DOI:10.11992/tis.201509011
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160105.1526.002.html

结合 Copula 理论与贝叶斯决策理论的分类算法

钱冬[1],王蓓[1],张涛[2],王行愚[1]
(1.华东理工大学 信息科学与工程学院,上海 200237; 2.清华大学 自动化系,北京 100086)

摘 要:传统的贝叶斯决策分类算法易受类条件概率密度函数估计的影响,可能会对分类结果造成干扰。对此本文提出来一种改进的贝叶斯决策分类算法,即 Bayesian-Copula 判别分类器(BCDC)。该方法无需对类条件概率密度函数的形式进行假设,而是将 Copula 理论和核密度估计相结合进行函数构建,利用核密度估计平滑特征的概率分布,概率积分变换将特征的累计概率分布转化为均匀分布,Copula 函数构建 2 个类别的边缘累积分布之间的相关性。随后,用极大似然估计方法确定 Copula 函数的参数,贝叶斯信息准则(BIC)用于选择最合适的 Copula 函数。通过生物电信号的仿真实验进行模型验证,结果表明相比传统的概率模型,提出的分类算法在分类精度和 AUC 两个性能指标上表现较好,鲁棒性更强,说明了 BCDC 模型充分利用 Copula 理论和核密度估计的优点,提高了估计的准确性和灵活性。

关键词:机器学习;贝叶斯决策理论;Copula 理论;核密度估计;生物电信号
中图分类号:TP391.4 **文献标志码**:A **文章编号**:1673-4785(2016)01-0078-06

中文引用格式:钱冬,王蓓,张涛,等.结合 Copula 理论与贝叶斯决策理论的分类算法[J]. 智能系统学报, 2016, 11(1): 78-83.
英文引用格式:QIAN Dong, WANG Bei, ZHANG Tao, et al. Classification algorithm based on Copula theory and Bayesian decision theory[J]. CAAI Transactions on Intelligent Systems, 2016, 11(1): 78-83.

Classification algorithm based on Copula theory and Bayesian decision theory

QIAN Dong[1], WANG Bei[1], ZHANG Tao[2], WANG Xingyu[1]
(1. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China; 2. Department of Automation, Tsinghua University, Beijing 100086, China)

Abstract: Traditional Bayesian decision classification algorithm is easily affected by the estimation of class-conditional probability densities, a fact that may result in incorrect classification results. Therefore, this paper proposes an improved classification algorithm based on Bayesian decision, i.e., Bayesian-Copula Discriminant Classifier (BCDC). This method constructs class-conditional probability densities by combining Copula theory and kernel density estimation instead of making assumptions on the form of class-conditional probability densities. Kernel density estimation is used to smooth the probability distribution of each feature. By performing probability integral transform, continuous distribution is converted to random variables having a uniform distribution. Then, Copula functions are used to construct the dependency structure between these probability distributions for two categories. Moreover, the maximum likelihood estimation is applied to determine the parameters of Copula functions, and two well-fitted Copula functions for two categories are selected based on Bayesian information criterion. The BCDC method was validated with experimental datasets of physiological signals. The obtained results showed that the proposed method outperforms other traditional methods in terms of classification accuracy and AUC as well as robustness. Moreover, it takes full advantage of Copula theory and kernel density estimation and improves the accuracy and flexibility of the estimation.

Keywords: machine learning; Bayesian decision theory; Copula theory; kernel density estimation; physiological signals

机器学习在人工智能领域的研究中具有十分重要的地位。目前,其应用已遍及人工智能的各个分支,如模式识别、计算机视觉、数据挖掘、医学诊断、自然语言处理等领域[1-6]。概率模型则是模式识别中被研究较多的一类模型,它给予了数据产生的复杂现象和内在机理的描述方式。其中,贝叶斯理论是基于概率表达的机器学习的主要工具,其认

收稿日期:2015-09-06. 网络出版日期:2016-01-05.
基金项目:上海市科委科技创新行动计划-生物医药领域产学研医合作资助项目(12DZ1940903).
通信作者:王蓓. E-mail:beiwang@ecust.edu.cn.

为:先验信息反映了试验前对总体参数分布的认识,在观察到样本信息后,对此认识有了改变,其结果反映在后验信息中,后验信息综合了样本信息和参数的先验信息[7]。

产生式模型(generative model)和判别式模型(discriminative model)是 2 个比较常见的有监督学习的分类模型。产生式模型可以指定数据结构的先验信息,但需要对观测数据建立正确的模型,而不是对类别分布进行建模,如贝叶斯决策理论;判别式模型则是通过最大化类别的概率学习模型,如 Logistic Regression (LR)[8-9]。然而,在实际使用中,贝叶斯决策理论仍然存在着一定的局限性。

贝叶斯决策理论是解决模式分类问题的一种基本统计方法。该理论的出发点是利用概率的不同分类决策与相应的决策代价之间的定量折中;目的则是对未知的数据所属的类别做出判决[10]。由于缺乏对于数据结构的信息,贝叶斯决策理论中类条件概率密度函数通常是很难准确估计的。

目前,估计类条件概率密度函数的方法主要有 2 种,但两者都是基于一定的假设条件。第一种是假设类条件概率密度函数服从多元高斯分布,简称为高斯判别分类器(Gaussian discriminant classifier, GDC)[11]。然而,多元高斯分布的边缘分布是一元高斯分布,该一元高斯分布并非和实际特征的概率分布相吻合。所以,该假设条件并不能准确地表现出多元变量的依赖结构。更重要的是,多元高斯分布中的协方差矩阵只能反映出各个特征之间的线性关系,难以精确地描述特征之间的非线性关系。第 2 种则是基于朴素贝叶斯条件独立的特点,假设类条件概率密度函数服从若干个一元高斯分布,简称为高斯朴素分类器(Gaussian naive Bayes classifier, GNBC)[12]。该假设条件虽然可以有效地减少参数估计的个数,但它过于简单,直接忽略了各个特征之间的依赖结构。因此,该方法也不能准确地估计出多个特征的联合分布。

由上述可知,现有的估计方法都存在着一定的不足和局限性。本文考虑了特征之间存在的依赖关系,提出了将贝叶斯决策理论和 Copula 理论相结合的分类器,简称为 Bayesian-Copula 判别分类器。该模型将 Copula 函数和核密度估计相结合构建类条件概率密度函数。Copula 函数能够描述变量间的线性或者非线性相关性,该理论表明多元联合分布函数可以通过 Copula 函数和任意的随机变量的边缘分布函数构建[13-15]。而核密度估计则是一种非参数估计方法,它不需要假设概率分布的形式,可以直接计算得到概率密度值[16]。最后,将改进的 BCDC 算法用于生物电信号分类识别的实际问题中进行模型的验证。由于从生物电信号中提取的特征之间存在依赖关系,在分类精度和 AUC 两个指标上,相比于传统的 GDC、GNBC 和 LR 模型,所提出的方法呈现出更好的分类效果。因此,该模型可以被用于处理特征间存在一定的相关性的实际问题,为机器学习问题提供了一种新的方法。

1　Bayesian-Copula 判别分类器

1.1　贝叶斯决策理论

贝叶斯决策理论表明对未知的数据 $\boldsymbol{x}$ 所属的类别做出判决,可以通过计算 $\boldsymbol{x}$ 属于某一个类别的概率值得到,因此通过贝叶斯公式,该概率值可表示为

$$P(C_k \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C_k)P(C_k)}{p(\boldsymbol{x})}$$

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x} \mid C_k)P(C_k)$$

$$k = 1,2,\cdots,K \tag{1}$$

式中:$\boldsymbol{x}$ 表示特征向量,即 $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$,$n$ 为特征的个数,K 为类别个数,$P(C_k)$ 是类别 C_k 的先验概率,$P(C_k \mid \boldsymbol{x})$ 则是相应的后验概率,$p(\boldsymbol{x} \mid C_k)$ 是类条件概率密度函数。此外,$p(\boldsymbol{x})$ 仅仅是一个标量,以保证各类别的后验概率总和为 1。贝叶斯公式表明,通过观察数据 $\boldsymbol{x}$,先验概率可以转换为后验概率。

根据最小化误差概率的准则,未知数据 $\boldsymbol{x}$ 将被归于后验概率 $P(C_k \mid \boldsymbol{x})$ 最大的类别。考虑到 $p(\boldsymbol{x})$ 只是一个标量因子,所以式(1)可以简化为

$$P(C_k \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid C_k)P(C_k) \tag{2}$$

注意到,在式(2)中,后验概率 $P(C_k \mid \boldsymbol{x})$ 主要由先验概率 $P(C_k)$ 和类条件概率密度函数 $p(\boldsymbol{x} \mid C_k)$ 的乘积所决定。先验概率 $P(C_k)$ 可以经验性地获得,计算在训练数据中属于某一类别的数据个数,再除以训练数据的总个数即可得到。

在下面小节中,我们将通过 Copula 函数和核密度估计的方法来构建类条件概率密度函数。

1.2　Copula 理论

近年来,在统计领域里,Copula 理论引起了研究者的关注。该理论可以理解为:多维随机变量的联合分布函数可以分解成若干个一维的分布函数和一个 Copula 函数,而 Copula 函数则将若干个分布函数连接起来,它可以描述随机变量间的依赖关系。目前,该理论被广泛应用于经济、金融等领域[17-18]。Sklar 定理是 Copula 理论的核心部分,也是 Copula 理论在统计学中应用的基础,在建立联合分布函数和它们相应边缘分布函数之间的关联中起着关键的作用。

定理　(Sklar 定理 (1959)):令 H 为 n 个随机

变量 X_1, X_2, …, X_n 的联合分布函数,令 $F_1(x_1)$, $F_2(x_2)$, …, $F_n(x_n)$ 为随机变量的边缘分布函数,如果所有的边缘分布函数都是连续的,那么存在唯一一个 Copula 函数 C 满足:

$$H(x_1,\cdots,x_n)=C(F_1(x_1),\cdots,F_n(x_n)) \tag{3}$$

联合密度函数 h 被定义为

$$h(x_1,\cdots,x_n)=c(F_1(x_1),\cdots,F_n(x_n))\cdot\prod_{i=1}^{n}f_i(x_i) \tag{4}$$

$$c(F_1(x_1),\cdots,F_n(x_n))=\frac{\partial C(F_1(x_1),\cdots,F_n(x_n))}{\partial F_1(x_1)\cdots\partial F_n(x_n)} \tag{5}$$

式中:$c(F_1(x_1),\cdots,F_n(x_n))$ 是一个 n 维的 Copula 密度函数,$f_i(x_i)$ 则是每个随机变量的密度函数。

推论 如果 C 是一个 Copula 函数,C 的值域为 $[0,1]^n$,$F_1(x_1)$, $F_2(x_2)$, …, $F_n(x_n)$ 为随机变量的边缘分布函数,那么 $C(F_1(x_1),\cdots,F_n(x_n))$ 可以定义一个联合分布函数。

通过 Copula 理论,式(2)可以被推导出

$$P(C_k\mid\boldsymbol{x})\propto c\{F_1(x_1),\cdots,F_n(x_n)\mid\theta;C_k\}\cdot\prod_{i=1}^{n}f_i(x_i\mid C_k)\cdot P(C_k) \tag{6}$$

式中:θ 是 Copula 密度函数的参数,右边第 1 项表示属于类别 C_k 的 Copula 密度函数,右边第 2 项表示属于类别 C_k 的核密度函数。

Copula 函数连接的是每个特征的累积分布函数 $F_i(x_i)$,而累积分布函数的值域是[0,1],因此,当每个特征都是连续的随机变量时,需对数据进行概率积分变换,计算出每个特征的经验累积分布,该方法可以使任意给定的分布转换为均匀分布。

1.3 边缘分布估计

式(4)表明,一个联合概率密度函数可以分解为一个 Copula 密度函数和 n 个边缘密度函数。非参数估计的方法,如直方图和核密度估计,可以直接利用样本来估计变量的密度函数。考虑到直方图的缺点,核密度估计被用来估计每个特征的概率密度函数。假设有 N 个样本 x_i,对于一个新来的样本 x,核密度估计的方法可以定义为

$$\hat{f}_h(x)=\frac{1}{N}\sum_{i=1}^{N}K_h(x-x_i)=\frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{x-x_i}{h}\right) \tag{7}$$

式中:$K(\cdot)$ 是核函数,h 是平滑参数,本文中,采用高斯核函数,因此,式(7)可以表示为

$$\hat{f}_h(x)=\frac{1}{\sqrt{2\pi}Nh}\sum_{i=1}^{N}\mathrm{e}^{-\frac{(x-x_i)^2}{2h^2}} \tag{8}$$

1.4 Copula 函数参数估计

采用极大似然估计的方法对 Copula 密度函数的参数 θ 进行估计,可以得到 θ 的估计值:

$$\theta^*=\arg\max_{\theta}\sum_{i=1}^{N}\log c\{F_{i1}(x_{i1}),\cdots,F_{in}(x_{in});\theta\} \tag{9}$$

此外,为了校准参数 θ,我们充分使用了随机数的性质,从估计的 Copula 密度函数中生成 10 000 个随机数,然后用极大似然估计的方法对生成的随机数重新进行参数拟合,估计出最终的参数 θ。

1.5 模型选择

目前广泛使用的 Copula 密度函数主要分为两大类:elliptical Copulas 和 Archimedean Copulas。在本文中,主要使用的是 elliptical Copulas 中的多元 Gaussian Copula 函数和多元 Student-t Copula 函数。

通常,Copula 模型的选择会对后续步骤造成一定的影响。因此,贝叶斯信息准则(Bayesian information criterion, BIC)用来对 Copula 模型进行选择,它是模型拟合程度和模型复杂度之间的权衡,BIC 值较小的 Copula 密度函数会被用于构建类条件概率密度函数。

$$\mathrm{BIC}=-2\log L(\theta^*)+m\log(k) \tag{10}$$

式中:$L(\theta^*)$ 是估计的似然值,m 表示 Copula 密度函数中参数的个数,k 表示数据的个数。

2 生物电信号的分类识别

通过检测受试者在白天短时睡眠过程中的困倦状态(drowsiness)和觉醒状态(alertness)这一个实际问题,验证所提出方法的有效性。通常对生物电信号问题进行分析和识别,需要经过信号的数据采集、特征提取和模式分类 3 个步骤[19-22]。考虑到从生物电信号中提取的特征参数能反映人的生理状态,而且特征之间可能存在一定的相关性,所以 BCDC 模型可以用于进行状态检测。

2.1 数据采集

共有 8 名受试者参与了白天短时睡眠的实验,将受试者安排在一个安静舒适的环境内,记录其午后 30 分钟的睡眠数据。原始睡眠数据的采集按照多导睡眠描记图(PSG, Ploysomnograph)的标准记录方式,包括了 4 导脑电信号(C_3-A_2, C_4-A_1, O_1-A_2, O_2-A_1),并同步采集了 2 导眼电信号(LOC-A_1, ROC-A_2),1 导肌电信号和 1 导心电信号。其中脑电、眼电和心电信号的采样频率为 100 Hz,肌电信号的采样频率为 200 Hz,高频截至频率是30 Hz,时间常数是0.3 s。本文主要分析 4 导脑电信号(EEG)和 2 导眼电信号(EOG)。

2.2　特征提取

考虑到在 20 s 的时间内,受试者的状态可能有所变化,因而特征参数可能也会有较大的波动,所以将受试者原始每段 20 s 的脑电和眼电信号进一步划分为 5 s 一段和 2.5 s 的重叠窗,提高特征参数的准确性,并对 5 s 的数据进行 512 个点的快速傅立叶变换(FFT),计算每个 5 s 内脑电信号和眼电信号的特征,对所有 5 s 的特征参数取平均值,将其作为 20 s 数据的特征参数,以减少干扰。选取的特征分别对应于 C_3/C_4 导联的 θ 波(4~8 Hz)和 O_1/O_2 导联的 α 波(8~13 Hz)的脑电能量占空比和左、右眼电信号的频域能量和(2~10 Hz),即特征向量 $\boldsymbol{x}=\{D_\theta, D_\alpha, S_{\mathrm{LOC}}, S_{\mathrm{ROC}}\}$。特征参数计算公式如表 1。

表 1　脑电信号和眼电信号中提取的特征参数

Table 1　Features extracted from EEG and EOG signals

信号	意义	特征参数
EEG	能量占空比/%	$D_\theta = \max\{\frac{S_\theta(C_3)}{S_T(C_3)}\times 100\%, \frac{S_\theta(C_4)}{S_T(C_4)}\times 100\%\}$ $D_\alpha = \max\{\frac{S_\alpha(O_1)}{S_T(O_1)}\times 100\%, \frac{S_\alpha(O_2)}{S_T(O_2)}\times 100\%\}$
EOG	频域能量和/μV^2	$S_{\mathrm{LOC}}(\mathrm{LOC})$, $S_{\mathrm{ROC}}(\mathrm{ROC})$

表 1 中 θ (4~8Hz),α(8~13 Hz),T(0.5~25 Hz);LOC, ROC(2~10 Hz)。

2.3　模式分类

2.3.1　参数优化和模型选择

首先,对数据集做归一化处理,随机选取 70% 的数据作为训练数据,30% 的数据作为测试数据进行分析。然后,针对每一个类别,通过概率积分变换计算训练数据中 4 个特征的经验累积分布,并用 kendall 秩相关系数表示两两特征之间的相关性。相关性如下所示:

$$\boldsymbol{C}_1^{\mathrm{tau}}=\begin{bmatrix}1 & -0.413\,7 & -0.275\,3 & -0.289\,5\\ -0.413\,7 & 1 & 0.228\,8 & 0.247\,0\\ -0.275\,3 & 0.228\,8 & 1 & 0.801\,8\\ -0.289\,5 & 0.247\,0 & 0.801\,8 & 1\end{bmatrix}$$

$$\boldsymbol{C}_2^{\mathrm{tau}}=\begin{bmatrix}1 & -0.539\,9 & -0.174\,5 & -0.187\,5\\ -0.539\,9 & 1 & 0.254\,1 & 0.198\,3\\ -0.174\,5 & 0.254\,1 & 1 & 0.728\,6\\ -0.187\,5 & 0.198\,3 & 0.728\,6 & 1\end{bmatrix} \quad (11)$$

从以上 2 个矩阵可知,每一个类别的特征之间存在正、负相关性,有些特征间的相关性比较微弱,这主要是由于不同的受试者对 2 个状态存在一定的差异性。

随后,对 Copula 密度函数的参数 θ 进行极大似然估计,并用随机数的性质重新校准参数 θ。最后,采用 BIC 选取最合适的 Copula 密度函数,并与核密度估计相结合,构建类条件概率密度函数,BIC 选取的模型如表 2 所示。

表 2　基于 BIC 选取的 2 个类别的 Copula 密度函数

Table 2　Copula density functions for two categories based on BIC

Copula 密度函数	觉醒状态(A)	困倦状态(D)
Gaussian Copula	−451.63	−477.25
Student-t Copula	−459.46	−471.83

BIC 值较小的 Copula 函数会被选择,所以针对 alertness 类别选取的是 Student-t Copula 函数,而 drowsiness 类别选取的是 Gaussian Copula 函数。

2.3.2　模式分类和模型比较

将改进的 BCDC 算法与 GDC、GNBC 和 LR 对测试数据进行分析和比较。ROC 曲线被用来表现分类器的性能,它通过将连续变量设定出多个不同的阈值来揭示真阳率(true positive rate, TPR)和假阳率(false positive rate, FPR)的相互关系。其横轴表示真阳率,纵轴表示假阳率,曲线下面积越大,分类器分类的能力越强。图 1 呈现出 4 个分类器在测试数据上的 ROC 曲线,其中连接点(0,0)和(1,1)的直线表示随机猜测。相比其他 3 个方法,BCDC 算法的曲线处于左上角,所以该方法表现出较好的分类能力。

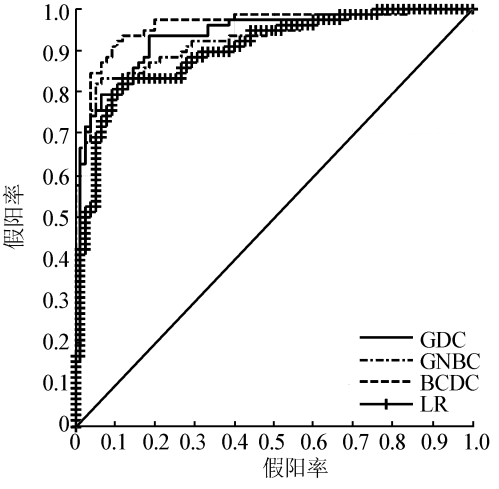


图 1　GDC、GNBC、BCDC、LR 的 ROC 曲线

Fig.1　ROC curves obtained by GDC, GNBC, BCDC, LR, respectively

为了进一步定量地检验 4 个分类器识别的准确性,通过分类精度和 AUC 两个性能指标对分类器进行评价。考虑到训练数据和测试数据是随机选取的,数据中存在的个体差异性可能会影响分类器的性能评估,所以将随机实验循环 50 次,得到分类器的平均分类精度和平均 AUC,如表 3 所示。

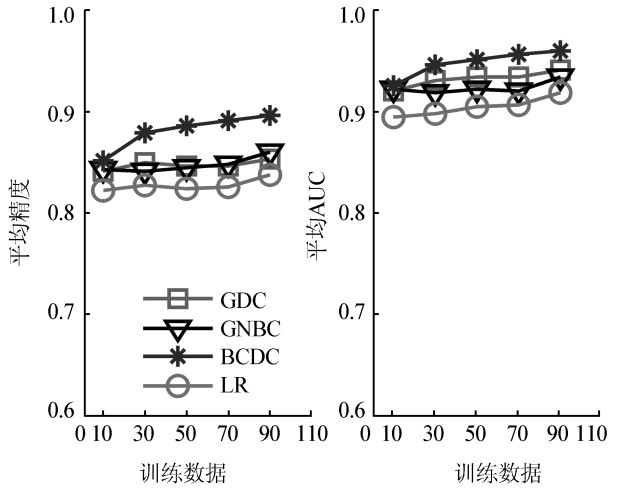
表 3 GDC、GNBC、BCDC、LR 的平均精度、平均 AUC 值和相应的标准差

Table 3 Average Accuracy, Average AUC and corresponding standard deviation obtained by GDC, GNBC, BCDC and LR, respectively

分类器	平均精度(标准差)	平均 AUC(标准差)
GDC	0.855 9(0.025 7)	0.940 8(0.012 9)
GNBC	0.858 8(0.025 8)	0.925 3(0.016 9)
LR	0.838 2(0.023 9)	0.912 0(0.017 3)
BCDC	0.902 6(0.017 9)	0.963 4(0.010 3)

从表 3 可知,本文提出的 BCDC 算法在两个分类指标上呈现出更好的分类表现。就平均精度而言,BCDC 识别的精度高于其他 3 个分类器大约 5% 左右,同时标准差也小于其他 3 个分类器。而对于 AUC,尽管 GDC 相对接近于 BCDC,但 BCDC 的 AUC 值大于其他 3 个方法,且标准差也较小,呈现出更强的稳定性。

为了了解不同分类器在不同数量的数据集上的分类能力,从数据中分别随机选取 10%、30%、50%、70%和 90%的数据作为训练数据,用剩余的测试数据评估 4 个分类方法,结果如图 2 所示。



(a)平均精度 (b)平均 AUC

图 2 GDC、GNBC、BCDC、LR 在不同训练数据个数下的平均精度和平均 AUC

Fig.2 Average accuracy and average AUC obtained by GDC, GNBC, BCDC, and LR based on the different subsets of the training data

分析数据可得:当训练数据较少时(10%),4 个方法表现出几乎相同的平均精度,BCDC 并没有产生显著的识别精度。当训练数据增加(30%),提出的方法的分类表现很快超越了其他 3 个分类器。当数据量大于 30%,BCDC 表现出更高的分类表现。总而言之,当 30%、50%、70%和 90%作为训练数据时,相比较 GDC、GNBC、LR,改进的 BCDC 的分类能力更强。由图 2 表明,增加训练数据个数能够提供更多的某种特定类别的信息,从而更加准确地判断类别。

作为一种监督式学习方法,BCDC 算法通过参数优化和模型选择提高了类条件概率密度函数估计的准确性。虽然训练时间大约是 10 s,但是在不同数据量的条件下,BCDC 算法呈现出更好的平均分类精度和平均 AUC。

3 结束语

本文提出了基于贝叶斯决策理论和 Copula 理论的分类算法。该算法在实际运用过程中,参数 Copula 模型和核密度估计相结合提升类条件概率密度函数估计的准确性。相比较其他传统的贝叶斯决策模型,Bayesian-Copula 判别分类器能够在实际的生物电信号分类识别问题中得到较好的分类效果。

Copula 模型的优势主要是不需要对边缘分布的形式进行假设,在模型中,我们仅仅计算每个特征的经验累积分布,用不同的 Copula 函数建立特征间的依赖结构。该模型简单、易懂,在对未知数据建立模型时,具有更多的灵活性。对于许多实际问题,概率模型中独立同分布的假设通常是不成立的。所以,通过 Copula 理论能够提高对联合分布估计的准确性。

参考文献:

[1] TIPPING M E. Sparse Bayesian learning and the relevance vector machine[J]. Journal of machine learning research, 2001, 1(3): 211-244.

[2] XUE Jinghao, HALL P. Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis? [J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(5): 1109-1112.

[3] FERNÁNDEZ-DELGADO M, CERNADAS E, BARRO S, et al. Do we need hundreds of classifiers to solve real world classification problems? [J]. Journal of machine learning research, 2014, 15(1): 3133-3181.

[4]RODRIGUEZ A, LAIo A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[5]李宏伟, 刘扬, 卢汉清, 等. 结合半监督核的高斯过程分类[J]. 自动化学报, 2009, 35(7): 888-895.
LI Hongwei, LIU Yang, LU Hanqing, et al. Gaussian processes classification combined with semi-supervised kernels[J]. Acta automatica sinica, 2009, 35(7): 888-895.

[6]BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2001, 3(4-5): 993-1022.

[7]BISHOP C M. Pattern Recognition and Machine Learning [M]. New York: Springer, 2006: 21-31.

[8]NG A Y, JORDAN M I. On discriminative vs. generative classifiers: a comparison of logistic regression and naïve Bayes[C]//Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada, 2002, 14: 841-848.

[9]李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 77-91.

[10]JAIN A K, DUIN R P W, MAO Jianchang. Statistical pattern recognition: a review[J]. IEEE transactions on pattern analysis and machine intelligence, 2000, 22(1): 4-37.

[11]DUDA R O, HART P E, STORK D G. Pattern Classification[M]. 2nd ed. New York: Wiley, 2001: 20-45.

[12]MURPHY K P. Machine Learning: A Probabilistic Perspective[M]. England: MIT, 2012: 82-87.

[13]NELSEN R B. An Introduction to Copulas[M]. 2nd ed. Springer: Berlin, 2006.

[14]GENEST C, FAVRE A C. Everything you always wanted to know about Copula modeling but were afraid to ask[J]. Journal of hydrologic engineering, 2007, 12(4): 347-368.

[15]EBAN E, ROTHSCHILD G, MIZRAHI A, et al. Dynamic Copula networks for modeling real-valued time series [C]//Proceedings of the 16th International Conference on Artificial Intelligence and Statistics. Scottsdale, AZ, USA, 2013, 4: 247-255.

[16]KRISTAN M, LEONARDIS A, SKOC AJ D. Multivariate online kernel density estimation with Gaussian kernels[J]. Pattern recognition, 2011, 44(10-11): 2630-2642.

[17]CHERUBINI U, LUCIANO E, VECCHIATO W. Copula Methods in Finance[M]. England: John Wiley & Sons, 2004.

[18]PATTON A J. A review of Copula models for economic time series[J]. Journal of multivariate analysis, 2012, 110: 4-18.

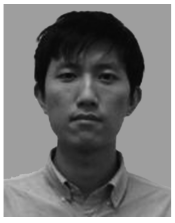
[19]AUBASI A. Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders[J]. Computers in biology and medicine, 2013, 43(5): 576-586.

[20]TAGLUK M E, SEZGIN N, AKIN M. Estimation of sleep stages by an artificial neural network employing EEG, EMG and EOG[J]. Journal of medical systems, 2010, 34(4): 717-725.

[21]CICHOCKI A, MANDIC D, DE LATHAUWER L, et al. Tensor decompositions for signal processing applications: from two-way to multiway component analysis[J]. IEEE signal processing, 2015, 32(2): 145-163.

[22]KHUSHABA R N, KODAGODA S, LAL S, et al. Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm[J]. IEEE transactions on biomedical engineering, 2011, 58(1): 121-131.

作者简介:



钱冬,男,1990 年生,硕士研究生,主要研究方向为机器学习、生物电信号。

王蓓,女,1976 年生,副研究员,主要研究方向为智能信息处理和模式分类、复杂系统及其在人工生命科学中的应用。曾参与国家自然科学基金、上海市科委科技创新行动计划等项目。发表学术论文 50 余篇,被 SCI、EI 检索 30 余篇。

张涛,男,1969 年生,教授,博士生导师,主要研究方向为控制理论及应用、信号处理、机器人控制等。主持或参与国家 973 项目、国家 863 项目、国家自然科学基金项目多项。曾获得教育部自然科学奖、军队科技进步奖、中国电子信息科学技术奖等。发表论文 200 余篇,其中被 SCI 检索 40 余篇,EI 检索 120 余篇。