验结果,第一列是特征库10次划分的说明;第二、三列是在特征库上执行10次后向传播神经网络算法平均所需时间和平均分类精确度,其中第二列的实验结果来自文[1];第四、五列是10次执行RNN算法平均所需时间和平均分类精确度。实验结果表明,RNN的平均分类精确度达到了92.37%,比文[1]中单独使用后向传播神经网络算法的平均分类精确度81.25%高;同时训练所需要的时间也明显减少。在实验中我们还计算了粗糙神经网络的灵敏性(sensitivity)和特效性(specificity)[10],这两个参数的计算公式分别是式(5)和式(6):

表1 RNN算法与后向传播神经网络算法在数据集MIAS上的实验结果比较

	后向传播神经网络算法		RNN算法	
	花费时间(hrs)	分类精确度(%)	花费时间(hrs)	分类精确度(%)
10次划分所得的平均值	8.72	81.25	1.05	92.37

$$Sensitivity=t\_pos/pos \quad (5)$$

$$Specificity=t\_neg/neg \quad (6)$$

其中t_pos表示真正样本数(被正确分类的正样本数),pos表示正样本数,t_neg表示真负样本数(被正确分类的负样本数),neg表示负样本数。从公式(5)(6)可以看出,sensitivity和specificity的值越接近100%,说明样本被正确分类率就越高。从图2中可以看到,我们实验中的灵敏性和特效性都接近100%,也就是表明,对正常和非正常乳腺X光图像的正确分类率都比较高;同时特效性的值更接近100%,这说明将非正常乳腺X光图像错误分类的可能性很小,这正是医学专家所期望的。

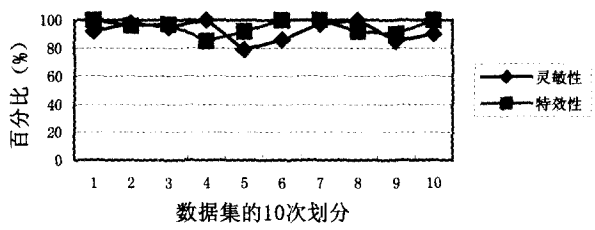


图2 MIAS被均分为10份的灵敏性和特效性度量

结论 文章基于后向传播神经网络算法和粗糙集中信息增益约简理论提出了一种粗糙神经网络算法RNN,并将其应用于乳腺X光图像数据集MIAS,实验结果表明,应用该方法对MIAS中的322幅图像做分类,能够得到92.37%的平均分类精确度,比文[1]中单独使用后向传播神经网络算法的平均分类精确度81.25%提高了近12%,并且训练数据所需要的时间也明显减少。将数据集随机均分成10份,超过一半的灵敏性值和特效性值接近100%,说明在对乳腺X光图像分类时,将非正常图像划分到正常范围的可能性很小,这是患者和医学专家期望的结果。在医学图像的自动化分类处理方面还有待于做进一步的研究,如与医学专家的合作能够得到更多有意义的结果;用不同的方法提取不同的特征、应用新的分类方法以及不同的特征数据库组织方式等能够使我们获得更好的结果。

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 Intl Workshop on Multimedia Data Mining (MDM/KDD'2001)in Conjunction with Seventh ACM SIGKDD, San Francisco, USA, 2001. 94～101

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

3 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

4 Sharma M, Singh S. Evaluation of Texture Methods for Image Analysis [C]. In: Proceedings of the 7th Australian and New Zeland Intelligent Information Systems Conference. Perth, November, 2001. 18～21

5 Yoshida H, Doi K, Nishikawa R, et al. Application of the Wavelet Transform to Automated Detection of Clustered Microcalcifications in Digital Mammograms [R]. In: Academic Reports of Tokyo Institute of Polytechnics, 1994. 24～37

6 Brazokovic D, Neskovic M. Mammogram Screening Using Multiresolution-based Image Segmentation [J]. International Journal of Pattern Recognition and Artificial Intelligence, 1993, 7(6): 1437～1460

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

8 http://www.wiau.man.ac.uk/services/MIAS/MIASweb.html [DB/OL]

9 Pawlak Z W. Rough sets and intelligent data analysis [J]. Information sciences, 2002. 1～12

10 Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 北京:高等教育出版社, 2001

11 Pan Dan, Zheng Qi-Lun, Zeng An, et al. A Novel Self-Optimizing Approach for Knowledge Acquisition[J]. IEEE Trans on Systems, Man and Cybernetics-Part A: Systems and Humans, 2002 (32): 505～514

12 Hu X, Cercone N. Data Mining Via Generalization, Discretization and Rough Set Feature Selection [J]. Knowledge and Information System: An International Journal, 1999, 1(1)