用机器学习方法解码脑图像数据

陈俊杰,赵 丽,相 洁
CHEN Junjie, ZHAO Li, XIANG Jie
太原理工大学 计算机与软件学院,太原 030024
College of Computer and Software, Taiyuan University of Technology, Taiyuan 030024, China

CHEN Junjie, ZHAO Li, XIANG Jie. Decoding brain image data using machine learning. Computer Engineering and Applications, 2012, 48(10):222-225.

Abstract: Feature selection and classification is the core issue for high-dimensional functional Magnetic Resonance Imaing(fMRI) data analysis. Feature selection is divided into two steps. Region of interesting is selected in brain in the first place, and the voxels of the most distinguishing stimulating task are selected among them. The method is simple, stable and consistent with human logic. Gaussian Naive Bayes(GNB) and Support Vector Machine(SVM) classifier is used to evaluate the method. The experimental results show that the method is feasible, and to compare classification methods, SVM is superior overall to GNB.

Key words: Gaussian Naive Bayes; Support Vector Machine; functional Magnetic Resonance Imaging(fMRI); feature selection

摘 要:特征选择和分类是脑功能磁共振成像(fMRI)数据分析的核心问题。针对fMRI高维数据,特征选择分两步,选取感兴趣脑区,选择最能区分刺激任务体素。该方法简单,稳定,符合人的思维逻辑。分类器选择高斯朴素贝叶斯(GNB)和支持向量机(SVM),评估该特征选择方法。实验结果表明,该方法有效提高了分类速度,分类准确度也得到很大提高。对分类方法进行比较,SVM总体上优于GNB。

关键词:高斯朴素贝叶斯;支持向量机;功能磁共振成像;特征选择
文章编号:1002-8331(2012)10-0222-04 **文献标识码:**A **中图分类号:**TP181

1 引言

功能磁共振成像(fMRI)的出现,推动了脑科学的巨大发展。它的特点是无创伤性、高空间分辨率和重复看到活体脑部工作神经活动。目前,fMRI已经成为研究大脑模式与神经激活之间关系的强大工具[1]。血氧依赖水平功能磁共振(BLOD fMRI)作为最常用的一种技术,当大脑受到任务刺激,会激发大脑神经系统的活动,造成血液中血氧含量发生变化,BLOD fMRI间接测量这种变化,并产生大量的数据。研究者们通过这些数据来分析大脑的模式。

fMRI数据分析方法有很多,根据研究目的不同可以分为两类:一类是脑激活;另一类是数据分类。前者是观察刺激任务下被激活的脑区,即脑功能定位,主要使用统计检验方法(一般线性模型GLM)[2]和基于数据信息挖掘方法(聚类分析clustering等)[3]。后者与前者问题正好相反,主要根据脑部产生的不同激活模式分析大脑所受到刺激任务,即大脑当前的认知状态。当前的研究主要是针对瞬时认知状态[4]。

机器学习和模式识别方法已经广泛用于fMRI数据分类,探索大脑思维与认知状态,例如简单种类:脸、工具和房子[5],情感[6],甚至说谎[7]等。机器学习分析fMRI数据主要涉及到两种理论,特征选择与分类。特征选择就是选择那些能较高分类fMRI实验任务类别(工具、房子与脸)的体素。fMRI分析的特征

基金项目:国家自然科学基金(No.60970059);山西省自然科学基金(No.20100110202)。
作者简介:陈俊杰(1956—),男,教授,博士生导师,主要研究领域为数据挖掘、人工智能及其应用;赵丽(1982—),女,硕士研究生;相洁(1970—),女,硕士研究生。E-mail:zhaoli.computer@qq.com
收稿日期:2010-10-19 **修回日期:**2011-01-18 CNKI出版日期:2011-05-18
DOI:10.3778/j.issn.1002-8331.2012.10.050 http://www.cnki.net/kcms/detail/11.2127.TP.20110518.1116.007.html

选择比较困难,主要是因为特征与分类对象之间比例太大,体素之间具有相关性[8]。fMRI分类的主要工作是找到大脑对某一精神状态的激活模式,因此标识大脑状态首先要考虑分类的准确率[9]。典型研究模式是以体素作为特征,大脑对刺激的反应作为事件(分类的对象)。

针对fMRI高维数据问题,本文提出结构像与功能像,单体素与多体素模式分析相结合方法。首先选用感兴趣区域脑区体素作为特征,然后采用单体素(特征)分类器选择特征,并对体素进行分级,设定不同集合大小体素,选择高斯朴素贝叶斯(Gaussian Naive Bayes,GNB)和支持向量机(Support Vector Machine,SVM)分类算法,检验特征选择是否可行。最后对这两种分类算法进行性能比较。

2　实验与数据

实验采用事件相关性设计,被试在实验过程中完成一个sudoku变形游戏,棋盘由4个2×2宫格组成,被试在"?"位置添写数字,使行、列、宫格中不存在重复的数字。刺激任务分为4类,根据解题步骤和复杂度分为一步简单、一步复杂、二步简单、二步复杂。其任务的形式如图1所示。

1		2	3
		4	?
3			

图1(a)　一步简单

2			
		2	4
1			*
		3	?

图1(b)　二步复杂

简单任务只需考虑宫格、列、行其中的一个因素就可以得到答案,复杂任务则需要综合考虑三个因素。二步任务需先求出"*"位置的解,然后再求"?"。本实验共19人参与实验,去除4个头动较大的被试,剩余15个被试参与数据分析。

本文只做二分类,比较一步简单和二步复杂的刺激任务。因为这两种刺激任务相差最大,相应脑激活模式区别最大。在实验过程中,被试完成一个实验需要18秒的时间,在第3秒出现刺激任务,被试知道答案后,点击按钮进入休息,取第8秒到17秒时间间隔作为fMRI数据,因为感兴趣神经激活体素fMRI BLOD信号一般持续8~12秒,这样选取时间间隔内BLOD峰值达到最高,各个感兴趣BLOD模式区别最大[10]。

实验只研究单个被试分类器。每个分类器由60个左右的样本组成,每个类大约30个左右样本。在功能像中,每个3D脑图像由64×64×32=131 072个体素组成,在特征选择之前,每个样本由131 072×5个特征组成(fMRI图像每2秒种收集一次,fMRI数据由5个图像组成)。

3　方法

3.1　数据预处理

采用对脑图像的每个体素值进行标准化处理,使其每个脑图像中所有体素的平均值为0,标准差为1,这种标准化处理对所有脑图像都是独立进行的。通过标准化减少缓慢时间漂移所带来的影响及其他噪音,以便提高贝叶斯及其他分类器的性能。标准化公式如下所示:

$$Y_i=\frac{X_i-\mu_i}{\sigma_i} \tag{1}$$

X_i,Y_i,是标准化之前和之后的数据,表示第i个3D图像的体素值,μ_i表示第i个3D图像的所有选择体素的平均值,σ_i表示第i个3D图像所有选择体素的标准差。

3.2　学习方法

高斯朴素贝叶斯(GNB)。它利用训练集估计每个样本对应的类标签(被试认知状态)所发生的概率。在脑图像数据中,每一个体素值由很多神经元组成,由于磁共振磁场的不稳定性及头动等因素使其带有噪声,根据经验,假设每个体素服从高斯分布,这样每一个样本服从多变量高斯分布。朴素贝叶斯要求各个属性之间独立,但是在实际应用中,在属性不独立的情况下也能取得很好的效果。文献[11]经过多个实验得出结论,朴素贝叶斯分类器的表现与属性的独立性没有必然的联系。因此,在分析脑数据时,为了简化数据计算的复杂度,假设体素之间是独立的。

样本$X=(X_1,X_2,\cdots,X_n)$,计算样本X,认知状态C_i发生的概率$P(C_i|X)$,$P(C_i|X)$由贝叶斯规则计算,如下:

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}$$
$$P(C_i)=\frac{S_i}{S},\ X=(X_1,X_2,\cdots,X_n) \tag{2}$$

S_i为C_i的个数,S为训练样本的总数。

由于$P(X)$对于所有的类都为常数,只需求$P(C_i|X)P(C_i)$最大即可:

$$P(C_i|X)=\Pi_j P(x_j|C_i)$$
$$P(x_j|C_i)\sim N(u,\sigma) \tag{3}$$

$$f(x|\mu_{ci},\sigma_{ci}^2)=\frac{1}{\sqrt{2\pi\sigma_{ci}^2}}\exp[-\frac{1}{2}(\frac{x-\mu_{ci}}{\sigma_{ci}})^2] \tag{4}$$

通过高斯分布概率密度求得,特征x_j的均值(u)为类标签C_i所有样本的平均值,方差(σ)同上,即每

个类体素的均值和方差分开估计[9,12-13]。

支持向量机在数据挖掘和机器学习领域得到了广泛的研究和应用,模型的建立所需的先验干预较少,在小样本处理、高维模式识别中表现出了自己的优越性。本文使用线性核函数支持向量机,工具包为libsvm-mat-2.89-3,核函数采用多项式(polynomial)[14]。

3.3 特征选择

针对fMRI海量数据,本实验采用两步降维方式,下面将详细介绍。

第一步,生成特征模板,完成对fMRI初步降维。从结构像生成特征模板(mask),主要是因为结构像比功能像分辨率高,选取特征更加准确。共生成两种:一种是感兴趣区域(ROI)脑区,包含1 156个体素,主要包括前额叶(PFC)、后侧顶叶(PPC),前扣带回(ACC)等。PFC、PPC、ACC是问题确决的主要脑区[15]。另一种是全脑(whole brain),去除3D结构像中非脑组织,包含45 546个体素。根据特征模板(mask),把3D功能像转换为一个长的特征向量。每个被试结构像与标准脑做过对齐,可以忽略头部大小的不同,所以特征模板对所有被试通用。

第二步,进一步对特征模板产生的数据进行降维。

选取最能区分不同任务的体素。对所有的体素分别训练单体素分类器,用单体素的时间序列作为特征,分类器采用高斯朴素贝叶斯,因为此算法简单,训练速度快[9]。由于数据集是小样本,评估分类器的真实准确率采用10折交叉验证的方法,准确率作为每个体素的分值,按分值对体素降序排序,选取期望的体素。设置体素集合,大小分别为top(20,40,80,100,200,400,600,800,1 000,1 156)。用这些特征集合进行多体素特征分类。

3.4 评估过程

评估分类器采用交叉验证(Cross Validation,CV)方法,考虑到计算复杂性,采用10_kold交叉验证,数据分为10个互不相交的组,每组包含样本的个数基本相等。在每组中,每个类所含样本的数目也是基本一致的。训练和测试共进行10次。在每次迭代中,首先选择1组数据作为测试,其他9组数据作为训练。然后在训练集中,进行特征选择,最后选择不同集合大小特征训练分类器。平均10次分类结果作为分类器的真实评估准确率,整个评估过程如图2所示。

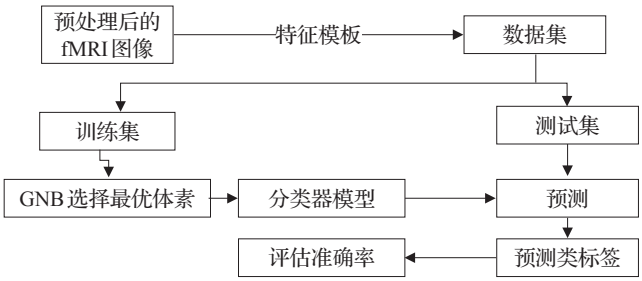

图2 评估过程

4 实验结果与分析

4.1 脑图像数据标准化

采用感兴趣脑区(ROI)中分类效果最好的80个体素,使用GNB、SVM分类器进行分类,数据集是标准化之前和之后的数据,数据分类准确率的结果如图3所示。

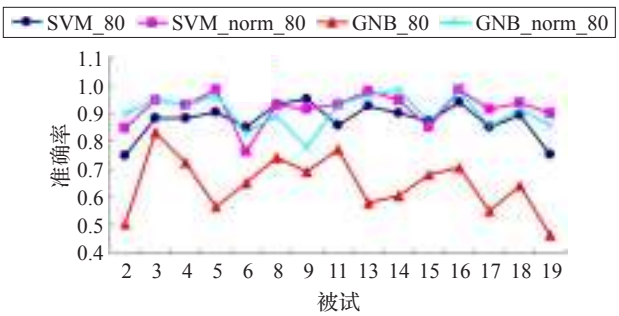

图3 脑图像标准化之后和之前比较

从图3可以得到,标准化之后,GNB分类器的准确率大幅度提高,所有被试(subject)平均提高了40%左右。标准化之前,所有被试平均准确率为64%,标准化之后,所有被试平均准确率为90%。SVM分类器虽然没有像GNB分类器大幅度提高,除了编号为6,9,15以外,大部分被试都有相应的提高,平均从原来的88%提高到92%,提高了4%。

脑图像标准化后,每个脑图像的数据更加服从正态分布,减少噪音,这可能是贝叶斯分类器性能提高的主要原因。

4.2 体素个数

选择感兴趣脑区(ROI)中分类最优的体素,分别选择20,40,60,80,100,200,400,600,800,1 156个体素集合大小作为分类器的特征。为了得到一个较真实分类器性能的结果,研究中对多个被试的平均准确率进行了计算,并将其作为每个体素集合大小对应的准确率。这样,每个分类器由11个不同体素集合大小所表示,分类结果如图4所示。

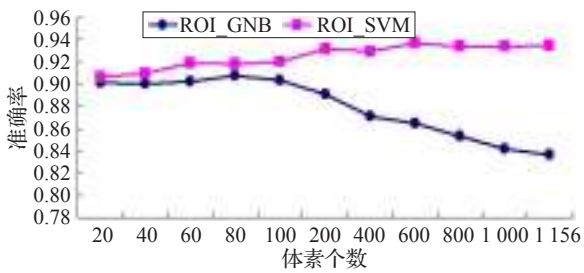

图4 体素个数比较

从图4可以看到,GNB分类器随着体素个数的增多,分类准确率逐步下降。SVM则相反。出现这种情况可能是体素之间存在相关性所引起的,将在今后的工作中继续深入研究。

4.3　不同分类器性能比较

为了使结果更接近于真实评估准确率,减少误差,决定对不同特征个数(20,40,60,80,100)求平均准确率,所有被试采用GNB、SVM分类,准确率的结果如图5所示。

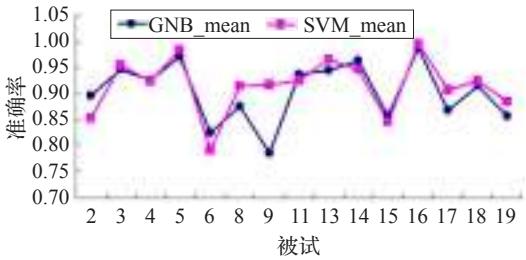

图5　不同分类器性能

从图5可知,SVM分类效果比GNB稍微好一点。GNB平均准确率是90%,SVM平均准确率是91.5%,15个中有11个高于GNB。

5　结论

利用结构像与功能像相结合的方法,选取感兴趣的脑区,减少特征的维数,然后再利用单体素分类器选取最能区别任务的体素(特征)。由实验的结果表明,此种方法是一种非常有效的特征选择方法,分类的准确率远远高于随机预测,相比全脑,平均准确度降低3个百分点,但分类器的速度大幅度提升,耗时不到原来的4%。

在分类器性能对比方面,SVM与GNB差不多,但是在高维特征(体素)上SVM优于GNB。在这点上与其他的研究领域,例如文本分类领域的研究结论是相一致的。

参考文献:

[1] Norman K A, Polyn A M, Detre G J, et al.Beyond mind-reading:multi-voxel pattern analysis of fMRI data[J]. Trends Cognitive Science,2006,10.

[2] Friston K J.Statistical parametric maps in functional imaging:a general linear approach[J].Human Brain Mapping, 1995,2:189-210.

[3] Goutte C, Toft P, Rostrup E, et al.On clustering fMRI time series[J].NeuroImage,1998,9:298-310.

[4] Mitchell T M, Hutchinson R, Just M, et al.Classifying instantaneous cognitive states from fMRI data[C]//Proceedings of the 2003 Americal Medical Informatics Association Annual Symposium, Washington D C,2003.

[5] Haxby J V, Gobbini M, Furey M L, et al.Distributed and overlapping representation of faces and objects in ventral temporal cortex[J].Science,2001,293.

[6] Hardoon D R.Unsupervised analysis of fMRI data using kernel canonical correlation[J].NeuroImage,2007,37(4).

[7] Langleben D D.Telling truth from lie in individual subjects with fast event-related fMRI[J].Human Brain Mapping,2005,26(4):262-272.

[8] Kuncheva L I, Rodríguez J J.Classifier ensembles for fMRI data analysis: an experiment[J].Magnetic Resonance Imaging,2010,28:583-593.

[9] Pereira F, Mitchell T, Botvinick M.Machine learning classifiers and fMRI:a tutorial overview[J].Neuroimage,2009, 45:199-209.

[10] 相洁.启发式问题解决认知神经机制及fMRI数据分析方法研究[D].太原:太原理工大学,2010.

[11] 范金金,刘鹏.朴素贝叶斯分类器的独立性假设研究[J].计算机工程与应用,2008,44(34):139-141.

[12] Han Jiawei, Kamber M.Data mining:concepts and techniques[M].范明,孟小峰,译.北京:机械工业出版社,2001.

[13] Mitchell T M, Hutchinson R, Niculescu R S, et al.Learning to decode cognitive states from brain images[J].Machine Learning,2004,57:145-175.

[14] Chang Chih-Chung, Lin Chih-Jen.LIBSVM—a library for support vector machines[EB/OL].(2010).http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html.

[15] 相洁,陈俊杰.基于SVM的fMRI数据分类:一种解码思维的方法[J].计算机研究与发展,2010,47(2):286-291.